Advanced Multilevel Monte Carlo Methods

BY AJAY JASRA[1], KODY LAW[2], & CARINA SUCIU[3]

[1,3]Computer, Electrical and Mathematical Sciences & Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, KSA.

E-Mail: ajay.jasra@kaust.edu.sa; oana.suciu@kaust.edu.sa

[2]School of Mathematics, University of Manchester, Manchester, M13 9PL, UK.

E-Mail: kodylaw@gmail.com

Abstract

This article reviews the application of some advanced Monte Carlo techniques in the context of Multilevel Monte Carlo (MLMC). MLMC is a strategy employed to compute expectations which can be biased in some sense, for instance by using the discretization of an associated probability law. The MLMC approach works with a hierarchy of biased approximations which become progressively more accurate and more expensive. Using a telescoping representation of the most accurate approximation, the method is able to reduce the computational cost for a given level of error versus i.i.d. sampling from this latter approximation. All of these ideas originated for cases where exact sampling from couples in the hierarchy is possible. This article considers the case where such exact sampling is not currently possible. We consider some Markov chain Monte Carlo and sequential Monte Carlo methods which have been introduced in the literature and we describe different strategies which facilitate the application of MLMC within these methods.

Key words: Multilevel Monte Carlo, Markov chain Monte Carlo, Sequential Monte Carlo, Ensemble Kalman filter, Coupling.

1 Introduction

Let E be a space, π be a probability density on E and $\varphi : \mathsf{E} \to \mathbb{R}$ be a $\pi-$integrable function. In this article we are concerned with the computation of

$$\mathbb{E}_{\pi}[\varphi(U)] = \int_{\mathsf{E}} \varphi(u)\pi(u)du \qquad (1)$$

for many different $\pi-$integrable functions φ. In particular, let $\kappa : \mathsf{E} \to \mathbb{R}_+$, then we consider

$$\pi(u) = \frac{\kappa(u)}{Z}, \tag{2}$$

where $Z = \int_{\mathsf{E}} \kappa(u)du < +\infty$ is not known, and calculating Z is also of interest. For example, the case of (2) could correspond to a Bayesian inference problem where κ is the likelihood multiplied by the prior and Z is the marginal likelihood on the observed data. To elaborate further, the computation of (1) can be associated to posterior expectations, and the value of Z can be used for Bayesian model selection; see [89]. Many of these problems are found in many real applications, such as meteorology, finance and engineering; see [75, 89]. Later in this article, we will expand upon the basic problem here.

We focus on the case when simulating from π requires the solution of a complex continuum problem. For instance, it may require solution of a continuous-time stochastic process or the solution of a partial differential equation (PDE). However, the methodology described in this article is not constrained to such examples. Also, we will assume that:

1. One must resort to numerical methods to approximate (1) or Z.

2. One can, at best, hope to approximate expectations w.r.t. some *biased* version of π, call it π_L. It is explicitly assumed that this bias is associated with a *scalar* parameter $h_L \in \mathbb{R}_+$ and that the bias disappears as $h_L \to 0$. An example of what we mean here, is that, if (2) holds and κ is a function of the solution of a PDE that can only be numerically solved up-to accuracy h_L, then one works with the κ which uses this numerical solution. Then one may have (the point will be clarified in details later in the article)

$$\Big| \int_{\mathsf{E}} \varphi(u)\pi(u)du - \int_{\mathsf{E}} \varphi(u)\pi_L(u)du \Big| \leq Ch_L^{\alpha}$$

for some $C < +\infty$, $\alpha > 0$.

3. When using state-of-the-art Monte Carlo methods, exact sampling from π_L is not possible; that is, one cannot sample i.i.d. from π_L.

Examples of models satisfying 1. & 2. include laws of stochastic differential equations (SDE) for which one cannot sample exactly (e.g. [73]), and one resorts to Euler or Milstein discretization. Examples satisfying 1.-3. include for example general Bayesian instances of the above (outside of very specific cases where sampling may be possible), where one updates the prior probability distribution based on noisy data to obtain the posterior conditional on the observed data (e.g. [53]) or general models where approximate Bayesian computation (e.g. [80]) must be used. Concrete examples will be given in Section 2.

1.1 Methodology Reviewed

The objective of this article is to provide a review of *advanced* methodology used to implement the MLMC method, of which we will explain in details in Section 3, but some basic notions are mentioned here. We reiterate that the standard MLMC method to be discussed in Section 3 (as in [44, 45, 51]) cannot be (easily) used in the contexts of our points 1-3 in the above paragraph. The MLMC method works with a hierarchy of biased probability densities $\pi_1, \ldots, \pi_L$ associated to scalar accuracy parameters $0 < h_L < \cdots < h_1 < +\infty$. In the original approaches of [44, 45, 51], point 3 above does not hold, so ordinary Monte Carlo approximation of (1) is possible by i.i.d. sampling from π_L which leads to a numerical error. In some examples, one can show that using MLMC to achieve the same numerical error as Monte Carlo, the associated computational cost is reduced. This is the main attractive feature of MLMC which one would like to replicate in other examples.

One of the critical ingredients of the MLMC method is sampling dependent couples of the pairs (π_l, π_{l-1}), $2 \leq l \leq L$. This means sampling a pair of dependent random variables X and Y, so that marginally X is distributed according to π_l and Y is distributed according to π_{l-1} and the random variables X and Y are not statistically independent. In addition to this, one would like $\mathbb{E}(X - Y)^2$ to be small, and the motivation for this will be described in detail in Section 3. This task, one might argue, is even more challenging than sampling from π_L for a single given L. In the context of interest in this paper (points 1.-3. above), one might use Markov chain Monte Carlo (MCMC – see e.g., [89, 92]) or Sequential Monte Carlo (SMC – see e.g., [25, 26, 35, 34]) to overcome the challenge of not being able to obtain i.i.d. samples from π_L. However, it is non-trivial to extend these methods for application of the MLMC method. The main issue is how can one utilize an MLMC approach in this context so that one reduces the cost relative to exact sampling from π_L, for a given numerical error. There have been many works on this topic and the objective of this article is to review these ideas as well as to identify important areas which could be investigated in the future.

The challenge lies not only in the design and application of the method, but in the subsequent analysis of the method, i.e., verifying that indeed it yields an improvement in cost for a given level of error. For instance, the analysis of MCMC and SMC rely upon techniques in Markov chains (e.g., [81, 92]) and Feynman-Kac formulae (e.g., [25, 26]), respectively. We highlight these techniques during our review.

1.2 Structure

This article is structured as follows. In Section 2, we give a collection of motivating examples from applied mathematics and statistics which are of interest from a practical perspective, and for which the application of multilevel methods would make sense. The problems chosen are simple toy problems which exemplify the types of problems one might encounter in practice. In Section 3 the basic MLMC method is reviewed. In Section 4, we give a short review of the single level version of some of the computational methods which this review is focussed on. In Section 5, we review several methods which have been adopted in the literature to date, mentioning some benefits and drawbacks of each approach. In Section 6, some discussion of the potential for future work is provided.

We end this introduction by mentioning that this review is not intended to be comprehensive. For instance, we do not discuss quasi-Monte Carlo methods or debiasing methods (e.g., [88]). An effort is made to discuss as much work as possible that exists under the umbrella of advanced MLMC methods. Furthermore, we mention here that this article tells a story. In particular, the culmination is in Section 5, however it may be difficult for even an expert to jump directly into this Section. Rather, Section 5 details some strategies for leveraging the fundamental benefits of the framework described in Section 3 in the context of the methods of Section 4, and as applied to the problems laid out in Section 2.

2 Motivating Examples

2.1 Bayesian Inverse Problems

We consider the following example as it is described in [9] (see also [53] and the references therein).

We introduce the nested spaces $V := H^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega) =: V^*$, where the domain Ω will be defined later. Furthermore, denote by $\langle \cdot, \cdot \rangle, \|\cdot\|$ the inner product and norm on L^2, and by $\langle \cdot, \cdot \rangle, |\cdot|$ the finite dimensional Euclidean inner product and norms. 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).

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

$$-\nabla \cdot (\hat{u} \nabla p) = f \quad \text{in } \Omega \ , \tag{3}$$

$$p = 0 \quad \text{on } \partial\Omega \ , \tag{4}$$

where

$$\hat{u}(x) = \bar{u}(x) + \sum_{k=1}^{K} u_k \sigma_k \phi_k(x) \ . \tag{5}$$

Define $u = \{u_k\}_{k=1}^K$, with $u_k \sim \mathcal{U}[-1,1]$ i.i.d. (the uniform distribution on $[-1,1]$). This determines the prior distribution for u. We note that for the methodology to be described later on, the prior need not be uniform, but this is the prior that will adopted in this article for simplicity of exposition. Assume that $\bar{u}, \phi_k \in C^\infty$ for all k and that $\sup_x |\phi_k(x)| = 1$. In particular, assume $\{\sigma_k\}_{k=1}^K$ decay with k. The state space is $\mathsf{E} = \prod_{k=1}^K [-1,1]$. Assume the following property holds: $\inf_x \hat{u}(x) \geq \inf_x \bar{u}(x) - \sum_{k=1}^K \sigma_k \geq u_* > 0$ so that the operator on the left-hand side of (3) is uniformly elliptic. Let $p(\cdot; u)$ denote the weak solution of (3) 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, $Y \in \mathbb{R}^m$

$$Y = \mathcal{G}(p) + \xi \ , \quad \xi \sim \mathcal{N}_m(0, \Gamma) \ , \quad \xi \perp u \ , \tag{6}$$

where $\mathcal{N}_m(0, \Gamma)$ denotes the $m-$dimensional Gaussian distribution with mean 0 and covariance Γ, and $\perp$ denotes independence. The unnormalized density for $u \in \mathsf{E}$ is then is given by:

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

2.1.1 Approximation

Consider the triangulated domains (with sufficiently regular triangles) $\{\Omega^l\}_{l=1}^\infty$ approximating Ω, where l indexes the number of nodes $d_l \propto h_l^{-D}$, for triangulation diameter h_l, so that we have $\Omega^1 \subset \cdots \subset \Omega^l \subset \Omega^\infty := \Omega$. Furthermore, consider a finite element discretization on Ω^l consisting of H^1 functions $\{\psi_\ell\}_{\ell=1}^{d_l}$. Denote the corresponding space of functions of the form $f = \sum_{\ell=1}^{d_l} v_\ell \psi_\ell^l$ by V^l, and notice that $V^1 \subset V^2 \subset \cdots \subset V^l \subset V$. By making Assumption 7 of [53] that the weak solution $p(\cdot; u)$ of (3)-(4) 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 $p(\cdot; u)$, with a rate of convergence in V or L^2, independently of u. Thus, the sequence of densities of interest in this context is:

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

One is also interested in computing Z_L, for instance, to perform model selection or averaging.

Exact sampling of this sequence of posterior distributions is not possible in general, and one must resort to an advanced method such as MCMC, which is introduced in section 4.1, or SMC samplers, which is introduced in section 4.2.1. But it is not obvious how one can leverage the MLMC approach for this application. Several strategies for accomplishing this are suggested later on in the article in section 5.1.

2.2 Partially Observed Diffusions

The following model is considered, as described in [64, 68] for example. We begin with a diffusion process:

$$dU_t \quad = \quad a(U_t)dt + b(U_t)dW_t, \tag{7}$$

with $U_t \in \mathbb{R}^d$, $t \geq 0$, U_0 given, $a : \mathbb{R}^d \to \mathbb{R}^d$ (denote the $j^{th}-$element as $a^j(U_t)$), $b : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ (denote the $j^{th}, k^{th}-$element as $b^{j,k}(U_t)$) and $\{W_t\}_{t\in[0,T]}$ a Brownian motion of $d-$dimensions. Appropriate assumptions are made as in [64, 68] to ensure that the diffusion has a solution; see [64, 68] for details.

The diffusion process itself is not observed directly, but only via some data observed discretely in time. To be precise, it will be assumed that the data are regularly spaced observations $y_{1:n} = (y_1, \ldots, y_n)$, with $y_k \in \mathbb{R}^m$. It will be assumed that conditional on $U_{k\delta} = u_{k\delta}$, for $1 \geq \delta > 0$, Y_k is independent of all other random variables and has probability density $G(u_{k\delta}, y_k)$. For simplicity of notation, let $\delta = 1$ (which can always be done by rescaling time), so $U_k = U_{k\delta}$. The joint probability density of the observations and the unobserved diffusion at the observation times is then

$$\prod_{i=1}^{n} G(u_i, y_i)Q(u_{(i-1)}, u_i),$$

where $Q(u_{(i-1)}, u)$ is the transition density of the diffusion process as a function of u, i.e., the density of the solution U_1 of Eq. (7) at time 1 given initial condition $U_0 = u_{(i-1)}$.

In this problem, one wants to *sequentially* approximate a probability on a fixed space. For $k \in \{1, \ldots, n\}$, the objective is to approximate the filter

$$\pi(u_k|y_{1:k}) = \pi^k(u_k) = \frac{\int_{\mathbb{R}^{(k-1)d}} \prod_{i=1}^{k} G(u_i, y_i)Q(u_{(i-1)}, u_i)du_{1:k-1}}{\int_{\mathbb{R}^{kd}} \prod_{i=1}^{k} G(u_i, y_i)Q(u_{(i-1)}, u_i)du_{1:k}},$$

with $u_{1:k} = (u_1, \ldots, u_k)$ and $y_{1:k} = (y_1, \ldots, y_k)$. The shorthand notation $\pi^k(\cdot) = \pi(\cdot|y_{1:k})$ is used above and in what follows. It is also of interest to estimate the normalizing constant, or marginal likelihood

$$Z = Z^k \int_{\mathbb{R}^{kd}} \prod_{i=1}^{k} G(u_i, y_i)Q(u_{(i-1)}, u_i)du_{1:k}\,.$$

Note that the filtering problem has many applications in engineering, statistics, finance, and physics (e.g., [13, 23, 34] and the references therein)

2.2.1 Approximation

There are several issues associated to the approximation of the filter and marginal likelihood, sequentially in time. The simplest context is when Q is known numerically, in the sense that for any $v \in E$ one can (i) sample $u \sim Q(v, \cdot)$, and (ii) evaluate $Q(v, u)$ (or at least a non-negative unbiased estimator of it). Even in this context, some advanced numerical method such as the particle filter (e.g., [35, 40]) is required for most cases of practical interest. Our context is even more complicated as it is assumed that one can do neither (i) nor (ii). This is because Q must be approximated by some discrete time-stepping method [73] (for time-step $h_l = 2^{-l}$, for simplicity). Note that if (ii) is true, up-to a non-negative unbiased estimator, then particle filters can be adopted directly. In the work [40] and also [41] it is shown how this can be done for a specific class of SDE (7) (although this latter method suffers from the quantum sign problem).

For simplicity and illustration, Euler's method [73] will be considered. One has

$$\begin{aligned} U_k^l(m+1) &= U_k^l(m) + h_l a(U_k^l(m)) + \sqrt{h_l} b(U_k^l(m))\xi_k(m)\,, \\ \xi_k(m) &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_d(0, I_d)\,, \end{aligned} \tag{8}$$

for $m = 0, \ldots, k_l - 1$, where $k_l = 2^l$ and $\mathcal{N}_d(0, I_d)$ is the $d-$dimensional normal distribution with mean zero and covariance the identity (when $d = 1$ we omit the subscript). Here $U_k^l(k_l) = U_k^l$, $U_k^l(0) = U_{k-1}^l = U_{k-1}^l(k_l)$. The numerical scheme gives rise to its own transition density between observation times $Q^l(u_{k-1}^l, u_k^l)$.

Therefore, one wants to approximate for $k \in \{1, \ldots, n\}$ the filter

$$\pi_L(u_k|y_{1:k}) = \frac{\int_{\mathbb{R}^{(k-1)d}} \prod_{i=1}^k G(u_i, y_i) Q^L(u_{(i-1)}, u_i) du_{1:k-1}}{\int_{\mathbb{R}^{kd}} \prod_{i=1}^k G(u_i, y_i) Q^L(u_{(i-1)}, u_i) du_{1:k}}\,,$$

and marginal likelihood

$$Z_L = \int_{\mathbb{R}^{kd}} \prod_{i=1}^k G(u_i, y_i) Q^L(u_{(i-1)}, u_i) du_{1:k}.$$

In section 4.2 we will consider how this task can be performed via the particle filter (an SMC method) and later in section 5.3 how that in turn can be extended to the MLMC context.

2.2.2 Parameter Estimation

Suppose that there is a static parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ in the model, so

$$dU_t = a_\theta(U_t)dt + b_\theta(U_t)dW_t \, ,$$

and G_θ is the likelihood function above. If one assumes a prior ν on θ, then one might be interested in, for k *fixed*:

$$\pi(\theta|y_{1:k}) = \nu(\theta)\frac{\int_{\mathbb{R}^{kd}} \prod_{i=1}^k G_\theta(u_i, y_i)Q_\theta(u_{(i-1)}, u_i)du_{1:k}}{\int_{\mathbb{R}^{kd}\times\Theta} \prod_{i=1}^k G_\theta(u_i, y_i)Q_\theta(u_{(i-1)}, u_i)\nu(\theta)d\theta du_{1:k}}$$

and the associated discretization. In section 4.3 we consider how the latter task is possible using MCMC for a given level, and later in section 5.2 how that is extended for the MLMC case.

3 Monte Carlo and Multilevel Monte Carlo

For now, let us assume that only 1. & 2. of Section 1 paragraph 2 apply, and 3. does not apply. In this context, one can obtain i.i.d. samples $U^1, \ldots, U^N \sim \pi_L$ and use the standard Monte Carlo estimator $\frac{1}{N}\sum_{i=1}^N \varphi(u^i) \approx \mathbb{E}_{\pi_L}[\varphi(U)] = \int_{\mathsf{E}} \varphi(u)\pi_L(u)du$. To explain what will follow, we will suppose the following.

Assumption 3.1 (Cost and discretization error)**.** *There are $\alpha, \zeta > 0$ such that*

- *The cost of simulating one sample $U \sim \pi_L$ is $\mathcal{O}(h_L^{-\zeta})$.*
- *The bias is $|\int_{\mathsf{E}} \varphi(u)\pi(u)du - \int_{\mathsf{E}} \varphi(u)\pi_L(u)du| = \mathcal{O}(h_L^\alpha)$.*

The terms ζ and α parameterize the cost and bias in Assumption 3.1, in terms of an accuracy parameter h. This parameterization is adopted because it often occurs in practical examples. Roughly speaking, the accuracy is expected to be proportional to h and the cost is expected to be proportional to the number of degrees of freedom h^{-1}. For instance, consider the SDE (7) with appropriate regularity conditions (as mentioned below (7)) and an Euler discretization with time-step h, as in (8). If the target π_L is the marginal at time 1, then for appropriate φ one has that Assumption 3.1 holds with $\alpha = \zeta = 1$.

For simplicity suppose, $h_L = 2^{-L}$. Consider the mean square error (MSE) associated to the Monte Carlo estimator

$$\mathbb{E}\Big[\Big(\frac{1}{N}\sum_{i=1}^{N}\varphi(U^i) - \mathbb{E}_\pi[\varphi(U)]\Big)^2\Big],$$

where $\mathbb{E}[\cdot]$ is the expectation operator w.r.t. the distribution of the samples $(U^1,\ldots,U^N)$. Note that $\mathbb{E}[\varphi(U^i)] = \mathbb{E}_{\pi_L}[\varphi(U)]$. Adding and subtracting $\mathbb{E}_{\pi_L}[\varphi(U)]$, and assuming φ has a second moment w.r.t. π_L, one has that the MSE is equal to

$$\frac{1}{N}\mathbb{V}\mathrm{ar}_{\pi_L}[\varphi(U)] + (\mathbb{E}_{\pi_L}[\varphi(U)] - \mathbb{E}_\pi[\varphi(U)])^2,$$

which is the standard variance plus bias squared. Now let $1 > \epsilon > 0$ be given, and suppose one wants to control the MSE, so that it is $\mathcal{O}(\epsilon^2)$. One begins by controlling the bias, by setting L. The constraint that $2^{-2\alpha L} = \mathcal{O}(\epsilon^2)$ can be satisfied by choosing $L \propto -\frac{\log(\epsilon)}{\alpha\log(2)}$. Then the constraint that the variance is $\mathcal{O}(\epsilon^2)$ can be satisfied by choosing $N \propto \epsilon^{-2}$. The cost can then be controlled by $2^{\zeta L}\epsilon^{-2} = \mathcal{O}(\epsilon^{-2-\frac{\zeta}{\alpha}})$. In the case of an Euler discretization of a sufficiently regular SDE, one can asymptotically obtain an MSE of $\mathcal{O}(\epsilon^2)$ for a cost of $\mathcal{O}(\epsilon^{-3})$.

The multilevel Monte Carlo method is designed to improve over the cost of Monte Carlo. As above, suppose $h_l = 2^{-l}$. The idea is to consider a hierarchy, $\infty > h_1 > \cdots > h_L > 0$ and consider the respresentation

$$\mathbb{E}_{\pi_L}[\varphi(U)] = \sum_{l=1}^{L}\{\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}\}[\varphi(U)], \tag{9}$$

where for $1 \leq l \leq L$, $\mathbb{E}_{\pi_l}$ is the expectation w.r.t. π_l (i.e., the biased approximation with parameter h_l) and for $l = 1$, $\mathbb{E}_{\pi_{l-1}}[\varphi(U)] := 0$. This will be referred to in what follows as the ML identity. Here, it is assumed that for each probability π_l one is only interested in a marginal on E, even if the entire space must be enlarged to facilitate the biased approximation. So, for instance, π_l may be defined on a larger space than E, but it admits a marginal on E which approaches π as l grows. For simplicity of exposition in this section, we will assume that π_l is a probability on E, $1 \leq l \leq L$. To approximate the first term in the summation of (9), one samples $U^{1,1}, \ldots, U^{1,N_1}$ i.i.d. from π_1 and one uses the standard Monte Carlo estimator $\frac{1}{N}\sum_{i=1}^{N} \varphi(u^{1,i})$. Note that the notation $U^{1,i} \in \mathsf{E}$, has superscript 1 to denote one is sampling from π_1 and $i \in \{1, \ldots, N_1\}$ is the sample counter. For the remainder of the terms $2 \leq l \leq L$, we suppose that it is possible to sample a (dependent) *coupling* of (π_l, π_{l-1}). That is, a pair of random variables $\check{U}^l = (\overline{U}^l, \underline{U}^l) \in \mathsf{E} \times \mathsf{E}$ such that Assumption 3.2 below holds. The notation $(\overline{U}^l, \underline{U}^l)$ is used in the following way. The superscript l denotes that one is considering a coupling of (π_l, π_{l-1}) and that $\overline{U}^l \sim \pi_l$ and $\underline{U}^l \sim \pi_{l-1}$.

Assumption 3.2 (Variance)**.** *There is a $\beta > 0$ such that the variance w.r.t. the coupling of (π_l, π_{l-1}),*

$$\mathbb{V}\mathrm{ar}_{(\pi_l,\pi_{l-1})}[\varphi(\overline{U}^l) - \varphi(\underline{U}^l)] = \mathcal{O}(h_l^{\beta}).$$

The main point of this assumption is that ensures that, if satisfied, one has constructed a coupling of (π_l, π_{l-1}) which can lead to a variance reduction in estimation, relative to sampling from π_L. For instance, one can always construct a coupling of (π_l, π_{l-1}) such as the independent one, but this will seldom (if ever) provide any improvement, in terms of variance, over sampling from π_L. In essense, one requires the coupling to be 'good enough' which is what this assumption tries to encapsulate. Couplings which satisfy Assumption 3.2 exist, for instance, in the context of SDEs. Again consider the SDE (7) with appropriate regularity conditions, an Euler discretization as in (8), and let π_l be the marginal at time 1 approximated with step-size h_l. Then one can construct a coupling so that $\beta = 1$, or if $b(U_t)$ in (7) is a constant function, then $\beta = 2$. In order to approximate the summands in (9), independently, for $2 \leq l \leq L$, draw N_l i.i.d. samples $(\overline{U}^{l,1}, \underline{U}^{l,1}), \ldots, (\overline{U}^{l,N_l}, \underline{U}^{l,N_l})$ from the coupling (π_l, π_{l-1}), and use the following unbiased estimator of $\{\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}\}[\varphi(U)]$:

$$\frac{1}{N_l}\sum_{i=1}^{N_l}\{\varphi(\overline{u}^{l,i}) - \varphi(\underline{u}^{l,i})\}\,.$$

The multilevel estimator is thus

$$\frac{1}{N_1}\sum_{i=1}^{N_1}\varphi(u^{1,i}) + \sum_{l=2}^{L}\Big(\frac{1}{N_l}\sum_{i=1}^{N_l}\{\varphi(\overline{u}^{l,i}) - \varphi(\underline{u}^{l,i})\}\Big).$$

The following construction is now, given $\epsilon > 0$ arbitrary, to select L, the number of samples $N_1, \dots, N_L$ so that the MSE is $\mathcal{O}(\epsilon^2)$ and to minimize the cost ('optimal cost') as a function of ϵ. The procedure is termed a 'general MLMC theorem' and is the scheme which one would like to replicate in other scenarios. One can analyze the MSE as above. It is equal to

$$\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\pi_1}[\varphi(U)] + \sum_{l=2}^{L}\frac{1}{N_l}\mathbb{V}\mathrm{ar}_{(\pi_l,\pi_{l-1})}[\varphi(\overline{U}^l) - \varphi(\underline{U}^l)] + (\mathbb{E}_{\pi_L}[\varphi(U)] - \mathbb{E}_{\pi}[\varphi(U)])^2\,,$$

and the associated cost is $\sum_{l=1} N_l h_l^{-\zeta}$, where we assume that the cost of sampling the coupling (π_l, π_{l-1}) is at most the cost of sampling π_l. Since we have assumed $h_1 = \mathcal{O}(1)$, then $\frac{1}{N_1}\mathbb{V}\mathrm{ar}_{\pi_1}[\varphi(U)] \leq \frac{C}{N_1}$, for $\infty > C > 0$ a constant independent of l. Now let $1 > \epsilon > 0$ be given, and suppose one wants to control the MSE, so that it is $\mathcal{O}(\epsilon^2)$. One controls the bias as above by letting

$$L \propto -\frac{\log(\epsilon)}{\alpha\log(2)}. \tag{10}$$

Then one seeks to minimize the cost $\sum_{l=1} N_l h_l^{-\zeta}$ in terms of $N_1, \dots, N_L$, subject to the constraint

$$\sum_{l=1}^{L}\frac{h_l^{\beta}}{N_l} = \mathcal{O}(\epsilon^2).$$

This constrained optimization problem is solved in e.g. [21] and has the solution $N_l \propto h_l^{(\beta+\zeta)/2}$ to obtain a MSE of $\mathcal{O}(\epsilon^2)$. Solving for the Lagrange multiplier, with equality above, one has that

$$N_l = \epsilon^{-2}h_l^{(\beta+\zeta)/2}K_L\,, \tag{11}$$

where $K_L = \sum_{l=1}^{L} h_l^{(\beta-\zeta)/2}$. Note that K_L may depend upon L, depending on the values of β, ζ. In the Euler case $\beta = \zeta$. So, one is able to obtain an MSE of $\mathcal{O}(\epsilon^2)$ for the cost $\mathcal{O}(\epsilon^{-2}\log(\epsilon)^2)$. In the special case in which the diffusion coefficient is constant, one obtains the Milstein method with $\beta > \zeta$, so the cost can be controlled by $\mathcal{O}(\epsilon^{-2})$.

The MLMC framework discussed above is considered in various different guises in this paper. The cost will always scale as in Assumption 3.1, and some analogue of the bias from Assumption 3.1 will determine L as defined in (10). We will then require rates on different quantities analogous to Assumption 3.2, in order to ensure the choice of N_l in (11) is optimal.

We stress here that we consider primarily expectations of the type $\mathbb{E}_{\pi_L}[\varphi(U)]$, where φ does not depend upon L, consistent with the applications of Section 2. For an appropriate class of functions φ, the extension of many of the results to the context where one considers $\mathbb{E}_{\pi_L}[\varphi_L(U)]$ can require very little extra (mathematical) analysis. See for instance [7, Proposition 3.1], which is an extension of [9, Theorem 3.1] in a similar context. In some cases, it may require additional work (see Section 5.1.1).

In the present work (with the exception of Section 5.1.1) we assume φ does not depend on L, however we note that this can be of importance in many application areas.

4 Some Computational Methods

The following section gives a basic introduction to some computational methods that can be used for the examples of the previous section. The review is quite basic, in the sense that there are numerous extensions in the literature, but, we try to provide pointers to the literature, where relevant. We first consider basic MCMC in Section 4.1. Then, in Section 4.2, SMC is discussed in context of filtering and the resulting algorithm is often termed a *particle filter*. Then, SMC samplers (e.g., [27]) is reviewed, which is used for different problems and uses MCMC algorithms within it. We remark that particle filters and SMC samplers are simply both examples of the *same* algorithm, but used in different contexts with different interpretations. Hence they are both SMC algorithms. In Section 4.3, we discuss particle MCMC [1], which uses SMC within MCMC proposals. Finally, in Section 4.4, we discuss ensemble Kalman filter approaches.

4.1 Markov chain Monte Carlo

We consider a target probability π_L on space $\mathsf{E} = \mathbb{R}^d$. The idea of MCMC is to build an ergodic Markov chain of invariant distribution π_L. That is, samples of the Markov chain

$U^1, \ldots, U^N$ have the property that for $\varphi : \mathsf{E} \to \mathbb{R}$, $\pi-$integrable, estimates of the form

$$\frac{1}{N}\sum_{i=1}^{N}\varphi(u^i)$$

will converge almost surely as $N \to \infty$ to $\mathbb{E}_{\pi_L}[\varphi(U)]$. A generalization of this case is when π_L is a marginal of the invariant distribution of the Markov chain; this case is relevant for the particle MCMC, but in this section we will concentrate on the standard case.

There are many ways to produce the Markov chain. Here we will only describe the standard random walk Metropolis-Hastings algorithm. At the end of the description, references for more recent work are given. Suppose at time i of the algorithm, u^i is the current state of the Markov chain. A new candidate state $U'|u^i$ is proposed according to

$$U' = u^i + Z,$$

where $Z \sim \mathcal{N}_d(0, \Sigma)$ independently of all other random variables. The proposed value is accepted ($u^{i+1} = u'$) with probability:

$$\min\left\{1, \frac{\pi_L(u')}{\pi_L(u^i)}\right\}$$

otherwise it is rejected and $u^{i+1} = u^i$. The scaling of the proposal, i.e., the proposal covariance Σ, is often chosen so that the acceptance rate is about 0.234 [93], although there are adaptive methods for doing this; see [2, 49].

The algorithm mentioned here is the most simple approach. The ideas can be extended to alternative proposals, such as Langevin [91], Hamiltonian Monte Carlo ([37], see also [83]), pre-conditioned Crank Nicholson (e.g., [22] and the references therein). The algorithm can also be used in infinite dimensions (e.g., Hilbert spaces [22]) and each dimension need not be updated simultaneously - for example, one can use Gibbs and Metropolis-within-Gibbs approaches; see for example [89, 92]. There are also population-based methods (e.g., [60] and the references therein) and non-reversible approaches (e.g., [85, 86, 12] and the references therein). Even this list is not complete: the literature is so vast, that one would require a book length introduction, which is well out of the scope of this work - the reader is directed to the previous papers and the references therein for more information. There is also a

Algorithm 1 The Particle Filter

- **Initialize**. Set $k = 1$, for $i \in \{1, \ldots, N\}$ sample u_1^i from q_1 and evaluate the weight

$$w_1^i = \Big(\frac{G(u_1^i, y_1)Q^L(u_0, u_1^i)}{q_1(u_1^i)}\Big)\Big(\sum_{j=1}^{N}\frac{G(u_1^j, y_1)Q^L(u_0, u_1^j)}{q_1(u_1^j)}\Big)^{-1}$$

- **Iterate**: Set $k = k + 1$,
  - Resample $(\hat{u}_{k-1}^1, \ldots, \hat{u}_{k-1}^N)$ according to the weights $(w_{k-1}^1, \ldots, w_{k-1}^N)$.
  - Sample $u_k^i|\hat{u}_{k-1}^i$ from q_k, for $i \in \{1, \ldots, N\}$, and evaluate the weight

$$w_k^i = \Big(\frac{G(u_k^i, y_k)Q^L(\hat{u}_{k-1}^i, u_k^i)}{q_k(\hat{u}_{k-1}^i, u_k^i)}\Big)\Big(\sum_{j=1}^{N}\frac{G(u_k^j, y_k)Q^L(\hat{u}_{k-1}^j, u_k^j)}{q_k(\hat{u}_{k-1}^j, u_k^j)}\Big)^{-1}.$$

well established convergence theory; see [81, 92] for information. We remark also that the cost of such MCMC algorithms can be quite reasonable if d is large, often of polynomial order (e.g., [93]) and can be dimension free when there is a well-defined limit as d grows [10, 22]. Finally, note that it is not simple to use 'standard' MCMC to estimate normalizing constants.

4.2 Sequential Monte Carlo

Here we will consider the particle filter. To assist our discussion, we focus on the filtering density from Section 2.2.1

$$\pi_L^k(u_k) = \pi_L(u_k|y_{1:k}) \propto \int_{\mathbb{R}^{(k-1)d}} \prod_{i=1}^{k} G(u_i, y_i)Q^L(u_{(i-1)}, u_i)du_{1:k-1}\,. \tag{12}$$

Here there are no static parameters to be estimated. To facilitate the discussion, we will suppose that Q^L can be evaluated pointwise, although of course, it is generally not the case in our application. Moreover, it will be assumed that $G(u_i, y_i)Q^L(u_{(i-1)}, u_i) > 0$ for each $i \geq 1$, u_{i-1}, u_i.

We suppose we have access to a collection of proposals $q_1(u_1), q_2(u_1, u_2), q_3(u_2, u_3), \ldots,$ where $q_j(u_{j-1}, u_j)$ is a positive probability density in u_j, for each value of u_{j-1}. The particle filter is then given in Algorithm 1.

The resampling step can be performed using a variety of schemes, such as systematic, multinomial, residual etc; the reader is referred to [35] for more details. For $\varphi : \mathbb{R}^d \to \mathbb{R}$, and φ π_L^k-integrable, one has the consistent estimate of $\mathbb{E}_{\pi_L^k}[\varphi(U_k)]$:

$$\sum_{i=1}^{N} w_k^i \varphi(u_k^i). \tag{13}$$

In addition, the marginal likelihood Z_L^k is unbiasedly [25] estimated by

$$\widehat{Z}_L^k := \prod_{i=1}^{k} \Big(\frac{1}{N} \sum_{j=1}^{N} \frac{G(u_i^j, y_i) Q^L(\hat{u}_{i-1}^j, u_i^j)}{q_i(\hat{u}_{i-1}^j, u_i^j)} \Big), \tag{14}$$

with the abuse of notation that $\hat{u}_0^i = u_0$. In principle, for $\varphi : \mathbb{R}^{kd} \to \mathbb{R}$, and φ π_L^k-integrable, one could also try to estimate $\mathbb{E}_{\pi_L^k}[\varphi(U_{1:k})]$ but this does not work well in practice due to the well-known path degeneracy problem; see [35, 72].

The algorithm given here is one of the most basic and many modifications can enhance the performance of this algorithm; see [25, 26, 35, 72] for some ideas. The theoretical validity of the method has been established in many works; see e.g., [18, 25, 26, 33]. The algorithm performs very well w.r.t. the time parameter k. Indeed L_p-errors for estimates such as (13) are $\mathcal{O}(N^{-1/2})$ where the constant is independent of time and the relative variance of (14) is $\mathcal{O}(k/N)$ (if $N > Ck$ for C some constant independent of k); see [26] and the references therein.

One of the main issues with particle filters/SMC methods in this context is that they do not always perform well in high dimensions (i.e., for large d) often having an exponential cost in d [102]. There have been some methods developed for high-dimensional filtering (e.g., [8, 82, 87]), however they are only useful for a small class of models. In addition, if there a well-defined limit of the model (in some sense) as d grows, useful particle filters can be developed (see e.g., [71, 78]). As noted in [14], the key criterion which needs to be satisfied is that the target distribution

$$\frac{\prod_{i=1}^{k} G(u_i, y_i) Q^L(u_{(i-1)}, u_i)}{\int \prod_{i=1}^{k} G(u_i, y_i) Q^L(u_{(i-1)}, u_i) du_{1:k}}$$

and the importance distribution

$$q_1(u_1) \prod_{i=2}^{k} q_i(u_{i-1}, u_i)$$

do not become mutually singular in the limit $d \to \infty$ (or equivalently that the symmetrized Kullback-Liebler distance between these distributions does not explode with d).

```
**Algorithm 2** SMC Sampler
• Initialize. Set $l = 1$, for $i \in \{1, \dots, N\}$ sample $u_1^i$ from $\pi_1$.

• Iterate: Set $l = l + 1$. If $l = L + 1$ stop.

    – Resample $(\hat{u}_{l-1}^1, \dots, \hat{u}_{l-1}^N)$ using the weights $(w_l^1, \dots, w_l^N)$ where, for $i \in \{1, \dots, N\}$,
          $w_l^i = \Big(\frac{\kappa_l(u_{l-1}^i)}{\kappa_{l-1}(u_{l-1}^i)}\Big)\Big(\sum_{j=1}^N \frac{\kappa_l(u_{l-1}^j)}{\kappa_{l-1}(u_{l-1}^j)}\Big)^{-1}.$
    – Sample $u_l^i | \hat{u}_{l-1}^i$ from $M_l$ for $i \in \{1, \dots, N\}$.
```

4.2.1 Sequential Monte Carlo Samplers

Consider a sequence of distributions $\pi_1, \dots, \pi_L$ on a common measurable space. In addition to this suppose we have Markov kernels $M_2, \dots, M_L$ of invariant densities $\pi_2, \dots, \pi_L$. This is possible if the densities are known up-to a constant, simply by using using MCMC. The SMC sampler algorithm (e.g., [27]) can be used to approximate expectations w.r.t. $\pi_1, \dots, \pi_L$, as well as to estimate ratios of normalizing constants. The un-normalized densities of $\pi_1, \dots, \pi_L$ are written $\kappa_1, \dots, \kappa_L$. To ease the notational burden, we suppose one can sample from π_1, but this is not necessary. The algorithm is given in Algorithm 2.

One can estimate expectations w.r.t. π_l, for $\varphi : \mathsf{E} \to \mathbb{R}$, π_l-integrable. The following estimator of $\mathbb{E}_{\pi_l}[\varphi(U)]$ converges almost surely as $N \to \infty$:

$$\frac{1}{N}\sum_{i=1}^N \varphi(u_l^i)\,.$$

In addition, for any $l \geq 2$, we have the unbiased estimator of Z_k/Z_1

$$\prod_{l=2}^k \Big(\frac{1}{N}\sum_{i=1}^N \frac{\kappa_l(u_{l-1}^i)}{\kappa_{l-1}(u_{l-1}^i)}\Big)$$

which converges almost surely as $N \to \infty$.

The basic algorithm goes back to at least [57]. Several versions are found in [17, 84], with a unifying framework in [27] and a rediscovery in [16]. Subsequently, several refined and improved versions of the algorithm have appeared [20, 28, 52, 67, 95], including those which allow algorithmic parameters to be set adaptively, that is, without user specification.

When $\mathsf{E} = \mathbb{R}^d$, this method indeed performs quite well w.r.t. the dimension d with only polynomial cost in d, in contrast to particle filters; see [4, 5]. Whilst the underlying theory for this algorithm is very similar to particle filters and it is covered in [25, 26], there are some additional results in [6, 98, 106]. In particular, [6] establish that when one updates parameters adaptively, such as in [67, 95], then the algorithm is still theoretically correct. The method is very useful in the following scenaria: (i) if one wishes to compute ratios of normalizing constants, (ii) the available MCMC kernels do not mix particularly well, and/or (iii) the target is multimodal and the modes are separated by regions of very low probability.

4.3 Particle Markov chain Monte Carlo

We now consider the scenario of Section 2.2.2. In this context, the standard approach is to consider the extended target with density

$$\pi_L^k(\theta, u_{1:k}) \propto \nu(\theta) \prod_{i=1}^{k} G_\theta(u_i, y_i) Q_\theta^L(u_{(i-1)}, u_i).$$

Sampling this distribution is notoriously challenging. The particle marginal Metropolis-Hastings (PMMH) algorithm of [1] can be considered the gold standard for solving this problem. It is given in Algorithm 3. Notice that within Algorithm 3 one employs a particle filter, as introduced in Section 4.2. Any suitable proposal q_k can be used in Algorithm 1, as long as it provides well-defined importance ratios. The simplest option is to use the forward (Euler discretized) dynamics, and that will be used here. In Algorithm 3, one requires a proposal density on the space of the parameter θ. We denote by $r(\theta, \theta')$ the conditional density of $\theta' \in \Theta$ given θ. This density is selected by the user such that one can sample from the associated distribution and evaluate $r(\theta, \theta')$ for every $(\theta, \theta') \in \Theta^2$. In the particle filter that is run as part of the proposal mechanism in Algorithm 3, it is remarked that the associated proposals (transitions of the dynamics) Q_θ^L and likelihoods G_θ depend upon θ. Hence one first proposes a θ from r and then runs the particle filter with the proposed θ.

The samples of Algorithm 3 can be used to estimate expectations such as $\int_\Theta \varphi(\theta) \pi_L(\theta|y_{1:k}) d\theta$ with

$$\frac{1}{N} \sum_{i=1}^{N} \varphi(\theta^i) \, .$$

Algorithm 3 A Particle MCMC Algorithm

```
• Initialize. Set $i = 0$ and sample $\theta^0$ from the prior. Given $\theta^0$ run the particle filter in
  Algorithm 1 and record the estimate of $\widehat{Z}^k_{L,\theta^0}$ from eq. (14).
• Iterate:
    – Set $i = i + 1$ and propose $\theta'$ given $\theta^{i-1}$ from a proposal $r(\theta^{i-1}, \cdot)$ ($r$ is explained
      in the main text).
    – Given $\theta'$ run the particle filter in Algorithm 1 and record the estimate $\widehat{Z}^k_{L,\theta'}$ (as
      in (14)).
    – Set $\theta^i = \theta'$ with probability
          $\min\left\{1, \frac{\widehat{Z}^k_{L,\theta'}\pi_\theta(\theta')r(\theta', \theta^{i-1})}{\widehat{Z}^k_{L,\theta^{i-1}}\pi_\theta(\theta^{i-1})r(\theta^{i-1}, \theta')}\right\}$
      otherwise $\theta^i = \theta^{i-1}$.
```

This works because one is effectively implementing a standard Metropolis-Hastings MCMC on an extended state-space, for which the marginal on θ is correct [1, 3]. Under minimal conditions this estimator converges almost surely as $N \to \infty$. The algorithm can also be extended to allow estimation of the hidden states $u_{1:k}$ as well. There are many parameters of the algorithm, such as the number of samples of the SMC algorithm, and tuning of these parameters has been discussed in [1, 36].

The PMMH algorithm is the most basic in [1]. Several enhancements are in [1] and numerous algorithms that improve upon this method can be found in [31, 101].

4.4 Ensemble Kalman filter

Similarly to the particle filter, the idea of the ensemble Kalman filter (EnKF) is to approximate the filtering distribution (12) using an ensemble of particles [38]. As such they are sometimes also referred to as sequential Monte Carlo methods, but it is important to distinguish them from the methods described in subsection 4.2, as will become clear from the description below. It will be assumed here for simplicity that the observation selection function is given by:

$$G(u_i, y_i) \propto \exp(-\frac{1}{2}|\Gamma^{-\frac{1}{2}}(Hu_i - y_i)|^2)\,, \tag{15}$$

where H is a linear operator and Γ is an invertible and symmetric linear operator.

We briefly recall the Kalman filter [70], which recovers the filtering distribution if the dynamics of the observations and hidden state are linear and Gaussian, i.e.

$$u_{k+1} = A^L u_k + \xi_{k+1}, \quad \xi_{k+1} \sim \mathcal{N}(0, \Sigma) \;\; \text{i.i.d.}.$$

It is easy to see that if $\pi_L^k(u_k) = \nu(u_k; \widehat{m}_k, \widehat{C}_k)$, where $\nu(u; m, C)$ is the Gaussian density of u with mean m and covariance C, then $\pi_L(u_{k+1}|y_1, \ldots, y_k) = \nu(u_{k+1}; A^L\widehat{m}_k, A^L\widehat{C}_k(A^L)^T + \Sigma)$. Furthermore, defining $m_{k+1} = A^L\widehat{m}_k$ and $C_{k+1} = A^L\widehat{C}_k(A^L)^T + \Sigma$, it is easy to calculate the updated mean and covariance $\widehat{m}_{k+1} = (I_d - K_{k+1}H)m_{k+1} + K_{k+1}y_{k+1}$ and $\widehat{C}_{k+1} = (I_d - K_{k+1}H)C_{k+1}$, where $K_{k+1} = C_{k+1}H^T(HC_{k+1}H^T + \Gamma)^{-1}$ (see for example [75]).

The EnKF consists of using the sample covariance of the ensemble of particles to perform an update on the particles such that the sample mean and covariance coincide with that which arises from an exact Kalman update of the approximate finite-sample predicted Gaussian. In other words, the observations are incorporated as though the process were linear and Gaussian. Hence the method is consistent for the linear Gaussian scenario described above [75]. In fact, it is only consistent in this scenario. Nonetheless it provides an estimate of the state and error in general, and can be tuned to perform reasonably well in tracking and forecasting. Additionally, it is robust even in high dimensions [39]. It has therefore become very popular among practitioners.

The EnKF is executed in a variety of ways and only one will be considered here, the *perturbed observation* EnKF:

$$\text{Prediction} \quad \begin{cases} u_{k+1}^{(n)} & \sim Q^L(\widehat{u}_k^{(n)}, \cdot\,)\,,\; n = 1, ..., N, \\ m_{k+1}^N & = \frac{1}{N}\sum_{n=1}^N u_{k+1}^{(n)}\,, \\ C_{k+1}^N & = \frac{1}{N-1}\sum_{n=1}^N (u_{k+1}^{(n)} - m_{k+1}^N)(u_{k+1}^{(n)} - m_{k+1}^N)^T\,. \end{cases}$$

$$\text{Analysis}\quad\begin{cases} S_{k+1} &= HC_{k+1}^N H^T + \Gamma, \\ K_{k+1} &= C_{k+1}^N H^T S_{k+1}^{-1}, \\ \widehat{u}_{k+1}^{(n)} &= (I - K_{k+1}H)u_{k+1}^{(n)} + K_{k+1}y_{k+1}^{(n)},\ n = 1, ..., N, \\ y_{k+1}^{(n)} &= y_{k+1} + \xi_{k+1}^{(n)},\ n = 1, ..., N. \end{cases}$$

Here $\xi_k^{(n)}$ are i.i.d. draws from $\mathcal{N}(0, \Gamma)$. Perturbed observation refers to the fact that each particle sees an observation perturbed by an independent draw from $\mathcal{N}(0, \Gamma)$. This procedure ensures that in the linear Gaussian case described above the Kalman Filter is obtained in the limit of infinite ensemble [75]. In other words, if we define

$$\widehat{m}_{k+1}^N = \frac{1}{N}\sum_{n=1}^N \widehat{u}_{k+1}^{(n)}, \qquad \widehat{C}_{k+1}^N = \frac{1}{N-1}\sum_{n=1}^N (\widehat{u}_{k+1}^{(n)} - \widehat{m}_{k+1}^N)(\widehat{u}_{k+1}^{(n)} - \widehat{m}_{k+1}^N)^T\,,$$

then the distribution $\mathcal{N}(\widehat{m}_{k+1}^N, \widehat{C}_{k+1}^N)$ converges to the Gaussian filtering distribution $\mathcal{N}(\widehat{m}_{k+1}, \widehat{C}_{k+1})$ in the limit as $N \to \infty$ almost surely [79], and one can derive rates of convergence as well [77]. Notice that the ensemble is not prescribed to be Gaussian, even though it is updated as though it were. So in the general nonlinear and/or non-Gaussian case the limiting target is some non-Gaussian $\widehat{\pi}_L^k$, which is in general not equal to the density π_L^k defined by (12) (see e.g., [77, 76]).

5 Approaches for MLMC Estimation

We now consider various ways in which the MLMC method can be used in these challenging situations, where it is non-trivial to construct couplings of the targets. The codes for these examples are available from the authors. It should be noted that we have written (in collaboration with past co-authors) long and detailed bespoke codes (in C++) for the examples and have not used any generic packages, such as in R. It remains an open research area to create a computer package which can run all of these techniques as have been implemented in this article.

5.1 Importance Sampling

In this case, we investigate the ML identity where the sequence of targets $\pi_1, \dots, \pi_L$ are defined on a common space and are known up-to a normalizing constant; i.e., $\pi_l(u) = \kappa_l(u)/Z_l$ as in Section 2.1 and 4.2.1.

In this scenario [9] (see also [7, 29, 30, 59]) investigate the simple modification

$$\begin{aligned}
\mathbb{E}_{\pi_L}[\varphi(U)] &= \sum_{l=1}^{L} \{\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}\}[\varphi(U)] \\
&= \mathbb{E}_{\pi_1}[\varphi(U)] + \sum_{l=2}^{L} \mathbb{E}_{\pi_{l-1}}\Big[\Big(\frac{\kappa_l(U)Z_{l-1}}{\kappa_{l-1}(U)Z_l} - 1\Big)\varphi(U)\Big]. \qquad (16)
\end{aligned}$$

The idea here is simple. If one does not know how to construct a coupling of the targets, then one replaces coupling by importance sampling. The key point is that the targets π_l and π_{l-1} are very closely related by construction, and therefore the change of measure formula above should facilitate an importance sampling procedure that *performs well.* Just as for standard MLMC (for instance as described in Section 3) where the coupling has to be good enough, the change of measure needs to be chosen appropriately to ensure that this approach can work well. Recall from Section 4.2.1 that SMC samplers can be designed to sequentially approximate $\pi_1, \dots, \pi_L$, and the ratios Z_l/Z_{l-1}. Therefore the change of measure in (16) is very natural here.

The approach in [9] is to simply run the algorithm of Section 4.2.1, except at step l one resamples $N_{l+1} < N_l$ particles, where the schedule of numbers $N_{1:L-1}$ is chosen using a similar principle as for standard MLMC. The identity (16) can be approximated via:

$$\sum_{l=3}^{L} \Big\{ \frac{\sum_{i=1}^{N_{l-1}} \varphi(u_{l-1}^i) \frac{\kappa_l(u_{l-1}^i)}{\kappa_{l-1}(u_{l-1}^i)}}{\sum_{i=1}^{N_{l-1}} \frac{\kappa_l(u_{l-1}^i)}{\kappa_{l-1}(u_{l-1}^i)}} - \frac{1}{N_{l-1}} \sum_{i=1}^{N_1} \varphi(u_{l-1}^i) \Big\} + \frac{\sum_{i=1}^{N_1} \varphi(u_1^i) \frac{\kappa_2(u_1^i)}{\kappa_1(u_1^i)}}{\sum_{i=1}^{N_1} \frac{\kappa_2(u_1^i)}{\kappa_1(u_1^i)}}. \qquad (17)$$

Note that the algorithm need only be run up-to level $L-1$. [9] not only show that this is consistent, but also give a general MLMC theorem using the theory in [25] with some additional work and assumptions (which are relaxed in [30]). In the context of the example of Section 2.1, [9] show that the work to compute expectations relative to standard SMC samplers (as in Section 4.2.1) is reduced to achieve a given MSE. It is noted that, under appropriate assumptions, the MSE of the estimate (17), for the problem in Section 2.1, is upper-bounded by (roughly, see [9, eq. (12)] for the exact expression)

$$\frac{1}{N_1} + \sum_{l=3}^{L} \frac{h_l^{\beta}}{N_l} \, ,$$

where h_l is the finite element mesh diameter. Therefore error does not accumulate as l grows.

The main reason why this approach can work well, can be explained by terms that look like

$$\frac{\kappa_l(U)Z_{l-1}}{\kappa_{l-1}(U)Z_l} - 1.$$

In the context of the problem in Section 2.1, this term will tend to zero at a rate h_l^{β}, under suitable assumptions and in an appropriate norm, just as the variance terms in the coupling of Section 3. In more standard importance sampling language, the weight tends to one as the sequence of target distributions gets more precise. This particular approach exploits this property, and the success of the method is dependent upon it. In particular, the key quantities for which one needs to obtain rates α and β for are summarized in Table 1.

To illustrate the MLSMC sampler, we consider the example in Section 2.1. A 1D version of the elliptic PDE problem is considered. Let $\Omega = [0, 1]$ 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 functional of interest is taken as the solution of the forward problem at the midpoint of the domain, that is $p(0.5; u)$. 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 $\Gamma = 0.25^2 I$ with I the identity matrix. The data are generated from the model with these settings (under a numerical finite elements solution with $h = 2^{-20}$).

A plot of log cost against log MSE for MLSMC and SMC samplers (Algorithm 2) is given in Figure 1. The plots indicate that for a MSE of $\mathcal{O}(\epsilon^2)$, MLSMC costs $\mathcal{O}(\epsilon^{-2})$ and SMC costs $\mathcal{O}(\epsilon^{-3})$. SMC samplers use all the same settings except N is the same at each level, and only the samples at level L are used to construct the estimator. N is chosen to achieve the same MSE as the MLSMC algorithm. The graph corresponds to different choices of L,

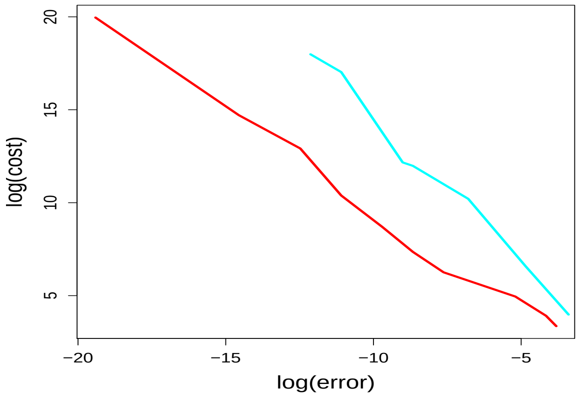


Figure 1: Log Computational cost against log mean squared error for a MLSMC sampler in comparison to a SMC sampler for the Bayesian inverse problem from subsection 2.1 [9]. The MLSMC is the red line.

ranging from $L = 1$ to $L = 10$ for MLSMC and $L = 1$ to $L = 7$ for SMC samplers. Exact details on implementation can be found in [9, Section 5.2].

Rate parameter	Relevant quantity
α	$(\mathbb{E}_{\pi_L} - \mathbb{E}_{\pi})(\varphi)$
β	$\sup_{u\in\mathsf{E}} \left\lvert \frac{\kappa_l(u)Z_{l-1}}{\kappa_{l-1}(u)Z_l} - 1 \right\rvert$

Table 1: The key rates of convergence required for MLSMC samplers.

The method of [9] has been extended to the computation of normalizing constants [30] and has been applied to other examples, such as non-local equations [59] and approximate Bayesian computation [69]. The method has also been extended to the case that the accuracy of the approximation improves as the dimension of the target grows, i.e. infinite dimensional E; see [7]. Another extension is in [74] where the authors use a temperiing sequence along with the level of discretization. Several well known methods such as adaptation are incorporated and, as one might expect, empirical improvements can be seen over just using the levels of discretization (see [74, Figure 5.11]). The idea of interpolating tempering steps has been used in many places, such as in [71] (see also the references therein).

5.1.1 Other instances of importance sampling

The importance sampling idea has also been considered in other articles such as [53, 99].

First we provide a simplified exposition of the seminal work [53], which is the first work to apply the MLMC identity in the context of MCMC. Consider a sequence of approximations of a target π on space E, $\pi_1, \ldots, \pi_L$, exactly as in the above exposition on importance sampling. We also suppose that one can only evaluate a discretized version of the function $\varphi : \mathsf{E} \to \mathbb{R}$, denoted by $\varphi_l : \mathsf{E} \to \mathbb{R}$ for each $1 \le l \le L'$, where here $L' \in \mathbb{N}$. We will assume that $\mathbb{E}_{\pi_l}[|\varphi_{l'}(U)|] < +\infty$ for every $(l, l') \in \{1, \ldots, L\} \times \{1, \ldots, L'\}$. One has that

$$\mathbb{E}_{\pi_L}[\varphi_{L'}(U)] = \sum_{l=1}^{L} \sum_{l'=1}^{L'} \{[\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}][\varphi_{l'}(U) - \varphi_{l'-1}(U)]\}$$

with the convention that $\varphi_0(u) = 0 \ \forall u \in \mathsf{E}$ and $\mathbb{E}_{\pi_0}[\cdot]$ is equal to zero for any real-valued measurable function. Then [53] show that

$$\begin{aligned}\mathbb{E}_{\pi_L}[\varphi_{L'}(U)] \quad = \quad & \sum_{l=1}^{L} \sum_{l'=1}^{L'} \Big\{ \mathbb{E}_{\pi_l}\Big[\Big(1 - \frac{\kappa_{l-1}(U)}{\kappa_l(U)}\Big)(\varphi_{l'}(U) - \varphi_{l'-1}(U))\Big] + \\ & \mathbb{E}_{\pi_l}\Big[\Big(\frac{\kappa_{l-1}(U)}{\kappa_l(U)} - 1\Big)\Big]\mathbb{E}_{\pi_{l-1}}[\varphi_{l'}(U) - \varphi_{l'-1}(U)]\Big\}.\end{aligned}$$

The work [53] replaces L' on the right hand side with judiciously chosen $L'(l)$ and approximates the resulting expression with two independent MCMC chains for each summand: one targeting π_l and the other π_{l-1}. This estimator is shown to provide the optimal cost (in the sense of Section 3) as a function of error under appropriate assumptions.

We remark that if φ does not upon on l then one has the identity

$$\begin{aligned}\mathbb{E}_{\pi_L}[\varphi(U)] \quad = \quad & \sum_{l=1}^{L} \Big\{ \mathbb{E}_{\pi_l}\Big[\Big(1 - \frac{\kappa_{l-1}(U)}{\kappa_l(U)}\Big)\varphi(U)\Big] + \\ & \mathbb{E}_{\pi_l}\Big[\Big(\frac{\kappa_{l-1}(U)}{\kappa_l(U)} - 1\Big)\Big]\mathbb{E}_{\pi_{l-1}}[\varphi(U)]\Big\}. \end{aligned} \tag{18}$$

It is worth noting that the identity (18) has two summands per level in l, rather than the single summand per level as in (16), which may be viewed as a benefit of the latter identity. It is also worth mentioning that the work [53] consider a limiting infinite-dimensional parameter space E, and account for the resulting approximation error in the method as well. This is dealt with by balancing the error arising from the finite dimensional approximation of the parameter space with the approximation error arising from the solution of the PDE. The works [32] and [7] employ a similar strategy to deal with the case of infinite-dimensional parameter space.

In [99], the numerator and denominator of (2) are approximated independently in terms of expectation w.r.t. the prior in a Bayesian context. This corresponds to a simple change of measure for Monte Carlo and Quasi Monte Carlo, although for MLMC it is slightly different, since MLMC is used separately for the numerator and the denominator. See [96, 97] for some contexts where this idea can be useful.

5.2 Approximate Coupling

The problem in Section 2.2.2 will be considered here in order to illustrate this idea. The method employed is based on the single level method introduced in Section 4.3. It is crucial to understand the example, as well as the basic single level method (in addition to the MLMC methodology described in Section 3) in order to follow this section. In particular, there is a stochastic process that is partially observed. It is assumed that the discretized dynamics of the associated discretized processes can be easily coupled. This is the case in Section 2.2.2 as the Euler dynamics can be coupled, for any fixed parameter value θ. We remark that the method which will be described below has use outside the case of Section 2.2.2, see [66] for example.

The main utility of the method to be described is in the case that sampling from (the distributions associated to) coupled pairs of discretizations (π_l, π_{l-1}) in a meaningful way (e.g. in the case of Assumption 3.2) is not currently possible. Again, this is the case in the example of Section 2.2.2. Given this, an *approximate* coupling is devised, which: (i) can be sampled from (using MCMC/SMC), and (ii) has marginals which are similar to, but not exactly equal to, the pair (π_l, π_{l-1}). Then the difference $\{\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}\}[\varphi(U)]$ is replaced with an importance sampling formula (change of measure w.r.t. the approximate coupling) and is approximated by sampling from the approximate coupling. The motivation for the idea will become clear as it is described in more detail. The approach is considered in [65] (see also [66]).

For the moment it suffices to consider a single level l. Now consider the fine and coarse densities for this l, given by:

$$\pi_p^k(\theta, u_{1:k}) \propto \nu(\theta) \prod_{i=1}^k G_\theta(u_i, y_i) Q_\theta^p(u_{(i-1)}, u_i) \,,$$

for consecutive levels $p = l, l-1 \geq 1$. Let $\check{u}_{1:k} = (\overline{u}_{1:k}, \underline{u}_{1:k}) \in \mathbb{R}^{2dk}$ and $\theta \in \Theta$. Recall again that the ideal situation is that we construct a coupled density $\pi_{l-1:l}(\check{u}_{1:k})$ which we can sample from such that for all sets A, we have

$$\int_{A \times \mathbb{R}^d} \pi_{l-1:l}(\check{u}_{1:k}) d\check{u}_{1:k} = \int_A \pi_l^k(\overline{u}_{1:k}) d\overline{u}_{1:k}$$

and

$$\int_{\mathbb{R}^d \times A} \pi_{l-1:l}(\check{u}_{1:k}) d\check{u}_{1:k} = \int_A \pi_{l-1}^k(\underline{u}_{1:k}) d\underline{u}_{1:k}$$

and recall that the issue is that in this case we do not know how to do this. We do however know that one can construct a good coupling of the two discretized kernels $Q_\theta^l, Q_\theta^{l-1}$ for any fixed θ by sampling the finer Gaussian increments and concatenating them for the coarser discretization (e.g., [44] or as written in [64, 68]). More precisly, given $\check{u}_{i-1} = (\overline{u}_{i-1}, \underline{u}_{i-1}) \in \mathbb{R}^{2d}$ and $\theta \in \Theta$, there is a Markov kernel $\check{Q}_\theta^{l,l-1}$ such that for any set A

$$\int_{A \times \mathbb{R}^d} \check{Q}_\theta^{l,l-1}(\check{u}_{i-1}, \check{u}_i) d\check{u}_i = \int_A Q_\theta^l(\overline{u}_{i-1}, \overline{u}_i) d\overline{u}_i \,, \tag{19}$$

and

$$\int_{\mathbb{R}^d \times A} \check{Q}_\theta^{l,l-1}(\check{u}_{i-1}, \check{u}_i) d\check{u}_i = \int_A Q_\theta^{l-1}(\underline{u}_{i-1}, \underline{u}_i) d\underline{u}_i \,, \tag{20}$$

and such that if $\mathbb{E}|\underline{u}_{i-1} - \overline{u}_{i-1}|^2$ is small and $\check{U}_i \sim \check{Q}_\theta^{l,l-1}(\check{u}_{i-1}, \cdot)$, then $\mathbb{E}|\underline{u}_i - \overline{u}_i|^2$ is small.

We then consider the joint probability on $\Theta \times \mathbb{R}^{2kd}$:

$$\check{\pi}_{l-1:l}^k(\theta, \check{u}_{1:k}) \propto \nu(\theta) \prod_{i=1}^k \check{G}_\theta(\check{u}_i, y_i) \check{Q}_\theta^{l,l-1}(\check{u}_{i-1}, \check{u}_i)$$

for any non-negative function $\check{G}_\theta(\check{u}_i, y_i)$. Whilst this function can be 'arbitrary', up-to some constraints, we set it as

$$\check{G}_\theta(\check{u}_i, y_i) = \max\{G_\theta(\overline{u}_i, y_i), G_\theta(\underline{u}_i, y_i)\}. \tag{21}$$

The motivation for this choice will be explained below. Let $\varphi : \Theta \times \mathbb{R}^{2kd} \to \mathbb{R}$ be π_l^k and π_{l-1}^k-integrable. Then we have

$$\mathbb{E}_{\pi_l^k}[\varphi(\theta, U_{1:k})] - \mathbb{E}_{\pi_{l-1}^k}[\varphi(\theta, U_{1:k})] =$$
$$\frac{\mathbb{E}_{\check{\pi}_{l-1:l}^k}[\varphi(\theta, \overline{U}_{1:k})\overline{H}_\theta(\check{U}_{1:k})]}{\mathbb{E}_{\check{\pi}_{l-1:l}^k}[\overline{H}_\theta(\check{U}_{1:k})]} - \frac{\mathbb{E}_{\check{\pi}_{l-1:l}^k}[\varphi(\theta, \underline{U}_{1:k})\underline{H}_\theta(\check{U}_{1:k})]}{\mathbb{E}_{\check{\pi}_{l-1:l}^k}[\underline{H}_\theta(\check{U}_{1:k})]}, \tag{22}$$

where

$$\begin{aligned} \overline{H}_\theta(\check{u}_{1:k}) &= \prod_{i=1}^k \frac{G_\theta(\overline{u}_i, y_i)}{\check{G}_\theta(\check{u}_i, y_i)}, \\ \underline{H}_\theta(\check{u}_{1:k}) &= \prod_{i=1}^k \frac{G_\theta(\underline{u}_i, y_i)}{\check{G}_\theta(\check{u}_i, y_i)}. \end{aligned}$$

Since we consider k fixed here, the dependence of H on k is suppressed. The difference can then be approximated by sampling from $\check{\pi}_{l-1:l}^k$ using the PMMH [1] from Section 4.3. This is done independently for each of the L summands in the ML identity, with the first summand (the coarsest discretization) sampled by a standard PMMH.

The basic idea is that one knows how to construct an exact coupling of the discretizations of the prior, i.e., the stochastic forward dynamics here, and it is natural to leverage this. However, as noted previously, exact couplings of the posterior are not trivial to sample. Instead, one aims to construct a joint probability that should have marginals which are close to, but not exactly equal to, the correct ones. As the coupling is not exact, one must correct for this fact with importance sampling, as in Section 5.1. Just as argued in that section, the associated weights of the importance sampling, here given by $(\overline{H}_\theta, \underline{H}_\theta)$, should be well behaved in some sense. This can be ensured by choosing the function $\check{G}_\theta$ so that the variance of the weights w.r.t. any probability density will remain bounded uniformly in time. This is ensured by the choice (21). [65] are able to prove, under suitable assumptions on the model and PMMH kernel, that the computational effort to estimate a class of expectations is reduced versus a single PMMH algorithm on the finest level, for a given MSE sufficiently small. The reduction in cost is a direct consequence of the prior coupling and well-behaved importance weights, in connection with the ML identity. Note that the results of [65] do not

[1] More precisely, one samples from a suitably extended distribution, as mentioned before and as described in [1, 3].

consider the dependence on the time parameter k, and this is an important consideration for future work. The quantities for which rates α and β need to be obtained in this context are given in Table 2. The idea has been further developed, analyzed and adapted for unbiased inference in [42] and optimal control [63].

To illustrate the ideas the following Langevin SDE is considered,

$$
\begin{aligned}
dU_t &= \frac{1}{2}\nabla \log \pi(U_t)dt + \sigma dW_t, \qquad U_0 = u_0\,, \\
Y_k|u_k &\sim \mathcal{N}(0, \tau^2 \exp u_k)\,,
\end{aligned}
$$

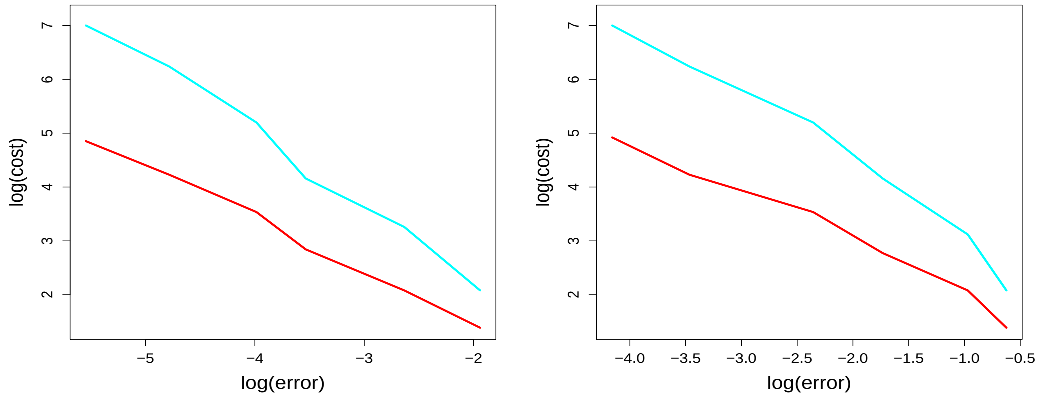
where $\pi(x)$ is the probability density function of a Student's t-distribution with ρ degrees of freedom. The parameters of interest are $\theta = (\rho, \sigma)$, and these are given priors (independently for ρ and σ), $\rho \sim \mathcal{G}(1,1), \sigma \sim \mathcal{G}(1,1)$ where $\mathcal{G}(a,b)$ denotes the Gamma distribution with shape a and scale b and $u_0 = 0$. A data set with 1000 observations is simulated with $\rho = 10$, $\sigma = 1$, and $\tau^2 = 1$.

The results are illustrated in Figure 2. They illustrate the improvement of the method in [65] against PMCMC on the most accurate target. The function estimated is the parameter itself. Figure 2 shows that for a MSE of $\mathcal{O}(\epsilon^2)$, the ML approach costs $\mathcal{O}(\epsilon^{-2})$ and PMCMC costs $\mathcal{O}(\epsilon^{-3})$. Complete details of the simulations are in [65].

Rate parameter	Relevant quantity
α	$(\mathbb{E}_{\pi_L^k} - \mathbb{E}_{\pi^k})(\varphi)$
β	$\int_{\Theta\times\mathbb{R}^{2kd}} \lvert\varphi(\theta, \overline{u}_{1:k}) - \varphi(\theta, \underline{u}_{1:k})\rvert^2 \check{\pi}_{l-1:l}^k(\theta, \check{u}_{1:k})d\theta d\check{u}_{1:k}$

Table 2: The key rates of convergence required for PMMH using MLMC.

We end this section by mentioning that the strategy described in this section is not the only one that could be adopted. For instance, the importance sampling approach of the previous section might also be considered. There are some reasons why the approximate coupling method described in this section might be preferred, for the example considered here. Firstly, the terms $\kappa_l(u)/\kappa_{l-1}(u)$ are not available pointwise for this example; this could possibly be dealt with by random weight ideas (e.g., [43, 94]), but it is still an issue.



(a) σ

(b) ρ

Figure 2: Log cost against Log MSE for the inference on two parameters. The red line is the ML approach and the blue a PMMH algorithm.

Secondly, there is a well-designed MCMC algorithm for the target (i.e., a PMMH kernel) and hence one would like to use this, as it is one of the gold standards for such models.

5.3 Coupling Algorithms

We now consider the case where one seeks to approximate the differences in the ML identity exactly by somehow trying to *correlate* or couple stochastic algorithms, rather than by constructing any joint coupling, either exact or approximate. We refer to this general strategy as coupling algorithms and it is explained further below.

5.3.1 Coupling MCMC

We begin by considering the method in [32], which is a Markov chain approach. Consider two probability densities (π_l, π_{l-1}), where the support of π_{l-1} is E and the support of π_l is $\mathsf{E} \times \mathsf{U}$. We focus on computing $\mathbb{E}_{\pi_l}[\varphi_l(U)] - \mathbb{E}_{\pi_{l-1}}[\varphi_{l-1}(U)]$ where $\varphi_l : \mathsf{E} \times \mathsf{U} \to \mathbb{R}$ and $\varphi_{l-1} : \mathsf{E} \to \mathbb{R}$ are π_l and π_{l-1} integrable.

Suppose we have a current state $(u_l, u_{l-1}) \in (\mathsf{E} \times \mathsf{U}) \times \mathsf{E}$. Define $u_l = (u_{l,c}, u_{l,f})$, where $u_{l,c} \in \mathsf{E}$ and $u_{l,f} \in \mathsf{U}$. The approach consists of two steps. First one samples $u'_{l-1} \sim \pi_{l-1}$ exactly. This state is accepted and is the next state of the approximation at level $l-1$. Now, one prescribes $u'_{l,c} = u'_{l-1}$, and a new state is proposed from a proposal $r_l(u_l, u'_l) = \pi_{l-1}(u'_{l,c}) R_l(u'_{l,f}|u_l, u'_{l,c})$, and is accepted or rejected according to the standard Metropolis-Hastings method [32, Algorithm 2]. This acceptance probability is, to move from u_l to $u'_l = (u'_{l-1}, u'_{l,f})$ (i.e. on level l)

$$\min\left\{1, \frac{\pi_l(u'_l)\pi_{l-1}(u_{l,c})R_l(u_{l,f}|(u'_{l-1}, u'_{l,f}), u_{l,c})}{\pi_l(u_l)\pi_{l-1}(u'_{l-1})R_l(u'_{l,f}|u_l, u'_{l-1})}\right\}.$$

In [32] the particular case of $r_l(u_l, u'_l) = \pi_{l-1}(u'_{l,c})R_l(u'_{l,f}|u_{l,f})$ is considered, i.e. the proposal for the fine degrees of freedom depends only on the fine degrees of freedom from the previous step, and does not depend on the coarse degrees of freedom at all. One can observe that the acceptance probability in this case is

$$\min\left\{1, \frac{\pi_l(u'_{l-1}, u'_{l,f})\pi_{l-1}(u_{l,c})r_l(u'_{l,f}, u_{l,f})}{\pi_l(u_{l,c}, u_{l,f})\pi_{l-1}(u'_{l-1})r_l(u_{l,f}, u'_{l,f})}\right\}.$$

It is clear that the samples are coupled and have the correct invariant distribution. [32] show that this approach can obtain the advantages of MLMC estimation for some examples (see for instance [32, Figure 4]). As noted in that article, it is not realistic to assume that exact sampling from π_{l-1} is feasible. The authors propose a subsampling method to assist with this. While this heuristic is sensible, there is no mathematical proof that this procedure is correct. As stated in [32, pp. 1086, par. 3], the mathematical analysis in that article does not correspond to the actual algorithm which is simulated in practice, but rather the ideal perfect algorithm ([32, Algorithm 2]) which can seldom be implemented in practice (for if it could be, one most likely would not consider MCMC).

Before concluding this section, we mention the recent related work [104] (see also [46]), which considers coupled Stochastic Gradient Langevin algorithms for some i.i.d. models in Bayesian statistics. Note that there the levels are from an Euler discretization associated to the algorithm, not the model per-se as described here. Recent work in [62] develops a new coupled MCMC method with convergence analysis, which is useful for rejection-free Markov chain simulation.

5.3.2 Coupling Particle Filters

Next we consider the context of Section 2.2.1 of filtering a discretely and partially observed diffusion processes. We describe an approach in [64] (see also [47, 56, 68, 100]), which consists of a strategy for employing the MLMC framework of Section 3 in the context of the particle filter, the single level version of which is presented in Section 4.2. For $l \geq 2$, $\varphi : \mathbb{R}^d \to \mathbb{R}$, π_l^k and π_{l-1}^k-integrable, we consider approximating the difference $\mathbb{E}_{\pi_l^k}[\varphi(U_k)] - \mathbb{E}_{\pi_{l-1}^k}[\varphi(U_k)]$ sequentially in time. Note that just running particle filters as in Algorithm 1, independently for π_l^k and π_{l-1}^k, will not achieve any correlation (in the sense of Assumption 3.2) and thus it is unlikely to bring any advantage versus just targeting π_L^k using Algorithm 1. This is the motivation for the approach outlined here. Some of the notations of Section 5.2 are also used. The parameter θ is also dropped from the notations, as it is assumed to be fixed here. For instance the Markov kernel $\check{Q}_\theta^{l,l-1}$ in (19)-(20) is used below, except we write $\check{Q}^{l,l-1}$.

The multilevel particle filter (MLPF) is described as follows. First, for $l = 1$, run a particle filter (Algorithm 1) for the coarsest discretization. Now, run Algorithm 4 independently for each $l \geq 2$. The notation of Algorithm 4 can be described as follows. $(\overline{U}_k^{l,i}, \underline{U}_k^{l,i}) \in \mathbb{R}^d \times \mathbb{R}^d$ is a *correlated pair of samples.* Independently for $i \in \{1, \ldots, N_l\}$, at time $k \geq 1$ and level $l \in \{2, \ldots, L\}$, $\overline{U}_k^{l,i}$ is a sample which is used (in a way to be described) to approximate expectations w.r.t. π_l^k, and $\underline{U}_k^{l,i}$ is a sample used to approximate expectations w.r.t. π_{l-1}^k. $(\overline{I}_k^{l,i}, \underline{I}_k^{l,i}) \in \{1, \ldots, N_l\}^2$ are a *correlated pair* of resampled indices, similarly to the iterate step of Algorithm 1. $\overline{I}_k^{l,i}$ (resp. $\underline{I}_k^{l,i}$) is associated to $\overline{U}_k^{l,i}$ (resp. $\underline{U}_k^{l,i}$). The correlation of these indices is the key to obtaining samples of the filter which remain correlated so that an estimate of the form in Assumption 3.2 holds.

The coupled resampling procedure for the indices $(\overline{I}_k^{l,i}, \underline{I}_k^{l,i})$ is described below. First let

$$\overline{w}_k^{l,i} = \frac{G(\overline{u}_k^{l,i}, y_k)}{\sum_{j=1}^{N_l} G(\overline{u}_k^{l,j}, y_k)} \qquad \text{and} \qquad \underline{w}_k^{l,i} = \frac{G(\underline{u}_k^{l,i}, y_k)}{\sum_{j=1}^{N_l} G(\underline{u}_k^{l,j}, y_k)} \,. \tag{23}$$

Now using for $a, b \in \mathbb{R}$, $a \wedge b = \min\{a, b\}$

a. with probability $\alpha_k^l = \sum_{i=1}^{N_l} \overline{w}_k^{l,i} \wedge \underline{w}_k^{l,i}$, draw $\overline{I}_k^{l,i}$ according to

$$\mathbb{P}(\overline{I}_k^{l,i} = j) = \frac{1}{\alpha_k^l}(\overline{w}_k^{l,j} \wedge \underline{w}_k^{l,j}), \qquad j \in \{1, \ldots, N_l\}\,,$$

Algorithm 4 A Coupled Particle Filter

```
For $i = 1, \dots, N_l$, draw $(\overline{U}_1^{l,i}, \underline{U}_1^{l,i}) \overset{\text{i.i.d.}}{\sim} \check{Q}^{l,l-1}((u_0, u_0), \cdot)$.
Initialize $k = 1$. Do
    (i) For $i = 1, \dots, N_l$, draw $(\overline{I}_k^{l,i}, \underline{I}_k^{l,i}) \in \{1, \dots, N_l\}^2$ according to the coupled resampling procedure below. Set $k = k + 1$.
    (ii) For $i = 1, \dots, N_l$, independently draw $(\overline{U}_k^{l,i}, \underline{U}_k^{l,i}) | (\overline{u}_{k-1}^{l,\overline{I}_k^{l,i}}, \underline{u}_{k-1}^{l,\underline{I}_k^{l,i}}) \sim \check{Q}^{l,l-1}((\overline{u}_{k-1}^{l,\overline{I}_k^{l,i}}, \underline{u}_{k-1}^{l,\underline{I}_k^{l,i}}),\ \cdot\ )$.
```

and let $\underline{I}_k^{l,i} = \overline{I}_k^{l,i}$.

b. otherwise, draw $(\overline{I}_k^{l,i}, \underline{I}_k^{l,i})$ independently according to the probabilities

$$\mathbb{P}(\overline{I}_k^{l,i} = j) = [\overline{w}_k^{l,j} - \overline{w}_k^{l,j} \wedge \underline{w}_k^{l,j}]/(\sum_{s=1}^{N_l} \overline{w}_k^{l,s} - \overline{w}_k^{l,s} \wedge \underline{w}_k^{l,s}), \qquad j \in \{1, \dots, N_l\},$$

$$\mathbb{P}(\underline{I}_k^{l,i} = j) = [\underline{w}_k^{l,j} - \overline{w}_k^{l,j} \wedge \underline{w}_k^{l,j}]/(\sum_{s=1}^{N_l} \underline{w}_k^{l,s} - \overline{w}_k^{l,s} \wedge \underline{w}_k^{l,s}), \qquad j \in \{1, \dots, N_l\}.$$

Note that by using the coupled kernel $\check{Q}^{l,l-1}$, one is sampling from the exact coupling of the discretized process, $(\overline{U}_k^{l,i}, \underline{U}_k^{l,i})$. Now one wants to maintain as much dependence as possible in the resampling, since resampling is necessary in particle filters. The coupled resampling described above maximizes the probability (conditional on the history) that the pair of samples remain coupled (see also [19]).

In the work [64], it is shown that

$$\sum_{i=1}^{N_l} \left\{ \varphi(\overline{u}_k^{l,i}) \overline{w}_k^{l,i} - \varphi(\underline{u}_k^{l,i}) \underline{w}_k^{l,i} \right\}$$

consistently approximates $\mathbb{E}_{\pi_l^k}[\varphi(U_k)] - \mathbb{E}_{\pi_{l-1}^k}[\varphi(U_k)]$. The MLPF estimator of $\mathbb{E}_{\pi_L^k}[\varphi(U_k)]$ is therefore given by

$$\sum_{i=1}^{N_1} w_k^{1,i} \varphi(u_k^{1,i}) + \sum_{l=2}^{L} \sum_{i=1}^{N_l} \left\{ \varphi(\overline{u}_k^{l,i}) \overline{w}_k^{l,i} - \varphi(\underline{u}_k^{l,i}) \underline{w}_k^{l,i} \right\}.$$

In the above displayed equation $\sum_{i=1}^{N_1} w_k^{1,i} \varphi(u_k^{1,i})$ is produced by a particle filter (Algorithm 1) that uses proposals Q^1 and resampling weights $w_k^{1,i} = G(u_k^{1,i}, y_k) / \sum_{j=1}^{N_1} G(u_k^{1,j}, y_k)$.

In the case of Euler-Maruyama discretization, [64] show that under suitable assumptions and for finite time the standard choice of L and $N_{1:L}$ as in (10) and (11) provides an MSE of $\mathcal{O}(\epsilon^2)$ for a cost of $\mathcal{O}(\epsilon^{-2.5})$ (for diffusions with constant diffusion coefficients, the cost is $\mathcal{O}(\epsilon^{-2}\log(\epsilon)^2)$). For a particle filter the cost required is $\mathcal{O}(\epsilon^{-3})$. The quantities for which rates α and β need to be obtained in this context are given in Table 3. It is important to note here that the rate of convergence of the increment estimators is halved with respect to standard forward MLMC estimation here, as a result of the coupled resampling step. It is of interest to obtain a filtering algorithm which preserves the forward rate. The theory is not limited to Euler discretizations, but the ultimate bound on the cost will depend on the convergence rate of the numerical method. The ideas here have been extended to the case of partially and discretely observed Lévy processes in [61].

This method is illustrated on the Langevin diffusion as in Section 5.2, except with θ fixed. Real daily S&P 500 log return data (from August 3, 2011 to July 24, 2015, normalized to unity variance) is used. The complexity results are illustrated in Figure 3 with $\varphi(u_k) = \tau^2 e^{u_k}$ as the function, considered at the final time. The figure compares the particle filter to the MLPF. Figure 3 illustrates that for a MSE of $\mathcal{O}(\epsilon^2)$, the MLPF costs $\mathcal{O}(\epsilon^{-2}\log(\epsilon)^2)$ and the particle filter costs $\mathcal{O}(\epsilon^{-3})$.

Rate parameter	Relevant quantity
α	$(\mathbb{E}_{\pi_L^k} - \mathbb{E}_{\pi^k})(\varphi)$
β	$\left(\int_{\mathbb{R}^{2d}} \lvert\varphi(\overline{u}_k) - \varphi(\underline{u}_k)\rvert^2 \check{\pi}_{l-1:l}^k(\check{u}_k) d\check{u}_k\right)^2$

Table 3: The key rates of convergence required for MLPF.

In [68] the approach and results are extended to the case of marginal likelihood estimation. We note that one drawback of the mathematical results in [64, 68] is that they do not consider the time-parameter (see [58] for some work on this). In [56, 100] the coupled resampling method is improved by using optimal transportation techniques [105]. Also [47] (see also [48]) obtain empirical results which indicate that more favorable convergence rates may be preserved in certain cases by replacing the resampling step with a deterministic

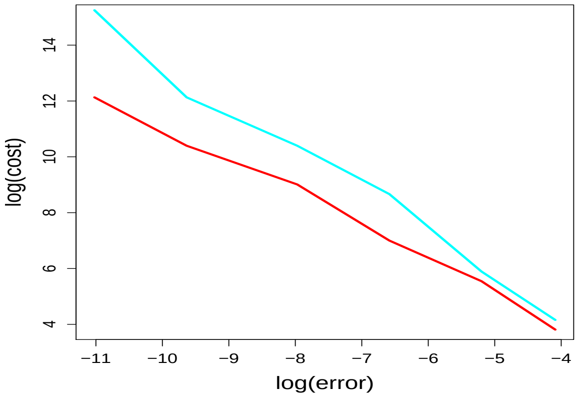


Figure 3: Log Cost against Log MSE for the MLPF (red) against the PF for a Langevin SDE of the form (7) [64].

linear transformation of the current population, derived from the optimal transportation. However, in [47, 48, 56, 100] there are no mathematical results which support the empirical results.

5.3.3 Coupling the EnKF

As discussed in subsection 4.4, the EnKF targets a different distribution than the filtering distribution in general, which has been denoted $\widehat{\pi}^k_L$. In between updates, this algorithm proceeds similarly to the MLPF of the previous section, propagating a pair of ensembles for each $l \geq 2$ using the kernel $\check{Q}^{l,l-1}$ as defined in equations (19) and (20). The fundamental difference in the update results in approximations of increments $\mathbb{E}_{\widehat{\pi}^k_l}[\varphi(U_k)] - \mathbb{E}_{\widehat{\pi}^k_{l-1}}[\varphi(U_k)]$. Therefore, the MSE will ultimately depend upon the difference $\mathbb{E}_{\widehat{\pi}^k_L}[\varphi(U_k)] - \mathbb{E}_{\pi^k}[\varphi(U_k)]$, which includes a Gaussian bias in addition to the discretization bias. In order to preserve the coupling of this algorithm after the update, the sample covariance is approximated using the entire multilevel ensemble in [54], as follows. Recall the functions $G(u_i, y_i)$ are assumed to take the form given in (15). The approach is in Algorithm 5.

The quantities for which rates α and β need to be obtained are given in table 4, and the complexity results are illustrated in Figure 4 for an example linear SDE of the form in (7).

Algorithm 5 Coupled ENKF

For $l = 2, \dots, L$, and $i = 1, \dots, N_l$, draw $(\overline{U}_1^{l,i}, \underline{U}_1^{l,i}) \overset{\text{i.i.d.}}{\sim} \check{Q}^{l,l-1}((u_0, u_0), \cdot)$. And draw $U_1^{1,i} \sim Q^1(u_0, \cdot)$.

Initialize $k = 1$. **While** $k \leq n - 1$ **Do**

(i) Compute the MLMC covariance estimator [11] :

$$
\begin{aligned}
C_k^{\text{ML}} =& \frac{1}{N_1} \sum_{i=1}^{N_1} u_k^{1,i} (u_k^{1,i})^T - \left(\frac{1}{N_1} \sum_{i=1}^{N_1} u_k^{1,i} \right) \left(\frac{1}{N_1} \sum_{i=1}^{N_1} u_k^{1,i} \right)^T \\
&+ \sum_{l=2}^{L} \Bigg[\frac{1}{N_l} \sum_{i=1}^{N_l} \left(\overline{u}_k^{l,i} (\overline{u}_k^{l,i})^T - \underline{u}_k^{l,i} (\underline{u}_k^{l,i})^T \right) - \\
&\left(\frac{1}{N_l} \sum_{i=1}^{N_l} \overline{u}_k^{l,i} \right) \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \overline{u}_k^{l,i} \right)^T + \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \underline{u}_k^{l,i} \right) \left(\frac{1}{N_l} \sum_{i=1}^{N_l} \underline{u}_k^{l,i} \right)^T \Bigg].
\end{aligned}
$$

(ii) Compute $K_k^{\text{ML}} = C_k^{\text{ML}} H^T (H C_{+,k}^{\text{ML}} H^T + \Gamma)^{-1}$, where $C_{+,k}^{\text{ML}}$ is a positive semi-definite modification of C_k^{ML}.

(iii) **For** $l = 2, \dots, L$, and $i = 1, \dots, N_l$, independently draw $Y_k^{l,i} \sim \mathcal{N}(y_k, \Gamma)$, and compute

$$
\widehat{\overline{u}}_k^{l,i} = (I - K_k^{\text{ML}} H) \overline{u}_k^{l,i} + K_k^{\text{ML}} y_k^{l,i} \,,
$$

and similarly for $\widehat{\underline{u}}_k^{l,i}$ and $\widehat{u}_k^{1,i}$. Set $k = k + 1$.

(iv) **For** $l = 2, \dots, L$, and $i = 1, \dots, N_l$, independently draw $(\overline{U}_k^{l,i}, \underline{U}_k^{l,i}) \overset{\text{i.i.d.}}{\sim} \check{Q}^{l,l-1}((\widehat{\overline{u}}_{k-1}^{i}, \widehat{\underline{u}}_{k-1}^{i}), \cdot)$. And draw $U_k^{1,i} \sim Q^1(\widehat{u}_{k-1}^{1,i}, \cdot)$.

In particular, consider the Ornstein–Uhlenbeck SDE problem

$$du_t = -u_t dt + \sigma dW_t, \qquad u(0) = 1.$$

The corresponding noisy observations are given by

$$y_n = u_n + \eta_n,$$

with $\eta_n \sim \mathcal{N}(0, \Gamma)$ i.i.d. The work [54] established that slightly modified choices of L and $N_{1:L}$ provide MSE at step k of $\mathcal{O}(|\log \epsilon|^{2k}\epsilon^2)$ for a cost of $\mathcal{O}(\epsilon^{-2}\tilde{K}_L^{3/2})$, where $\tilde{K}_L^{1/2} = \sum_{l=1}^{L} h_l^{(\beta-\zeta)/3}$. However, the numerical results indicate not only a time-independent rate of convergence without logarithmic penalty, but in fact also a time-uniform constant – see Figure 4. Presumably the penalty on the MSE is mostly a technical hurdle. The recent work [15] has extended this method to spatial processes, for example given by stochastic partial differential equations. This is the context where the EnKF is typically applied, for example in numerical weather prediction.

In the numerical example the covariance parameters are taken as $\Gamma = 0.04$ and $\sigma = 0.5$. The numerical method employed is Euler method, with $\alpha = \gamma = 1$ and $\beta = 2$ for this example, and we observe the rates $\mathcal{O}(\varepsilon^{-3})$ and $\mathcal{O}(\varepsilon^{-2})$ for EnKF and MLEnKF, respectively. Here the MSE of the covariance is computed over the observation times with $n = 100, 200$ and 400.

Rate parameter	Relevant quantity
α	$(\mathbb{E}_{\widehat{\pi}_L^k} - \mathbb{E}_{\pi^k})(\varphi)$
β	$\left(\int_{\mathbb{R}^{2d}} \|\varphi(\overline{u}_k) - \varphi(\underline{u}_k)\|^p \check{\widehat{\pi}}^k_{l-1:l}(\overline{u}_k, \underline{u}_k) d\overline{u}_k d\underline{u}_k\right)^{2/p}$

Table 4: The key rates of convergence required for MLEnKF, for all $p \geq 1$, where $\check{\widehat{\pi}}^k_{l-1:l}$ denotes the coupled density/distribution resulting from the algorithm above with the mean-field limiting gain K_k.

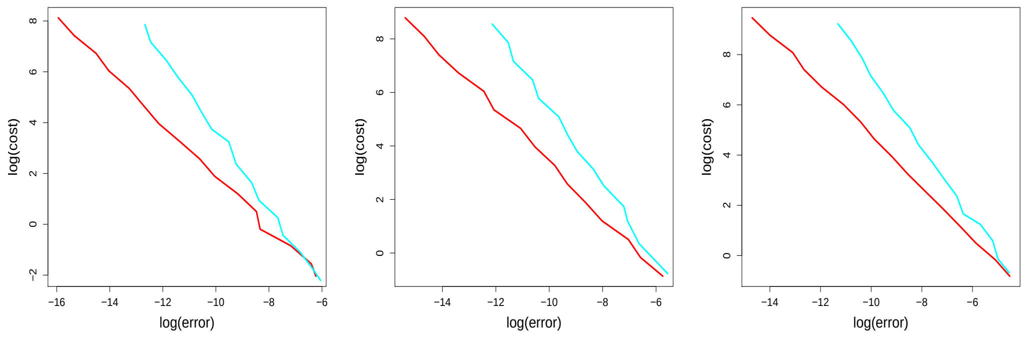


Figure 4: Comparison of the accuracy vs. computational cost when using the EnKF and MLEnKF methods on a linear Gaussian filtering problem of the form given in (7). The observations occur at times $1, \dots, n$. The error is measured in terms of the RMSE for the covariance, computed with $n = 100, 200$ and 400 observation times in the first, second and third column, respectively. The computational cost is measured in computer runtime. Again the blue line is EnKF, following $\mathcal{O}(\varepsilon^{-3})$ cost for MSE $\mathcal{O}(\varepsilon^{-2})$, and the red line is MLEnKF [54], following $\mathcal{O}(\varepsilon^{-2})$ cost for the same MSE.

6 Future Work

The main purpose of this article is to provide an exposition of some of the computational methods which can be used to apply MLMC. The relative merits of these methodologies, such as MCMC, particle filters, SMC samplers and so fourth are well-known in the literature (see for instance [23, 35, 90]). Given the knowledge of these ideas, some of which have been summarized in the article, how can one use them to expand the classes of problems for which MLMC could be beneficial? We have provided an introduction to some approaches that have been used so far, however there is still much work to be done.

There are many areas for possible exploration in future work. One strand consists of considering multi-dimension discretizations, such as in [50]. There are a small number of papers on this topic, such as [24, 66], but there seem to be many possible avenues for future work. Some of the theoretical results of [64, 68] for multilevel particle filtering, do not consider the time parameter. This important restriction should be dealt with. It is noted, however, that we expect the standard behaviour for particle filters with respect to time, as mentioned in Section 4.2. Another direction consists of a general method for sampling (e.g., by MCMC/SMC) exact (dependent) couplings of the targets in the ML identity. As we have commented, it does not appear to be trivial, but it may be far from impossible. Such a method would be very beneficial, as one could then appeal to existing literature in order to prove complexity results about MLMC and MIMC versions. Another very interesting avenue for future research is exact coupling, using optimal transport and i.i.d. sampling. For some model structures, the ideas of [103] could be very useful, as has been followed up in [55].

Finally, it is commented that it would be of interest to perform a systematic comparison on some benchmark problems such as the Bayesian inverse problem in Section 2.1 across the various existing MLMC methods for this type of problem, to asses their benefits and drawbacks in practice. For example such a comparison would include at least the 3 distinct MLMCMC (the third is a generalization of the method [65], introduced in [66] but not presented in this work), as well as MLSMC.

Acknowledgements

AJ and CS were supported under the KAUST Competitive Research Grants Program-Round 4 (CRG4) project, Advanced Multi-Level sampling techniques for Bayesian Inverse Problems with applications to subsurface, ref: 2584. KJHL was supported by Oak Ridge National Laboratory (ORNL) Directed Research and Development Seed funding and much of this work was performed while he was a staff member at ORNL. We thank the editor and two reviewers whose comments have greatly improved the paper.

References

[1] Andrieu, C., Doucet, A. & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc. Ser. B*, **72**, 269–342.

[2] Andrieu, C. & Moulines É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Prob.*, **16**, 1462–1505.

[3] Andrieu, C., & Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

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

[5] Beskos, A. , Crisan, D., Jasra, A. & Whiteley, N. (2014). Error bounds and normalizing constants for sequential Monte Carlo. *Adv. Appl. Probab.*, **46**, 279–306.

[6] Beskos, A., Jasra, A., Kantas, N. & Thiery, A. (2016). On the convergence of adaptive sequential Monte Carlo. *Ann. Appl. Probab.*, **26**, 1111-1146.

[7] Beskos, A., Jasra, A., Law, K., Marzouk, Y., & Zhou, Y. (2018). Multilevel Sequential Monte Carlo samplers with dimension independent likelihood informed proposals. *SIAM JUQ*, **6**, 762–786.

[8] Beskos, A., Crisan, D., Jasra, A., Kamatani, K., & Zhou, Y. (2017). A stable particle filter for a class of high-dimensional state-space models. *Adv. Appl. Probab.* **49**, 1-25.

[9] Beskos, A., Jasra, A., Law, K. J. H,, Tempone, R., & Zhou, Y. (2017). Multilevel Sequential Monte Carlo samplers. *Stoch. Proc. Appl.*, **127**, 1417-1440.

[10] Beskos, A., Roberts, G., Stuart, A., & Voss, J. (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics*, **8**(03), 319-350.

[11] Bierig, C., & Chernov, A. (2015). Convergence analysis of multilevel Monte Carlo variance estimators and application for random obstacle problems. *Numerische Mathematik*, **130**(4), 579–613.

[12] Bouchard-Cote, A., Vollmer, S. & Doucet, A. (2018). The bouncy particle sampler: A non-reversible rejection free MCMC method. *J. Amer. Statist. Assoc.* **113**, 855-867.

[13] Cappé, O., Ryden, T, & Moulines, É. (2005). *Inference in Hidden Markov Models.* Springer: New York.

[14] Chatterjee, S., & Diaconis, P. (2018). The sample size required in importance sampling. *Ann. Appl. Probab.* **28**, 1099-1135.

[15] Chernov, A., Hoel, H., Law, K., Nobile, F., & Tempone, R. (2016). Multilevel ensemble Kalman filtering for spatially extended models. arXiv preprint arXiv:1608.08558.

[16] Ching, J. & Chen, Y.-C. (2007). Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging, *J. Eng. Mech.*, **133**, 816–832.

[17] Chopin, N. (2002). A sequential particle filter for static models. *Biometrika*, **89**, 539–552.

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

[19] Chopin, N. & Singh, S. S. (2015). On particle Gibbs sampling. *Bernoulli*, **21**, 1855-1883.

[20] Chopin, N., Jacob, P. E., & Papaspiliopoulos, O. (2013). SMC^2: an efficient algorithm for sequential analysis of state space models. *J. R. Statist. Soc. B*, **75**, 397–426

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

[22] 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**, 424-446.

[23] Crisan, D., Rozovskii, B. (2011) The Oxford handbook of nonlinear filtering, *Oxford Univ. Press.* Oxford,

[24] Crisan, D, Del Moral, P., Houssineau, J., & Jasra, A. (2018). Unbiased Multi-Index Monte Carlo. *Stoch. Anal.*, **36**, 257-273.

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

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

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

[28] Del Moral, P., Doucet, A. & Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.*, **22**, 1009–1020.

[29] Del Moral, P., Jasra, A. & Law, KJH. (2017). Multilevel Sequential Monte Carlo: Mean Square Error Bounds under Verifiable Conditions. *Stoch. Anal.* **35**, 478–498.

[30] Del Moral, P, Jasra, A., Law, K. J. H. & Zhou, Y. (2017). Multilevel SMC samplers for normalizing constants. *TOMACS*, **27**, article 20.

[31] Deligiannidis, G., Doucet, A. & Pitt, M. (2018). The Correlated Pseudo-Marginal Method. *J. Roy. Statist. Soc. Ser. B*, **80**, 839-870.

[32] Dodwell, T. J., Ketelsen, C., Scheichl, R. & Teckentrup, A. L. (2015). A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. *SIAM/ASA J. Uncer. Quant.*, **3**, 1075–1108.

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

[34] Doucet, A., De Freitas, N. & Gordon, N. (2001). *Sequential Monte Carlo methods in practice*. Springer: New York.

[35] Doucet, A. & Johansen, A. (2011). A tutorial on particle filtering and smoothing: Fifteen years later. In *Handbook of Nonlinear Filtering* (eds. D. Crisan & B. Rozovsky), Oxford University Press: Oxford.

[36] Doucet, A., Pitt, M. K., Deligiannidis, G. & Kohn, R. (2015). Efficient Implementation of Markov chain Monte Carlo when Using an Unbiased Likelihood Estimator. *Biometrika*, **102**, 295-313.

[37] Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, **195**, 216–222.

[38] Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geo. Res.: Oceans*, **99**(C5), 10143–10162.

[39] Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dyn.*, **53**, 343–367.

[40] Fearnhead, P., Papaspiliopoulos, O. & Roberts, G. O. (2008). Particle filters for partially observed diffusions. *J. R. Stat. Soc. Ser. B* **70**, 755–777.

[41] Fearnhead, P., Latuszynzki, K., Roberts, G. O. & Sermaidis, G. (2017). Continuous-time importance sampling: Monte Carlo methods which avoid time discretization. arXiv preprint.

[42] Franks, J., Jasra, A., Law, K. J. H. & Vihola, M. (2018). Unbiased Inference for Hidden Markov Diffusion Models. arXiv preprint.

[43] Glasserman, P., & Staum, J. (2001). Conditioning on one-step survival for barrier options. *Op. Res.*, **49**, 923–937.

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

[45] Giles, M. B. (2015) Multilevel Monte Carlo methods. *Acta Numerica* **24**, 259-328.

[46] Giles, M. B., Nagapetyan, T., Szpruch, L., Vollmer, S., & Zygalakis, K. (2016). Multilevel Monte Carlo for Scalable Bayesian Computations. arXiv preprint.

[47] Gregory, A., Cotter, C., & Reich, S. (2016). Multilevel Ensemble Transform Particle Filtering. *SIAM J. Sci. Comp.* **38**, A1317-A1338.

[48] Gregory, A., & Cotter, C. (2016). A Seamless Multilevel Ensemble Transform Particle Filter. *SIAM J. Sci. Comp.*, **39**, A2684-A2701.

[49] Haario, H., Saksman, E. & Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.

[50] Haji-Ali, A. L., Nobile, F. & Tempone, R. (2016). Multi-Index Monte Carlo: When sparsity meets sampling. *Numerische Mathematik*, **132**, 767–806.

[51] Heinrich, S. (2001). Multilevel Monte Carlo methods. In *Large-Scale Scientific Computing*, (eds. S. Margenov, J. Wasniewski & P. Yalamov), Springer: Berlin.

[52] Heng, J., Doucet, A. & Pokern, Y. (2015). Gibbs Flow for Approximate Transport with Applications to Bayesian Computation. arXiv preprint.

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

[54] Hoel, H. Law, K. & Tempone, R. (2016). Multilevel ensemble Kalman filter. *SIAM J. Numer. Anal.*, **54**, 1813–1839.

[55] Houssineau, J., Jasra, & Singh, S. S. (2018). Multilevel Monte Carlo for smoothing via transport methods. *SIAM J. Sci. Comp.* **40**, A2315-A2335.

[56] Jacob, P. E., Lindsten, F. & Schonn, T. (2016). Coupling of particle filters. arXiv preprint.

[57] Jarzynski, C., (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693.

[58] Jasra, A., & Yu, F. (2018). Central limit theorems for coupled particle filters. arXiv:1810.04900.

[59] Jasra, A., Law, K. J. H. & Zhou, Y. (2016). Forward and inverse uncertainty quantification using multilevel Monte Carlo algorithms for an elliptic nonlocal equation. *Intl., J. Uncert. Quant.* **6**, 501–514.

[60] Jasra, A., Stephens, D. A. & Holmes C. C. (2007). On population-based simulation for static inference. *Statist. Comp.*, **17**, 263–279.

[61] Jasra, A., Law K. J. H. & Osei, P. P. (2019). Multilevel particle filters for Lévy driven stochastic differential equations. *Stat. Comp.* **29**, 775–789.

[62] Jasra, A., Law K. J. H. & Xu, Y. (2018). Markov chain simulation for multilevel Monte Carlo. arXiv preprint.

[63] Jasra, A., Heng, J., Bishop, A. & Xu, Y. (2018). A multilevel approach for stochastic nonlinear optimal control. arXiv preprint.

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

[65] Jasra, A., Kamatani, K., Law, K. J. H,& Zhou, Y. (2018). Bayesian Static Parameter Estimation for Partially Observed Diffusions via Multilevel Monte Carlo. *SIAM J. Sci. Comp.*, **40**, A887-A902.

[66] Jasra, A., Kamatani, K., Law, K. J. H., & Zhou, Y. (2018). A Multi-Index Markov Chain Monte Carlo Method. *Intl. J. Uncert. Quant.*, **8**, 61-73.

[67] Jasra, A., Stephens, D. A., Doucet, A. & Tsagaris, T. (2011). Inference for Lévy driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Statist.*, **38**, 1–22 .

[68] Jasra, A., Kamatani, K., Osei, P. P., & Zhou, Y. (2018). Multilevel particle filters: Normalizing Constant Estimation. *Statist. Comp.*, **28**, 47-60.

[69] Jasra, A., Jo, S., Nott, D., Shoemaker, C. & Tempone, R. (2019). Multilevel Monte Carlo in approximate Bayesian computation. *Stoch. Anal. Appl.*, **37**, 346-360.

[70] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, **82**(1), 35–45.

[71] Kantas, N., Beskos, A., & Jasra, A. (2014). Sequential Monte Carlo for inverse problems: a case study for the Navier Stokes equation. *SIAM/ASA JUQ*, **2**, 464–489.

[72] Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J. M. & Chopin, N. (2015) On Particle Methods for Parameter Estimation in General State-Space Models. *Statist. Sci.*, **30**, 328-351.

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

[74] Latz, J., Papaioannou, I. & Ullmann, E. (2018). Multilevel Sequential^2 Monte Carlo for Bayesian inverse problems. *J. Comp. Phys.*, **368**, 54-178.

[75] Law, K., Stuart, A. and Zygalakis, K. (2015). *Data Assimilation.* Springer-Verlag, New York.

[76] Law, K. J. H., Tembine, H., & Tempone, R. (2016). Deterministic mean-field ensemble Kalman filtering. *SIAM Journal on Scientific Computing*, **38**(3), A1251–A1279.

[77] Le Gland, F., Monbet, V., & Tran, V. D. (2009). Large sample asymptotics for the ensemble Kalman filter (Doctoral dissertation, INRIA).

[78] Llopis, F., Kantas, N., Beskos, A., & Jasra, A. (2018). Particle Filtering for stochastic Navier-Stokes signals observed with linear additive noise. *SIAM J. Sci. Comp.*, **40**, A1544-A1565.

[79] Mandel, J., Cobb, L., & Beezley, J. D. (2011). On the convergence of the ensemble Kalman filter. *Applications of Mathematics*, **56**(6), 533–541.

[80] Marin, J.-M., Pudlo, P., Robert, C.P. & Ryder, R. (2012). Approximate Bayesian computational methods. *Statist. Comp.*, **22**, 1167–1180.

[81] Meyn, S. & Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability.* Second edition, CUP: Cambridge.

[82] Naesseth, C. A., Lindsten, F., & Schön, T. B. (2015). Nested Sequential Monte Carlo Methods. *Proc. 32nd ICML.*

[83] Neal, R. M. (1996). *Bayesian Learning for Neural Networks.* Lecture Notes in Statistics, No. 118. Springer-Verlag.

[84] Neal, R. M. (2001). Annealed importance sampling. *Statist. Comp.*, **11**, 125–139.

[85] Ottobre, M., Pillai, N. S., Pinski, F. J., & Stuart, A. M. (2016). A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, **22**(1), 60-106.

[86] Ottobre, M. (2016). Markov Chain Monte Carlo and Irreversibility. *Reports on Mathematical Physics*, **77**(3), 267-292.

[87] Rebeschini, P. & Van Handel, R. (2015). Can local particle filters beat the curse of dimensionality? *Ann. Appl. Probab.*, **25**, 2809–2866.

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

[89] Robert, C. (2001). *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation.* Springer: New York.

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

[91] Roberts, G. O. & Tweedie, R. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.

[92] Roberts, G. O., & Rosenthal, J. (2004). General state-space Markov chains and MCMC algorithms. *Probab. Surveys*, **1**, 20–71.

[93] Roberts, G. O., Gelman, A. & Gilks W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, **7**, 110–120.

[94] Rousset, M., & Doucet, A. (2006). Discussion of Beskos et al. *J. R. Statist. Soc. B*, **68** 374–375.

[95] Schäfer, C. & Chopin, N. (2013). Adaptive Monte Carlo on binary sampling spaces. *Statist. Comp.*, **23**, 163–184.

[96] Schillings, C. & Schwab, C. (2016). Scaling limits in computational Bayesian inversion. *ESIAM Math. Mod. Numer. Anal.*, **50**, 1825–1856.

[97] Schillings, C., Sprungk, B., & Wacker, P. (2019). On the convergence of the Laplace approximation and noise level-robustness of Laplace-based Monte Carlo methods for Bayesian inverse problems. arXiv preprint.

[98] Schweizer, N. (2012). Non-asymptotic error bounds for sequential MCMC and stability of Feynman-Kac propagators. arXiv preprint, arXiv:1204.2382.

[99] Scheichl, R., Stuart A. & Teckentrup, A. L. (2016). Quasi-Monte Carlo and Multilevel Monte Carlo Methods for Computing Posterior Expectations in Elliptic Inverse Problems. arXiv preprint.

[100] Sen, D., Thiery, A., Jasra, A. (2018). On coupling particle filters. *Statist. Comp.*, **28**, 461-475.

[101] Singh, S. S., Lindsten, F. & Moulines, E. (2015). Blocking Strategies and Stability of Particle Gibbs Samplers. arXiv preprint.

[102] Snyder, C., Bengtsson, T., Bickel, P., & Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Month. Weather Rev.*, **136**, 4629–4640.

[103] Spantini, A., Bigoni, D. & Marzouk Y. (2017). Inference via low-dimensional couplings. arXiv preprint.

[104] Szpruch, L., Vollmer, S., Zygalakis, K. & Giles M. (2016). Multilevel Monte Carlo methods for the approximation of invariant distribution of Stochastic Differential Equations. arXiv preprint.

[105] Villani, C. (2008). *Optimal transport: old and new (Vol. 338)*. Springer Science & Business Media.

[106] Whiteley, N. P. (2012). Sequential Monte Carlo samplers: Error bounds and insensitivity to initial conditions. *Stoch. Anal.*, **30**, 774–798.