

# Advanced Multilevel Monte Carlo Methods

BY AJAY JASRA<sup>1</sup>, KODY LAW<sup>2</sup>, & CARINA SUCIU<sup>3</sup>

<sup>1</sup>Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG.

E-Mail: *staja@nus.edu.sg*

<sup>2</sup>Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, 37934 TN,

USA. E-Mail: *kodylaw@gmail.com*

<sup>3</sup>Center for Uncertainty Quantification in Computational Science & Engineering, King Abdullah

University of Science and Technology, Thuwal, 23955-6900, KSA.

E-Mail: *oana.suciu@kaust.edu.sa*

## Abstract

This article reviews the application of 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 a 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 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  $(E, \mathcal{E})$  be a measurable space,  $\pi$  be a probability measure on  $(E, \mathcal{E})$  and  $\varphi : E \rightarrow \mathbb{R}$  be a measurable and  $\pi$ -integrable function. In this article we are concerned with the computation of

$$\mathbb{E}_\pi[\varphi(U)] = \int_E \varphi(u)\pi(du) \quad (1)$$

for many different  $\pi$ -integrable functions  $\varphi$ . In addition, if  $\pi$  admits a density w.r.t. a dominating  $\sigma$ -finite measure  $du$  and if one can write for  $\kappa : E \rightarrow \mathbb{R}_+$

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

where  $Z$  is not known, but one can obtain  $\kappa$  up-to a non-negative unbiased estimator, then one is also interested in the computation of  $Z$ . These problems occur in a wide variety of real applications, often concerning Bayesian statistical inference. For instance, the computation of (1) can be associated to posterior expectations, and the value of  $Z$  can be used for Bayesian model selection; see [76]. Many of these problems are found in many real applications, such as meteorology, finance and engineering; see [65, 76]. Later in this article, we will expand upon the basic problem here.

We focus on the case when  $\pi$  is associated to some complex continuum problem, for instance, a continuous-time stochastic process or the solution of a partial differential equation (PDE), although 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  $\pi$ , call it  $\pi_L$ . It is explicitly assumed that this bias is associated with a *scalar* parameter  $h_L \in \mathbb{R}_+$  and that the bias disappears as  $h_L \rightarrow 0$ .
3. When using Monte Carlo methods, exact sampling from  $\pi_L$  is not possible; that is, one cannot sample i.i.d. from  $\pi_L$ .

Examples of models satisfying 1. & 2. include laws of stochastic differential equations (SDE) for which one cannot sample exactly (e.g., [64]), and one resorts to Euler or Milstein discretization. In addition, the law of a quantity of interest (QOI) resulting from the solution of a PDE associated to random input parameters, which cannot be solved exactly and needs to be numerically approximated. Examples satisfying 1.-3. include for example Bayesian instances of the above, where one updates the prior probability distribution based on noisy data to obtain the posterior conditional on the observed data (e.g., [48]) or general models where approximate Bayesian computation (e.g., [67]) must be used.

## 1.1 Monte Carlo and Multilevel Monte Carlo

For now, let us assume that only 1. & 2. apply. In this context, one could, for instance, sample  $U^1, \dots, U^N$  i.i.d. from  $\pi_L$  and one could use the standard Monte Carlo estimator  $\frac{1}{N} \sum_{i=1}^N \varphi(u^i)$ , which approximates  $\mathbb{E}_{\pi_L}[\varphi(U)] = \int_{\mathbb{E}} \varphi(u) \pi_L(du)$ . To explain what will follow, we will suppose the following.

**Assumption 1.1** (Cost and discretization error). *There are  $\alpha, \zeta > 0$  such that*

- *The cost of simulating one sample is  $\mathcal{O}(h_L^{-\zeta})$ .*
- *The bias is of order  $\mathcal{O}(h_L^\alpha)$ .*

This scenario would occur under an Euler discretization of a suitably regular SDE and for appropriate  $\varphi$ , with  $\alpha = \zeta = 1$ . For simplicity suppose,  $h_L = 2^{-L}$ . Consider the mean square error (MSE) associated to the Monte Carlo estimator

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

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

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

which is the standard variance plus bias squared. Now let  $1 > \epsilon > 0$  be given, and suppose one wants to control the MSE, e.g., so that it is  $\mathcal{O}(\epsilon^2)$ . One begin 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 (MLMC) method [39, 40, 46] is designed to improve over the cost of Monte Carlo. As above, suppose  $h_l = 2^{-l}$ . The idea is to consider a hierarchy,  $\infty > h_1 > \dots > h_L > 0$  and consider the respresentation

$$\mathbb{E}_{\pi_L}[\varphi(U)] = \sum_{l=1}^L \{\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}\}[\varphi(U)], \quad (3)$$

where for  $1 \leq l \leq L$ ,  $\mathbb{E}_{\pi_l}$  is the expectation w.r.t.  $\pi_l$  (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  $\pi_l$  one is only interested in a marginal on  $\mathbf{E}$ , even if the entire space must be enlarged to facilitate the biased approximation. So, for instance,  $\pi_l$  may be defined on a larger space than  $\mathbf{E}$ , but it admits a marginal on  $\mathbf{E}$  which approaches  $\pi$  as  $l$  grows. Furthermore, it is explicitly assumed that the cost of sampling or evaluating  $\pi_l$  grows with  $l$ ; again this would occur in most applications mentioned above. To approximate the first term in the summation of (3), one samples  $U^1(1), \dots, U^{N_1}(1)$  i.i.d. from  $\pi_1$  and one uses the standard Monte Carlo estimator  $\frac{1}{N} \sum_{i=1}^N \varphi(u^i(1))$ . For the remainder of the terms  $2 \leq l \leq L$ , we suppose that it is possible to sample a (dependent) *coupling* of  $(\pi_l, \pi_{l-1})$  with samples  $(U_l(l), U_{l-1}(l))$  such that the following holds.

**Assumption 1.2** (Variance). *There is a  $\beta > 0$  such that the variance w.r.t. the coupling of  $(\pi_l, \pi_{l-1})$ ,*

$$\text{Var}_{(\pi_l, \pi_{l-1})}[\varphi(U_l(l)) - \varphi(U_{l-1}(l))] = \mathcal{O}(h_l^\beta).$$

Note that by coupling, we mean that  $U_l(l) \sim \pi_l$  and  $U_{l-1}(l) \sim \pi_{l-1}$  (the random variables are generally dependent). Couplings which satisfy the above bullet exist, for instance, in the context of SDE. If the SDE is suitably regular and is approximated by Euler method, then  $\beta = 1$ , or  $\beta = 2$  if the diffusion coefficient is constant. In order to approximate the summands in (3) for  $2 \leq l \leq L$ , draw  $N_l$  i.i.d. samples  $(U_l^1(l), U_{l-1}^1(l)), \dots, (U_l^{N_l}(l), U_{l-1}^{N_l}(l))$  from the coupling  $(\pi_l, \pi_{l-1})$ , and use the unbiased estimator

$$\frac{1}{N_l} \sum_{i=1}^{N_l} \{\varphi(u_l^i(l)) - \varphi(u_{l-1}^i(l))\} \approx \{\mathbb{E}_{\pi_l} - \mathbb{E}_{\pi_{l-1}}\}[\varphi(U)].$$

The multilevel estimator is thus

$$\frac{1}{N_1} \sum_{i=1}^{N_1} \varphi(u_1^i(1)) + \sum_{l=2}^L \left( \frac{1}{N_l} \sum_{i=1}^{N_l} \{\varphi(u_l^i(l)) - \varphi(u_{l-1}^i(l))\} \right).$$

One can analyze the MSE as above. It is equal to

$$\frac{1}{N_1} \text{Var}_{\pi_1}[\varphi(U)] + \sum_{l=2}^L \frac{1}{N_l} \text{Var}_{(\pi_l, \pi_{l-1})}[\varphi(U_l(l)) - \varphi(U_{l-1}(l))] + (\mathbb{E}_{\pi_L}[\varphi(U)] - \mathbb{E}_{\pi}[\varphi(U)])^2,$$

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

$$L \propto -\frac{\log(\epsilon)}{\alpha \log(2)}. \quad (4)$$

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

$$\sum_{l=1}^L \frac{h_l^\beta}{N_l} \propto \epsilon^2.$$

This constrained optimization problem is solved in [39] 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

$$N_l = \epsilon^{-2} h_l^{(\beta+\zeta)/2} K_L, \quad (5)$$

where  $K_L = \sum_{l=1}^L h_l^{(\beta-\zeta)/2}$ . Note that  $K_L$  may depend upon  $L$ , depending upon the values of  $\beta, \zeta$ . 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 1.1, and some analogue of the bias from Assumption 1.1 will determine  $L$  as defined in (4). We will then require rates on different quantities analogous to Assumption 1.2, in order to ensure the choice of  $N_l$  in (5) is optimal.

## 1.2 Methodology Reviewed

In the above section, we have supposed only the points 1. and 2. However, in this article we are considering all three points. In other words, it is not possible to exactly sample from any of  $\pi_L$  or  $\pi_1, \dots, \pi_{L-1}$ . One of the critical ingredients of the MLMC method is sampling dependent couples of the pairs  $(\pi_l, \pi_{l-1})$ , which one might argue is even more challenging than sampling from  $\pi_L$  for a single given  $L$ . In the context of interest, one might use Markov chain Monte Carlo (MCMC – see e.g., [76, 78]) or Sequential Monte Carlo (SMC – see e.g., [23, 24, 32, 31]) to overcome the challenges of not being able to sample  $\pi_L$ . However, a simple procedure of trying to approximate the ML identify (3) by sampling independent MCMC chains targeting  $\pi_1, \dots, \pi_L$  would seldom lead to improvement over just sampling from  $\pi_L$ . So, in such contexts where also using the MLMC approach makes sense, the main issue is how can one utilize such methodology so that one reduces the cost relative to exact sampling from  $\pi_L$ , for a given MSE. 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 MSE. For instance, the analysis of MCMC and SMC rely upon techniques in Markov chains (e.g., [68, 78]) and Feynman-Kac formulae (e.g., [23, 24]). We highlight these techniques during our review.

### 1.3 Structure

This article is structured as follows. In Section 2, we give a collection of motivating examples from applied mathematics and statistics, for which the application of multilevel methods would make sense, and are of interest from a practical perspective. These examples are sufficiently complex that standard independent sampling is not currently possible, but advanced simulation methods such as those described in the previous subsection can be used. In Section 3, a short review of some of the computational methods for which this review is focussed on is given. In Section 4, we review several methods which have been adopted in the literature to date, mentioning the benefits and drawbacks of each approach. In Section 5, 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., [75]). An effort is of course made to discuss as much work as possible that exists under the umbrella of advanced MLMC methods.

## 2 Motivating Examples

### 2.1 Bayesian Inverse Problems

We consider the following example as it is described in [8] (see also [48] 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  $\Omega$  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  $\Omega$ :

$$-\nabla \cdot (\hat{u} \nabla p) = f \quad \text{in } \Omega, \quad (6)$$

$$p = 0 \quad \text{on } \partial\Omega, \quad (7)$$

where

$$\hat{u}(x) = \bar{u}(x) + \sum_{k=1}^K u_k \sigma_k \phi_k(x). \quad (8)$$

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  $u$ . Assume that  $\bar{u}, \phi_k \in C^\infty$  for all  $k$  and that  $\|\phi_k\|_\infty = 1$ . In particular, assume  $\{\sigma_k\}_{k=1}^K$  decay with  $k$ . The state space is  $\mathbf{E} = \prod_{k=1}^K [-1, 1]$ . 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 (6) is uniformly elliptic. Let  $p(\cdot; u)$  denote the weak solution of (6) for parameter value  $u$ . Define the following vector-valued function

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

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

$$Y = \mathcal{G}(p) + \xi, \quad \xi \sim \mathcal{N}(0, \Gamma), \quad \xi \perp u, \quad (9)$$

where  $\mathcal{N}(0, \Gamma)$  denotes the Gaussian random variable with mean 0 and covariance  $\Gamma$ , and  $\perp$  denotes independence. The unnormalized density for  $u \in \mathbf{E}$  is then is given by:

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

### 2.1.1 Approximation

Consider the triangulated domains (with sufficiently regular triangles)  $\{\Omega^l\}_{l=1}^\infty$  approximating  $\Omega$ , where  $l$  indexes the number of nodes  $d_l \propto h_l^{-D}$ , for triangulation diameter  $h_l$ , so that we have  $\Omega^1 \subset \dots \subset \Omega^l \subset \Omega^\infty := \Omega$ . Furthermore, consider a finite element discretization on  $\Omega^l$  consisting of  $H^1$  functions  $\{\psi_\ell\}_{\ell=1}^{d_l}$ . Denote the corresponding space of functions of the form  $\varphi = \sum_{\ell=1}^{d_l} v_\ell \psi_\ell^l$  by  $V^l$ , and notice that  $V^1 \subset V^2 \subset \dots \subset V^l \subset V$ . By making the further Assumption 7 of [48] that the weak solution  $p(\cdot; u)$  of (6)-(7) 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 distributions 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. But it is not obvious how one can leverage the MLMC approach for this application. Several strategies are suggested later on in the article.

## 2.2 Partially Observed Diffusions

The following model is considered, as described in [54, 59]. Consider the partially-observed diffusion process:

$$dU_t = a(U_t)dt + b(U_t)dW_t, \quad (10)$$

with  $U_t \in \mathbb{R}^d$ ,  $t \geq 0$ ,  $U_0$  given  $a : \mathbb{R}^d \rightarrow \mathbb{R}^d$  (denote the  $j^{\text{th}}$ -element as  $a^j(U_t)$ ),  $b : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$  (denote the  $j^{\text{th}}, k^{\text{th}}$ -element as  $b^{j,k}(U_t)$ ) and  $\{W_t\}_{t \in [0, T]}$  a Brownian motion of  $d$ -dimensions. Some assumptions are made in [54, 59] to ensure that the diffusion has an appropriate solution; see [54, 59] for details.

It will be assumed that the data are regularly spaced (i.e., in discrete time) observations  $y_1, \dots, y_n$ , with  $y_k \in \mathbb{R}^m$ . It is 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 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

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

where  $Q^\infty(u_{(i-1)}, u)$  is the transition density of the diffusion process as a function of  $u$ , i.e., the density of the solution  $U_1$  of Eq. (10) 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  $k \in \{1, \dots, n\}$ , the objective is to approximate the filter

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

with  $u_{1:k} = (u_1, \dots, u_k)$  and  $y_{1:k} = (y_1, \dots, y_k)$ . The shorthand notation  $\pi^k(\cdot) = \pi(\cdot | y_{1:k})$  is used above and in what follows. Note that we will use  $\pi_\infty$  as the notation for measure and density, with the use clear from the context. It is also of interest, to estimate the normalizing constant, or marginal likelihood

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

Note that the filtering problem has many applications in engineering, statistics, finance, and physics (e.g., [12, 21, 31] 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. Even if one knows  $Q^\infty$  pointwise, up-to a non-negative unbiased estimator, and/or can sample exactly from the associated law, advanced computational methods, such as particle filters (e.g., [32, 37]) – an exchangeable term for SMC when used in a filtering context – are often adopted in order to estimate the filter. In the setting considered in this paper, it is assumed that one cannot

- evaluate  $Q^\infty$  pointwise, up-to a non-negative unbiased estimator ;
- sample from the associated distribution of  $Q^\infty$ .

$Q^\infty$  and its distribution must be approximated by some discrete time-stepping method [64] (for time-step  $h_l = 2^{-l}$ ).

For simplicity and illustration, Euler’s method [64] will be considered. One has

$$\begin{aligned} U_k^l(m+1) &= U_k^l(m) + h_l a(U_k^l(m)) + \sqrt{h_l} b(U_k^l(m)) \xi_k(m), \\ \xi_k(m) &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}_d(0, I_d), \end{aligned} \tag{11}$$

for  $m = 0, \dots, k_l - 1$ , 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  $k \in \{1, \dots, n\}$  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

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

First, we consider how this task can be performed using SMC and 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  $G_\theta$  is the likelihood function above. If one assumes a prior  $\pi_\theta$  on  $\theta$ , then one might be interested in, for  $k$  fixed:

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

and the associated discretization. We consider how the latter task is possible using MCMC for a given level and how that is extended for the MLMC case.

## 3 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 the basic ideas, with pointers to the literature, where relevant. We first consider basic MCMC in Section 3.1.

Then, in Section 3.2, SMC is discussed in context of particle filtering. Here, a certain SMC algorithm, SMC samplers (e.g., [25]) is approached, which uses MCMC algorithms within it. In Section 3.3, we discuss particle MCMC [1], which combines SMC within MCMC proposals. Finally, in Section 3.4, we discuss ensemble Kalman filter approaches.

### 3.1 Markov chain Monte Carlo

We consider a target probability  $\pi_L$  on measurable space  $(\mathbf{E}, \mathcal{E}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ , with  $\mathcal{B}(\mathbb{R}^d)$  the Borel sets on  $\mathbb{R}^d$ ; extensions to other spaces is simple, but it is omitted for brevity. We write the target density of  $\pi_L$  w.r.t. Lebesgue measure as  $\pi_L$  also, following standard practice.

The idea of MCMC is to build an ergodic Markov kernel of invariant measure  $\pi_L$  or, at least, that  $\pi_L$  is a marginal of the invariant measure of the Markov chain – we concentrate on the former case. That is, samples of the Markov chain  $U^1, \dots, U^N$  have the property that for  $\varphi : \mathbf{E} \rightarrow \mathbb{R}$ ,  $\varphi$   $\pi$ -integrable, one has that estimates of the form

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

will converge almost surely as  $N \rightarrow \infty$  to  $\mathbb{E}_{\pi_L}[\varphi(U)]$ .

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 ( $u^{i+1} = u^i$ ) 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  $\Sigma$ , is often chosen so that the acceptance rate is about 0.234 [79], although there are adaptive methods for doing this; see [2, 44].

The algorithm mentioned here is the most simple approach. The ideas can be extended to alternative proposals, Langevin [77], Hamiltonian Monte Carlo ([34], see also [70]), pre-conditioned Crank Nicholson (e.g., [20] and the references therein). The algorithm can also be used in infinite dimensions (e.g., Hilbert spaces [20]) and each dimension need not be updated simultaneously - for example, one can use Gibbs and Metropolis-within-Gibbs approaches; see [76, 78] for some coverage. There are also population-based methods (e.g., [53] and the references therein) and non-reversible approaches (e.g., [72, 73, 11] 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 well established convergence theory; see [68, 78] 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., [79]) and can be dimension free when there is a well-defined limit as  $d$  grows [9, 20]. Note that finally, it is not simple to use ‘standard’ MCMC to estimate normalizing constants.

### 3.2 Sequential Monte Carlo

Here we will consider the standard SMC algorithm for filtering, also called the particle filter in the literature. To assist our discussion, we focus on the filtering density from Section 2.2.1

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

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  $q_1(u_1), q_2(u_1, u_2), q_3(u_2, u_3), \dots$ , 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 as follows:

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

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

- **Iterate:** Set  $k = k + 1$ ,

– Resample  $(\hat{u}_{k-1}^1, \dots, \hat{u}_{k-1}^N)$  according to the weights  $(w_{k-1}^1, \dots, w_{k-1}^N)$ .

– Sample  $u_k^i | \hat{u}_{k-1}^i$  from  $q_k$ , for  $i \in \{1, \dots, N\}$ , and evaluate the weight

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

The resampling step can be performed using a variety of schemes, such as systematic, multinomial residual etc; the reader is referred to [32] for more details. For  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ , and  $\varphi \pi_L^k$ -integrable, one has the consistent estimate:

$$\sum_{i=1}^N w_k^i \varphi(u_k^i) \approx \mathbb{E}_{\pi_L^k} [\varphi(U_k)]. \quad (13)$$

In addition, the marginal likelihood is unbiasedly [23] estimated by

$$\hat{Z}_L^k := \prod_{i=1}^k \left( \frac{1}{N} \sum_{j=1}^N \frac{G(u_i^j, y_i) Q^L(\hat{u}_{i-1}^j, u_i^j)}{q_i(\hat{u}_{i-1}^j, u_i^j)} \right) \approx Z_L^k, \quad (14)$$

with the abuse of notation that  $\hat{u}_0^i = u_0$ . In principle, for  $\varphi : \mathbb{R}^{kd} \rightarrow \mathbb{R}$ , and  $\varphi \pi_L^k$ -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 [32, 62].

The algorithm given here is one of the most basic and many modifications can enhance the performance of this algorithm; see [23, 24, 32, 62] for some ideas. The theoretical validity of the method has been established in many works; see e.g., [13, 17, 23, 24, 30]. 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 [24] and the references therein. One of the main issues with particle filters/SMC methods in this context is that they do not perform well in high dimensions (i.e., for large  $d$ ) often having an exponential cost in  $d$  [86]. Note however, that if there is a well-defined limit as  $d$  grows, SMC methods can be designed to perform quite well in practice on a finite time horizon (see e.g., [61] and

the next section). There have been some methods developed for high-dimensional filtering (e.g., [7, 69, 74]), however, they are only useful for a small class of models.

### 3.2.1 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 measures  $\pi_2, \dots, \pi_L$ . This is possible if the densities are known up-to a constant (and potentially also a non-negative unbiased estimator - although this is not considered at the moment), simply by using using MCMC. The SMC sampler algorithm (e.g., [25]) 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 (assumed to exist w.r.t. a common dominating measure) of  $\pi_1, \dots, \pi_L$  are written  $\kappa_1, \dots, \kappa_L$ . To ease the notational burden, we suppose one can sample from  $\pi_1$ , but this is not necessary.

The algorithm is as follows:

- **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  $i \in \{1, \dots, N\}$ ,
$$w_l^i = \left( \frac{\kappa_l(u_{l-1}^i)}{\kappa_{l-1}(u_{l-1}^i)} \right) \left( \sum_{j=1}^N \frac{\kappa_l(u_{l-1}^j)}{\kappa_{l-1}(u_{l-1}^j)} \right)^{-1}.$$
  - Sample  $u_l^i | \hat{u}_{l-1}^i$  from  $M_l$  for  $i \in \{1, \dots, N\}$ .

One can estimate expectations w.r.t.  $\pi_l$ , for  $\varphi : \mathbf{E} \rightarrow \mathbb{R}$ ,  $\varphi$   $\pi_l$ -integrable. The consistent estimator

$$\frac{1}{N} \sum_{i=1}^N \varphi(u_l^i) \approx \mathbb{E}_{\pi_l}[\varphi(U)]$$

converges almost surely as  $N \rightarrow \infty$ . In addition, for any  $l \geq 2$ , we have the unbiased estimator

$$\prod_{\ell=2}^l \left( \frac{1}{N} \sum_{i=1}^N \frac{\kappa_\ell(u_{\ell-1}^i)}{\kappa_{\ell-1}(u_{\ell-1}^i)} \right) \approx Z_l / Z_1,$$

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

The basic algorithm goes back to at least [51]. Several versions are found in [16, 71], with a unifying framework in [25] and a rediscovery in [15]. Subsequently, several refined and improved versions of the algorithm have appeared [19, 26, 47, 58, 81], including those which allow algorithmic parameters to be set adaptively, that is, without user specification.

Contrary to particle filters, when  $\mathbf{E} = \mathbb{R}^d$ , this method indeed performs quite well w.r.t. the dimension  $d$  with only polynomial cost in  $d$ ; see [3, 4]. Whilst the underlying theory for this algorithm is very similar to particle filters and it is covered in [23, 24], there are some additional results in [5, 82, 90]. In particular, [5] establish that when one updates parameters adaptively, such as in [58, 81], then the algorithm is still theoretically correct. The method is very useful in the following scenarios: (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.

### 3.3 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

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

Sampling this distribution is notoriously challenging. One recent method and its extensions can be considered the gold standard as we describe now. Note that the SMC algorithm (in Section 3.2) used below uses the Euler discretized dynamics as the proposal.

The particle marginal Metropolis-Hastings (PMMH) algorithm of [1] proceeds as follows.

- **Initialize.** Set  $i = 0$  and sample  $\theta^0$  from the prior. Given  $\theta^0$  run the SMC algorithm in Section 3.2 and record the estimate of  $\widehat{Z}_{L,\theta^0}^k$  from eq. (14).
- **Iterate:**
  - Set  $i = i + 1$  and propose  $\theta'$  given  $\theta^{i-1}$  from a proposal  $r(\theta^{i-1}, \cdot)$ .
  - Given  $\theta'$  run the SMC algorithm in Section 3.2 and record the estimate  $\widehat{Z}_{L,\theta'}^k$ .

– Set  $\theta^i = \theta'$  with probability

$$\min \left\{ 1, \frac{\widehat{Z}_{L,\theta'}^k \pi_\theta(\theta') r(\theta', \theta^{i-1})}{\widehat{Z}_{L,\theta^{i-1}}^k \pi_\theta(\theta^{i-1}) r(\theta^{i-1}, \theta')} \right\}$$

otherwise  $\theta^i = \theta^{i-1}$ .

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

with

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

Note that this is consistent as  $N$  grows, in the sense that it recovers the true expectation with probability 1, under minimal conditions. 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 them has been discussed in [1, 33], for example.

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 [29, 85].

### 3.4 Ensemble Kalman filter

The idea of the ensemble Kalman filter (EnKF) is to approximate the filtering distribution (12) using an ensemble of particles and their sample covariance [35]. As such they are sometimes also referred to as sequential Monte Carlo methods, but we believe it is important to distinguish them from the methods described in subsection 3.2. The observations are incorporated as though the process were linear and Gaussian, hence requiring only the covariance approximation. Hence, the method is consistent only in the case of a linear Gaussian model [65]. However, it is robust even in high dimensions [36] and can be tuned to perform reasonably well in tracking and forecasting. It has therefore become very popular among practitioners.

It will be assumed here for simplicity that the observation selection function is given by:

$$G(u_i, y_i) \propto \exp\left(-\frac{1}{2} |\Gamma^{-\frac{1}{2}} (H u_i - y_i)|^2\right), \quad (15)$$

where  $H$  is a linear operator. The linearity assumption is without any loss of generality since if  $h$  is nonlinear, one can always extend the system to  $(u, v)^\top$ , where  $v = h(u)$ .

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

$$\begin{cases} \text{Prediction} & \left\{ \begin{array}{l} u_{j+1}^{(n)} \sim Q^L(\hat{u}_j^{(n)}, \cdot), \quad n = 1, \dots, N, \\ m_{j+1} = \frac{1}{N} \sum_{n=1}^N u_{j+1}^{(n)}, \\ C_{j+1} = \frac{1}{N-1} \sum_{n=1}^N (u_{j+1}^{(n)} - m_{j+1})(u_{j+1}^{(n)} - m_{j+1})^T. \end{array} \right. \\ \\ \text{Analysis} & \left\{ \begin{array}{l} S_{j+1} = HC_{j+1}H^T + \Gamma, \\ K_{j+1} = C_{j+1}H^T S_{j+1}^{-1}, \\ \hat{u}_{j+1}^{(n)} = (I - K_{j+1}H)u_{j+1}^{(n)} + K_{j+1}y_{j+1}^{(n)}, \quad n = 1, \dots, N, \\ y_{j+1}^{(n)} = y_{j+1} + \xi_{j+1}^{(n)}, \quad n = 1, \dots, N. \end{array} \right. \end{cases}$$

Here  $\xi_j^{(n)}$  are i.i.d. draws from  $N(0, \Gamma)$ . Perturbed observation refers to the fact that each particle sees an observation perturbed by an independent draw from  $N(0, \Gamma)$ . This procedure ensures the Kalman Filter is obtained in the limit of infinite ensemble in the linear Gaussian case [65]. Notice that the ensemble is not prescribed to be Gaussian, even though it is updated as though it were, so the limiting target is some non-Gaussian  $\hat{\pi}_L^k$ , which is in general not equal to the density defined by (12) (see e.g., [66]).

## 4 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.

### 4.1 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 measurable 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 3.2.1.

In this scenario [8] (see also [6, 27, 28, 52]) investigate the simple modification

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

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 as the targets  $\pi_l$  and  $\pi_{l-1}$  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 1.1) 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 3.2.1 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 [8] is to simply run the algorithm of Section 3.2.1, except at step  $l$  one resamples  $N_{l+1} < N_l$  particles, where the schedule of numbers  $N_{0:L-1}$  is chosen using a similar principle as for standard MLMC. The identity (16) can be approximated via:

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

Note that the algorithm need only be run up-to level  $L - 1$ . [8] not only show that this is consistent, but also give a general MLMC theorem using the theory in [23] with some additional work and assumptions (which are relaxed in [28]). In the context of the example of Section 2.1, the authors show that the work to compute expectations relative to standard SMC samplers (as in Section 3.2.1) is reduced to achieve a given MSE, under the following assumptions.

**Assumption 4.1** (MLSMC samplers). *There is a  $\varepsilon > 0$  such that for all  $l = 1, \dots, L$ ,  $u, v \in \mathbf{E}$ , and  $A \in \mathcal{E}$*

- $\varepsilon < \kappa_l(u) < \varepsilon^{-1}$  ;

- $\varepsilon M_l(v, A) < M_l(u, A) < \varepsilon^{-1} M_l(v, A)$ , where we recall from subsection 3.2.1 that  $M_l$  is the Markov kernel with invariant measure proportional to  $\kappa_l$ , used to mutate the updated population of samples at step  $l$ .

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  $h_l^\beta$ , under suitable assumptions and in an appropriate norm, just as the variance terms in the coupling of Section 1.1. 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  $\alpha$  and  $\beta$  for are summarized in table 1, and the convergence rate is illustrated in Figure 1.

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

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

The method of [8] has been extended to the computation of normalizing constants [28] and has been applied to other examples, such as non-local equations [52] and approximate Bayesian computation [60]. The method has also been extended to the case that the accuracy of the approximation improves as the dimension of the target grows; see [6].

The importance sampling idea has also been considered in other articles such as [48, 83]. In [48], the change of measure appears in the context of MCMC and is not too dissimilar to the one presented in (16), although it is with respect to the higher level in the couple, and multiple discretization indices are considered as well (see also [45]). In [83], the numerator and denominator of (1), arising from the form (2), are approximated independently in terms of expectation w.r.t. the prior, in a Bayesian set up, and they find reasonably decent

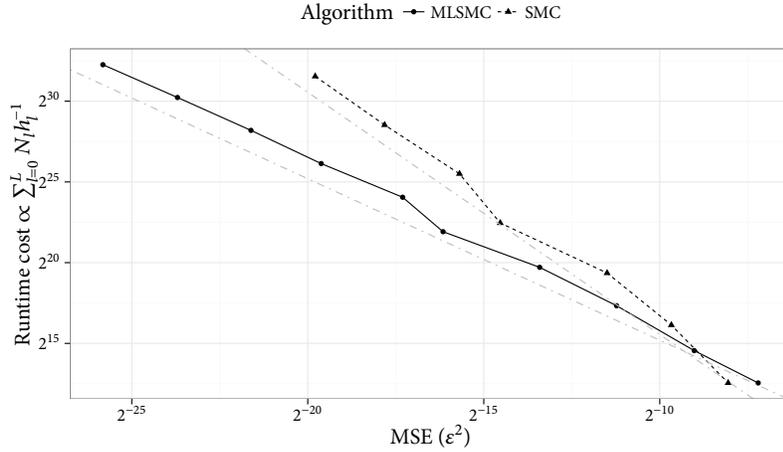


Figure 1: Computational cost against mean squared error for MLSMC sampler in comparison to SMC sampler for the Bayesian inverse problem from subsection 2.1 [8].

performance. This corresponds to a 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. In general such a procedure is not advisable unless the prior is very informative; one expects that the estimators for each of the numerator and the denominator will have very high variance.

## 4.2 Approximate Coupling

The case of Section 2.2.2 will be considered here in order to illustrate this idea. In particular, there is a stochastic process that is partially observed. It is assumed that the dynamics of the associated discretized processes can be easily coupled. 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  $(\pi_l, \pi_{l-1})$ . 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 [55] (see also [56]).

We begin by considering:

$$\pi_l^k(\theta, u_{1:k}) \propto \pi_\theta(\theta) \prod_{i=1}^k G_\theta(u_i, y_i) Q_\theta^l(u_{(i-1)}, u_i)$$

for two levels  $l, l-1 \geq 1$ . We know that one can construct a good coupling of the two discretized kernels  $Q_\theta^l, Q_\theta^{l-1}$  for any fixed  $\theta$  by sampling the finer Gaussian increments and concatenating them for the coarser discretization (e.g., [39] or as written in [54, 59]). More precisely, given  $\check{u}_{i-1} = (u_{i-1}, \bar{u}_{i-1}) \in \mathbb{R}^{2d}$  and  $\theta \in \Theta$ , there is a Markov kernel  $\check{Q}_\theta^{l,l-1}$  such that for any  $A \in \mathcal{B}(\mathbb{R}^d)$

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

and

$$\int_{\mathbb{R}^d \times A} \check{Q}_\theta^{l,l-1}(\check{u}_{i-1}, \check{u}_i) d\check{u}_i = \int_A Q_\theta^{l-1}(u_{i-1}, u_i) du_i.$$

Note that under the coupling considered, the discretized processes are not independent.

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

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

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

$$\check{G}_\theta(\check{u}_i, y_i) = \max\{G_\theta(\bar{u}_i, y_i), G_\theta(u_i, y_i)\}. \quad (18)$$

This will be explained below. Let  $\varphi : \Theta \times \mathbb{R}^{2kd} \rightarrow \mathbb{R}$  be  $\pi_l^k$  and  $\pi_{l-1}^k$ -integrable. Then we have (supressing the conditioning on  $y_{1:k}$  in the expectations)

$$\begin{aligned} & \mathbb{E}_{\pi_l^k}[\varphi(\theta, U_{1:k})] - \mathbb{E}_{\pi_{l-1}^k}[\varphi(\theta, U_{1:k})] = \\ & \frac{\mathbb{E}_{\tilde{\pi}_{l-1:l}^k}[\varphi(\theta, \bar{U}_{1:k}) \bar{H}_\theta(\check{U}_{1:k})]}{\mathbb{E}_{\tilde{\pi}_{l-1:l}^k}[\bar{H}_\theta(\check{U}_{1:k})]} - \frac{\mathbb{E}_{\tilde{\pi}_{l-1:l}^k}[\varphi(\theta, \underline{U}_{1:k}) \underline{H}_\theta(\check{U}_{1:k})]}{\mathbb{E}_{\tilde{\pi}_{l-1:l}^k}[\underline{H}_\theta(\check{U}_{1:k})]}, \end{aligned} \quad (19)$$

where

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

The difference can then be approximating by sampling from  $\check{\pi}_{l-1:l}^k$ , e.g., by using the PMMH<sup>1</sup> from Section 3.3. This is done independently for each summand in the ML identity, with the first summand (the coarsest discretization) sampled by PMMH.

We now explain the idea in more detail. 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 and, as in Section 4.1, use importance sampling. Just as argued in that section, the associated weights of the importance sampling, that is, the terms  $(\bar{H}_\theta, \underline{H}_\theta)$  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 measure will remain bounded uniformly in time. Hence the reason for its selection as 18. [55] 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 [55] do not consider the dependence on the time parameter  $k$ , and that is something that should be addressed. In particular, the required assumptions in this context are given below.

**Assumption 4.2** ((PMMH using MLMC)). *There is a  $\varepsilon > 0$  and probability  $\nu$  over  $\Theta \times \mathbb{R}^{2kd}$ , such that for all  $l = 1, \dots, L$ ,  $u \in \mathbb{R}^d$ ,  $y \in \mathbb{R}^m$ ,  $\theta \in \Theta$ ,  $A \in \sigma(\Theta)$ , and any  $\varphi : \Theta \times \mathbb{R}^{2kd}$  bounded and Lipschitz, and  $w \in \mathbb{W}$  (the space of all auxiliary variables involved in the higher-dimensional chain), the following hold*

- $\varepsilon < G_\theta(u, y) < \varepsilon^{-1}$  ;
- $Q^l(u, v) > \varepsilon$  ;

---

<sup>1</sup>More precisely, one samples from a suitably extended measure, as described in [1].

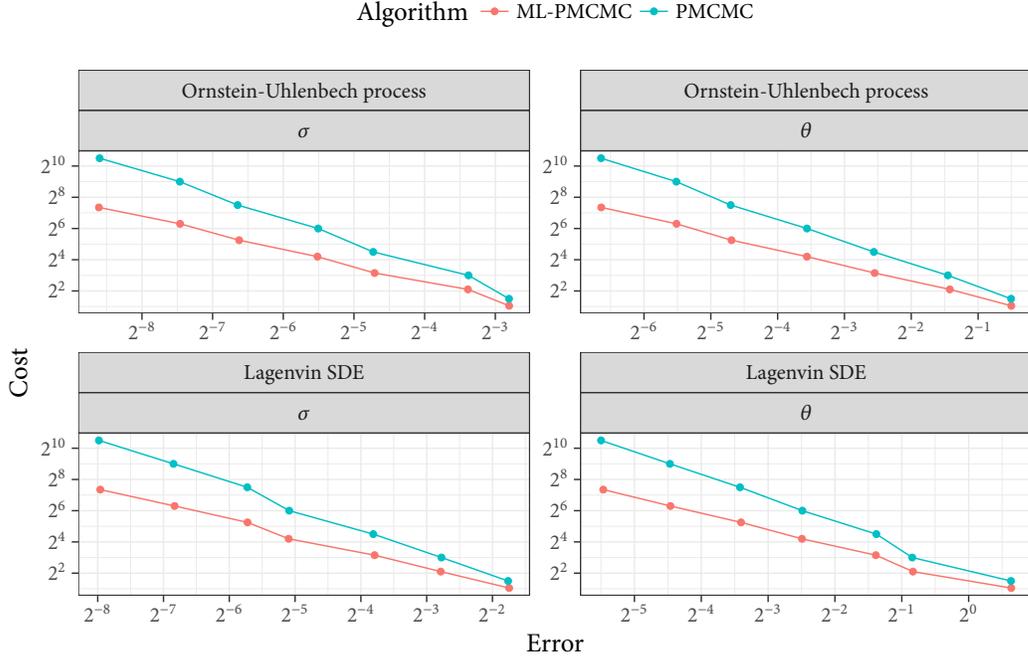


Figure 2: Cost vs. MSE for the inference of 2 parameters for each of 2 SDE examples [55].

- The final Metropolis kernel  $K$  on the extended space satisfies  $\int_{\mathbb{W}} \varphi(w')K(w, dw') \geq \varepsilon \int_{\Theta \times \mathbb{R}^{2kd}} \varphi(v)\nu(dv)$ ,

The quantities for which rates  $\alpha$  and  $\beta$  need to be obtained in this context are given in table 2, and the results are illustrated in Figure 2 for 2 example SDE of the type introduced in section 2.2.2, each with 2 unknown parameters  $\sigma$  and  $\theta$ .

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

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. Indeed that may be considered as an extreme example of approximate coupling in which  $\check{G}_\theta(\check{u}_i, y_i) = G_\theta(\underline{u}_i, y_i)\delta_{\bar{u}_i, \underline{u}_i}$ . 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., [38, 80]), but it is still an issue. 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. If one elects to use an SMC sampler to approximate (19), then PMMH kernels can be used as described in [19]. The algorithm of [19] can also be used for dynamic (sequential) inference on the parameter, in an MLMC context. The same principle as described above can be generalized to MCMC, and this has been done in the work [57].

### 4.3 Coupling Algorithms

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

#### 4.3.1 Coupling MCMC

We begin by considering the method in [63], which is a Markov chain approach. Consider two probability measures  $(\pi_l, \pi_{l-1})$ , where the support of  $\pi_{l-1}$  is  $\mathbf{E}$  and the support of  $\pi_l$  is  $\mathbf{E} \times \mathbf{U}$ . We focus on computing  $\mathbb{E}_{\pi_l}[\varphi_l(U)] - \mathbb{E}_{\pi_{l-1}}[\varphi_{l-1}(U)]$  where  $\varphi_l : \mathbf{E} \times \mathbf{U} \rightarrow \mathbb{R}$  and  $\varphi_{l-1} : \mathbf{E} \rightarrow \mathbb{R}$  are  $\pi_l$  and  $\pi_{l-1}$  integrable.

Suppose we have a current state  $(u_l, u_{l-1}) \in \mathbf{E} \times \mathbf{U} \times \mathbf{E}$ . The approach consists first by sampling from  $\pi_{l-1}$  exactly. Given this sample and the current  $u_l \in \mathbf{E}$  a new state is proposed from a proposal  $r$  and is accepted or rejected according to the standard Metropolis-Hastings method. It is clear that the samples are coupled and seemingly that they have the correct invariant distribution. [63] show that this approach indeed can obtain the advantages of MLMC estimation for some examples. As noted in that article, supposing exact sampling from  $\pi_{l-1}$  is feasible, is not realistic and the authors propose a subsampling method to assist

with this. See [63] for more details.

Before concluding this section, we mention the recent related works [50, 84, 88]. The work [88] (see also [41]) 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.

### 4.3.2 Coupling Particle Filters

We consider the context of Section 2.2.1 of filtering a discretely and partially observed diffusion processes. We describe an approach in [54] (see also [42, 50, 59, 84]). For  $l \geq 2$ ,  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $\pi_t^k$  and  $\pi_{t-1}^k$ -integrable we consider approximating the difference  $\mathbb{E}_{\pi_t^k}[\varphi(U_k)] - \mathbb{E}_{\pi_{t-1}^k}[\varphi(U_k)]$  sequentially in time. Some of the notations of Section 4.2 are also used. The parameter  $\theta$  is also dropped from the notations, as it is assumed to be fixed here.

The multilevel particle filter (MLPF) is described as follows. First, for  $l = 1$ , run a particle filter for the coarsest discretization. Now, run the following procedure independently for each  $l \geq 2$ .

**For**  $i = 1, \dots, N_l$ , draw  $(\bar{U}_1^{l,i}, \underline{U}_1^{l,i}) \stackrel{\text{i.i.d.}}{\sim} \check{Q}^{l,l-1}((u_0, u_0), \cdot)$ .

**Initialize**  $k = 1$ . **Do**

- (i) **For**  $i = 1, \dots, N_l$ , draw  $(\bar{I}_k^{l,i}, \underline{I}_k^{l,i})$  according to the coupled resampling procedure below. Set  $k = k + 1$ .
- (ii) **For**  $i = 1, \dots, N_l$ , independently draw  $(\bar{U}_k^{l,i}, \underline{U}_k^{l,i}) | (\bar{u}_{k-1}^{l,i}, \underline{u}_{k-1}^{l,i}) \sim \check{Q}^{l,l-1}((\bar{u}_{k-1}^{l,i}, \underline{u}_{k-1}^{l,i}), \cdot)$ .

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

$$\bar{w}_k^{l,i} = \frac{G(\bar{u}_k^{l,i}, y_k)}{\sum_{j=1}^{N_l} G(\bar{u}_k^{l,j}, y_k)} \quad \text{and} \quad \underline{w}_k^{l,i} = \frac{G(\underline{u}_k^{l,i}, y_k)}{\sum_{j=1}^{N_l} G(\underline{u}_k^{l,j}, y_k)}. \quad (20)$$

Now

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

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

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

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

$$\begin{aligned}\mathbb{P}(\bar{I}_k^{l,i} = j) &= [\bar{w}_k^{l,j} - \bar{w}_k^{l,j} \wedge \underline{w}_k^{l,j}] / (\sum_{s=1}^{N_l} \bar{w}_k^{l,s} - \bar{w}_k^{l,s} \wedge \underline{w}_k^{l,s}), \quad j \in \{1, \dots, N_l\}, \\ \mathbb{P}(\underline{I}_k^{l,i} = j) &= [\underline{w}_k^{l,j} - \bar{w}_k^{l,j} \wedge \underline{w}_k^{l,j}] / (\sum_{s=1}^{N_l} \underline{w}_k^{l,s} - \bar{w}_k^{l,s} \wedge \underline{w}_k^{l,s}), \quad j \in \{1, \dots, N_l\}.\end{aligned}$$

Note that by using the coupled kernel  $\check{Q}^{l,l-1}$ , one is sampling from the exact coupling of the discretized process,  $(\bar{U}_k^{l,i}, \underline{U}_k^{l,i})$ . 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 [18]).

In the work [54], it is shown that

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

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

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

In the case of Euler-Maruyama discretization, [54] it is shown that under suitable assumptions and for finite time the standard choice of  $L$  and  $N_{1:L}$  as in (4) and (5) provides an MSE of  $\mathcal{O}(\epsilon^2)$  for a cost of  $\mathcal{O}(\epsilon^{-2.5})$ . For a particle filter the cost required is  $\mathcal{O}(\epsilon^{-3})$ . 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.

Sufficient assumptions in this case are given by

**Assumption 4.3** (MLPF). *There is a  $\varepsilon > 0$  such that for all  $l = 1, \dots, L$ ,  $u, v \in \mathbb{R}^d$ , and  $y \in \mathbb{R}^m$ , the following hold*

- $\varepsilon < G_\theta(u, y) < \varepsilon^{-1}$  ;
- $Q^l(u, v) > \varepsilon$ .

The quantities for which rates  $\alpha$  and  $\beta$  need to be obtained in this context are given in table 2, and the complexity results are illustrated in Figure 3 for some example SDEs.

Rate parameter	Relevant quantity
$\alpha$	$(\mathbb{E}_{\pi_L^k} - \mathbb{E}_{\pi^k})(\varphi)$
$\beta$	$(\int_{\mathbb{R}^{2d}}  \varphi(\bar{u}_k) - \varphi(\underline{u}_k) ^2 \tilde{\pi}_{l-1:l}^k(\check{u}_k) d\check{u}_k)^2$

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

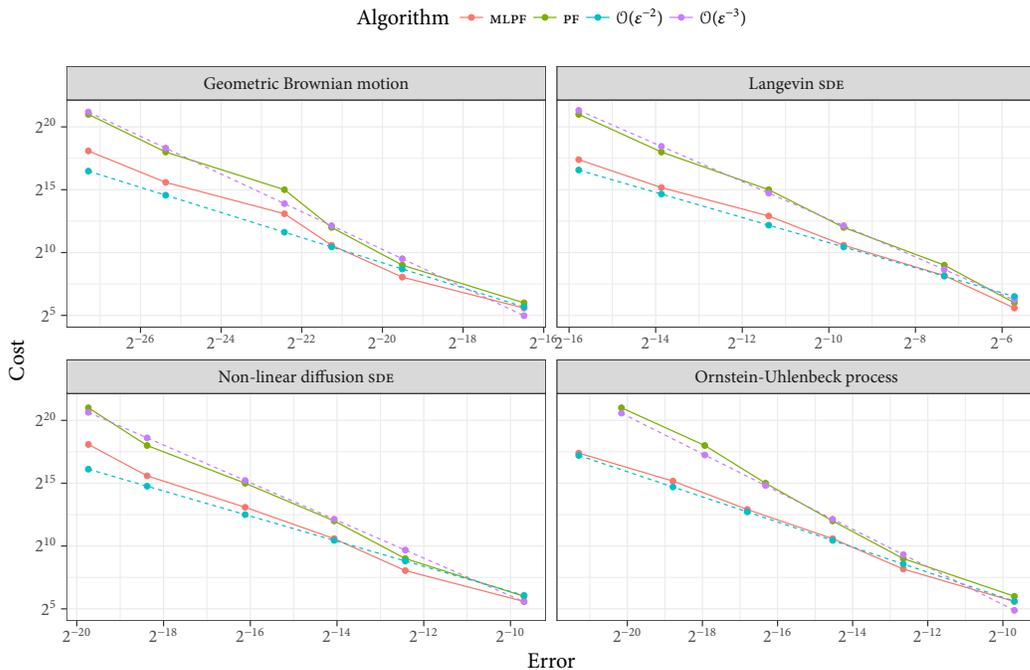


Figure 3: Cost rates as a function of MSE using MLPF in comparison to bootstrap particle filter for various SDE examples of the form (10) [54].

In [59] the approach and results are extended to the case of marginal likelihood estimation. We note that one drawback of the mathematical results in [54, 59] is that they do not consider the time-parameter. Note also that the method does not preserve the standard squared strong convergence of the forward numerical method which defines  $\tilde{Q}^{l,l-1}$ . A power of 1/2 is lost in the rate of convergence as a result of the coupled resampling (see (table 3). In [50, 84] the coupled resampling method is improved by using optimal transportation techniques [89]. Also [42] (see also [43]) obtain empirical results which indicate that more favorable convergence rates may be preserved in certain cases by replacing the resampling

step with a deterministic linear transformation of the current population, derived from the optimal transportation. However, in [42, 43, 50, 84] there are no mathematical results which support the encouraging empirical results; one expects that this is mainly a technical challenge.

### 4.3.3 Coupling the EnKF

As discussed in subsection 3.4, the EnKF targets a different distribution than the filtering distribution in general, which has been denoted  $\hat{\pi}_L^k$ . In between updates, this algorithm proceeds similarly to the MLPF of the previous section, propagating pairs of ensembles for each  $l \geq 2$ . The fundamental difference in the update results in approximations of increments  $\mathbb{E}_{\hat{\pi}_l^k}[\varphi(U_k)] - \mathbb{E}_{\hat{\pi}_{l-1}^k}[\varphi(U_k)]$ . Therefore, the MSE will ultimately depend upon the difference  $\mathbb{E}_{\hat{\pi}_L^k}[\varphi(U_k)] - \mathbb{E}_{\pi_L^k}[\varphi(U_k)]$ , 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 [49], as follows. Recall the functions  $G(u_i, y_i)$  are assumed to take the form given in (15).

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

**Initialize**  $k = 1$ . **Do**

(i) Compute the MLMC covariance estimator [10] :

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

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

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

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

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

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

And draw  $U_k^{1,i} \sim Q^1(\widehat{u}_{k-1}^{1,i}, \cdot)$ .

A sufficient assumption in this case is given by

**Assumption 4.4** (MLEnKF). *The coefficients of (10) are globally Lipschitz and the initial condition is in  $L^p$  for all  $p \geq 2$ .*

The quantities for which rates  $\alpha$  and  $\beta$  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 (10). The work [49] established that slightly modified choices of  $L$  and  $N_{1:L}$  provide MSE at step  $k$  of  $\mathcal{O}(|\log \epsilon|^{2n} \epsilon^2)$  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 [14] 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.

Rate parameter	Relevant quantity
$\alpha$	$(\mathbb{E}_{\widehat{\pi}_L^k} - \mathbb{E}_{\pi^k})(\varphi)$
$\beta$	$\left( \int_{\mathbb{R}^{2d}}  \varphi(\bar{u}_k) - \varphi(\underline{u}_k) ^p \widehat{\pi}_{l-1:l}^k(\bar{u}_k, \underline{u}_k) d\bar{u}_k d\underline{u}_k \right)^{2/p}$

Table 4: The key rates of convergence required for MLEnKF, for all  $p \geq 1$ , where  $\widehat{\pi}_{l-1:l}^k$  denotes the coupled measure resulting from the algorithm above.

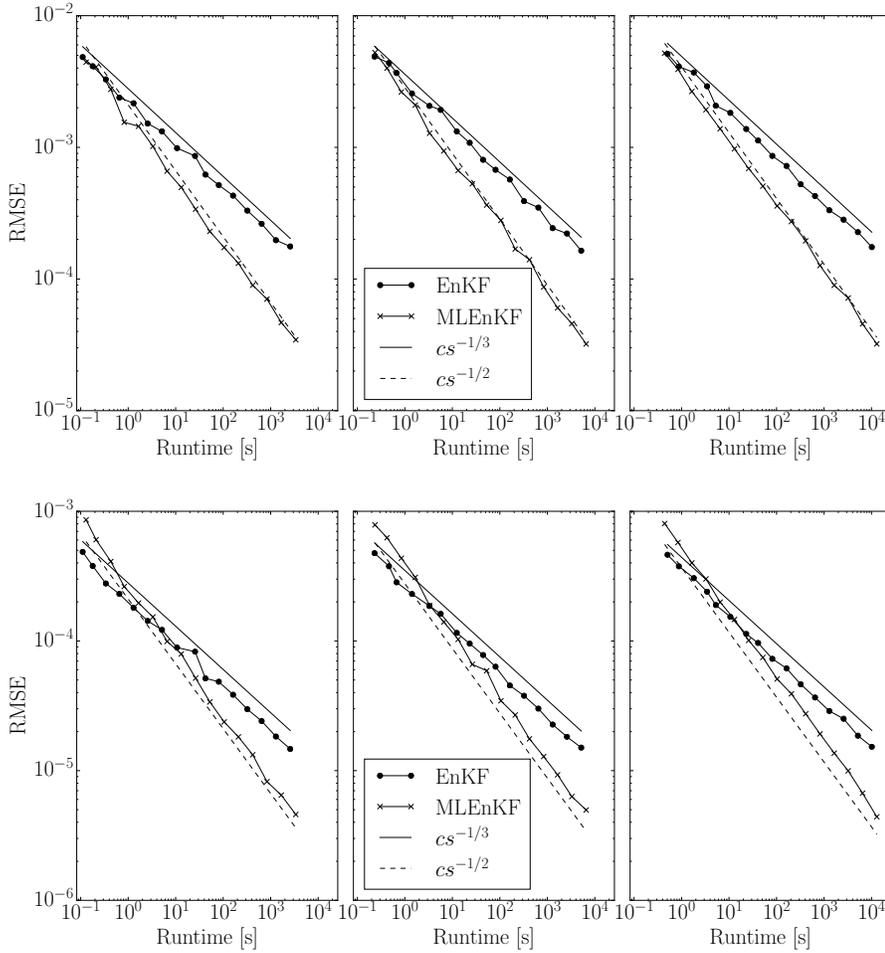


Figure 4: Comparison of the accuracy vs. computational cost when using the EnKF and MLEnKF methods on a linear Gaussian filtering problem of the form given in (10). The observations occur at times  $1, \dots, N$ . The error is measured in terms of the RMSE for the mean (top row) and covariance (bottom row), 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. [49]

#### 4.3.4 Discussion

Some examples of coupling algorithms have been reviewed. The main issues with such techniques are (i) coupling the algorithms correctly, so that the coupling is ‘good enough’, and (ii) mathematical analysis of such couplings to prove that they indeed provide a benefit. These challenges are crucial and must be further studied. It was already mentioned that the works [50, 84] consider coupling the pair of particle filters arising in a PMCMC algorithm, and empirical results are promising. However, establishing that indeed (i)-(ii) would occur in practice is not so easy and at least does not appear to have been done in publicly available research. In the MLMC context, the theoretical and numerical results of [55] indicate a loss of a power of  $1/2$  in the rate of strong convergence following from the coupled resampling of section 4.3.2, hindering the ultimate cost of the algorithm (see also table 3). However, the rates may be improved by the resampling based on optimal transportation from [50, 84]. Indeed the works [42, 43] numerically observe preservation of the strong rate of convergence using a deterministic transformation based on the optimal transportation coupling, in lieu of resampling.

## 5 Future Work and Summary

Here we examined some computational approaches to facilitate the application of the MLMC method in challenging examples, where standard (independent) sampling is not currently possible. Some review of the computational methods was provided, although as we have noted it is a large literature that one cannot hope to include a complete summary of all the methodology. We then detailed various approaches one can use to leverage MLMC within these methods.

There are many areas for possible exploration in future work. One strand consists of considering multi-dimension discretizations, such as in [45]. There are a small number of papers on this topic, such as [22, 57], but there seem to be many possible avenues for future work. 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. One final very interesting avenue for future research is exact coupling, using optimal transport and i.i.d. sampling. For some model structures, the ideas of [87] could be very useful.

### 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.” KJHL was supported by ORNL LDRD Seed funding grant number 32102582.

### 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. Probab.*, **16**, 1462–1505.
- [3] BESKOS, A., CRISAN, D. & JASRA, A. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.*, **24**, 1396–1445.
- [4] BESKOS, A. , CRISAN, D., JASRA, A. & WHITELEY, N. (2014). Error bounds and normalizing constants for sequential Monte Carlo. *Adv. Appl. Probab.*, **46**, 279–306.
- [5] BESKOS, A., JASRA, A., KANTAS, N. & THIERY, A. (2016). On the convergence of adaptive sequential Monte Carlo. *Ann. Appl. Probab.*, **26**, 1111-1146.
- [6] BESKOS, A., JASRA, A., LAW, K., MARZOUK, Y., & ZHOU, Y. (2017). Multilevel Sequential Monte Carlo samplers with dimension independent likelihood informed proposals. arXiv preprint.

- [7] 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.
- [8] BESKOS, A., JASRA, A., LAW, K. J. H., TEMPONE, R., & ZHOU, Y. (2017). Multilevel Sequential Monte Carlo samplers. *Stoch. Proc. Appl.* (to appear).
- [9] BESKOS, A., ROBERTS, G., STUART, A., & VOSS, J. (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics*, **8**(03), 319-350.
- [10] 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.
- [11] BOUCHARD-COTE, A., VOLLMER, S. & DOUCET, A. (2017). The bouncy particle sampler: A non-reversible rejection free MCMC method. *J. Amer. Statist. Assoc.* (to appear).
- [12] CAPPÉ, O., RYDEN, T., & MOULINES, É. (2005). *Inference in Hidden Markov Models*. Springer: New York.
- [13] CHAN, H. P. & LAI, T. L. (2013). A general theory of particle filters in hidden Markov models and some applications. *Ann. Statist.*, **41**, 2877-2904.
- [14] CHERNOV, A., HOEL, H., LAW, K., NOBILE, F., & TEMPONE, R. (2016). Multilevel ensemble Kalman filtering for spatially extended models. arXiv preprint arXiv:1608.08558.
- [15] 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.
- [16] CHOPIN, N. (2002). A sequential particle filter for static models. *Biometrika*, **89**, 539–552.

- [17] CHOPIN, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.*, **32**, 2385–2411.
- [18] CHOPIN, N. & SINGH, S. S. (2015). On particle Gibbs sampling. *Bernoulli*, **21**, 1855–1883.
- [19] CHOPIN, N., JACOB, P. E., & PAPASPILIOPOULOS, O. (2013). SMC<sup>2</sup>: an efficient algorithm for sequential analysis of state space models. *J. R. Statist. Soc. B*, **75**, 397–426
- [20] 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.
- [21] CRISAN, D., ROZOVSKII, B. (2011) The Oxford handbook of nonlinear filtering, *Oxford Univ. Press*. Oxford,
- [22] CRISAN, D, HOUSSEINEAU, J., & JASRA, A. (2017). Unbiased Multi-Index Monte Carlo. arXiv preprint.
- [23] DEL MORAL, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.
- [24] DEL MORAL, P. (2013). *Mean Field Simulation for Monte Carlo Integration* Chapman & Hall: London.
- [25] DEL MORAL, P., DOUCET, A. & JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436.
- [26] DEL MORAL, P., DOUCET, A. & JASRA, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statist. Comp.*, **22**, 1009–1020.
- [27] DEL MORAL, P., JASRA, A. & LAW, K. J. H. (2017). Multilevel Sequential Monte Carlo: Mean Square Error Bounds under Verifiable Conditions. *Stoch. Anal.* **35**, 478–498.
- [28] DEL MORAL, P, JASRA, A., LAW, K. J. H. & ZHOU, Y. (2016). Multilevel SMC samplers for normalizing constants. arXiv preprint.

- [29] DELIGIANNIDIS, G., DOUCET, A. & PITT, M. (2015). The Correlated Pseudo-Marginal Method. arXiv preprint.
- [30] DOUC, R. & MOULINES, E. (2008). Limit theorems for weighted samples with applications to sequential Monte Carlo methods. *Ann. Statist.*, **36**, 2344–2376.
- [31] DOUCET, A., DE FREITAS, N. & GORDON, N. (2001). *Sequential Monte Carlo methods in practice*. Springer: New York.
- [32] 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.
- [33] 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.
- [34] DUANE, S., KENNEDY, A. D., PENDLETON, B. J., & ROWETH, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, **195**, 216–222.
- [35] 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.
- [36] EVENSEN, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dyn.*, **53**, 343–367.
- [37] FEARNHEAD, P., PAPASPILOPOULOS, O. & ROBERTS, G. O. (2008). Particle filters for partially observed diffusions. *J. R. Stat. Soc. Ser. B* **70**, 755–777.
- [38] GLASSERMAN, P., & STAUM, J. (2001). Conditioning on one-step survival for barrier options. *Op. Res.*, **49**, 923–937.
- [39] GILES, M. B. (2008). Multilevel Monte Carlo path simulation. *Op. Res.*, **56**, 607–617.
- [40] GILES, M. B. (2015) Multilevel Monte Carlo methods. *Acta Numerica* **24**, 259–328.

- [41] GILES, M. B., NAGAPETYAN, T., SZPRUCH, L., VOLLMER, S., & ZYGALAKIS, K. (2016). Multilevel Monte Carlo for Scalable Bayesian Computations. arXiv preprint.
- [42] GREGORY, A., COTTER, C., & REICH, S. (2016). Multilevel Ensemble Transform Particle Filtering. *SIAM J. Sci. Comp.* **38**, A1317-A1338.
- [43] GREGORY, A., & COTTER, C. (2016). A Seamless Multilevel Ensemble Transform Particle Filter. arXiv preprint.
- [44] HAARIO, H., SAKSMAN, E. & TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, **7**, 223–242.
- [45] HAJI-ALI, A. L., NOBILE, F. & TEMPONE, R. (2016). Multi-Index Monte Carlo: When sparsity meets sampling. *Numerische Mathematik*, **132**, 767–806.
- [46] HEINRICH, S. (2001). Multilevel Monte Carlo methods. In *Large-Scale Scientific Computing*, (eds. S. Margenov, J. Wasniewski & P. Yalamov), Springer: Berlin.
- [47] HENG, J., DOUCET, A. & POKERN, Y. (2015). Gibbs Flow for Approximate Transport with Applications to Bayesian Computation. arXiv preprint.
- [48] HOANG, V., SCHWAB, C. & STUART, A. (2013). Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Prob.*, **29**, 085010.
- [49] HOEL, H. LAW, K. & TEMPONE, R. (2016). Multilevel ensemble Kalman filter. *SIAM J. Numer. Anal.*, **54**, 1813–1839.
- [50] JACOB, P. E., LINDSTEN, F. & SCHONN, T. (2016). Coupling of particle filters. arXiv preprint.
- [51] JARZYNSKI, C., (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693.
- [52] 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.

- [53] JASRA, A., STEPHENS, D. A. & HOLMES C. C. (2007). On population-based simulation for static inference. *Statist. Comp.*, **17**, 263–279.
- [54] JASRA, A., KAMATANI, K., LAW K. J. H. & ZHOU, Y. (2015). Multilevel particle filters. arXiv preprint, arXiv:1605.04963.
- [55] JASRA, A., KAMATANI, K., LAW, K. J. H. & ZHOU, Y. (2017). Bayesian Static Parameter Estimation for Partially Observed Diffusions via Multilevel Monte Carlo. arXiv preprint.
- [56] JASRA, A., KAMATANI, K., LAW, K. J. H. & ZHOU, Y. (2017). A multi-index Markov chain Monte Carlo method. arXiv preprint.
- [57] JASRA, A., KAMATANI, K., LAW, K. J. H., & ZHOU, Y. (2017). A Multi-Index Markov Chain Monte Carlo Method. Submitted. arXiv preprint.
- [58] 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 .
- [59] JASRA, A., KAMATANI, K., OSEI, P. P., & ZHOU, Y. (2017). Multilevel particle filters: Normalizing Constant Estimation. *Statist. Comp.* (to appear).
- [60] JASRA, A., JO, S., NOTT, D., SHOEMAKER, C. & TEMPONE, R. (2016). Multilevel Monte Carlo in approximate Bayesian computation. arXiv preprint.
- [61] 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.
- [62] 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.
- [63] 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.

- [64] KLOEDEN, P. E. & PLATEN, E. (1992) *Numerical Solution of Stochastic Differential Equations*. Springer: Berlin.
- [65] LAW, K., STUART, A. AND ZYGALAKIS, K. (2015). *Data Assimilation*. Springer-Verlag, New York.
- [66] LAW, K. J. H., TEMBINE, H., & TEMPONE, R. (2016). Deterministic mean-field ensemble Kalman filtering. *SIAM Journal on Scientific Computing*, **38**(3), A1251–A1279.
- [67] MARIN, J.-M., PUDLO, P., ROBERT, C.P. & RYDER, R. (2012). Approximate Bayesian computational methods. *Statist. Comp.*, **22**, 1167–1180.
- [68] MEYN, S. & TWEEDIE, R.L. (2009). *Markov Chains and Stochastic Stability*. Second edition, CUP: Cambridge.
- [69] NÆSSETH, C. A., LINDSTEN, F., & SCHÖN, T. B. (2015). Nested Sequential Monte Carlo Methods. *Proc. 32nd ICML*.
- [70] NEAL, R. M. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics, No. 118. Springer-Verlag.
- [71] NEAL, R. M. (2001). Annealed importance sampling. *Statist. Comp.*, **11**, 125–139.
- [72] 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.
- [73] OTTOBRE, M. (2016). Markov Chain Monte Carlo and Irreversibility. *Reports on Mathematical Physics*, **77**(3), 267–292.
- [74] REBESCHINI, P. & VAN HANDEL, R. (2015). Can local particle filters beat the curse of dimensionality? *Ann. Appl. Probab.*, **25**, 2809–2866.
- [75] RHEE, C. H., & GLYNN, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Op. Res.*, **63**, 1026–1043.

- [76] ROBERT, C. (2001). *The Bayesian Choice: from Decision-Theoretic Motivations to Computational Implementation*. Springer: New York.
- [77] ROBERTS, G. O. & TWEEDIE, R. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.
- [78] ROBERTS, G. O., & ROSENTHAL, J. (2004). General state-space Markov chains and MCMC algorithms. *Probab. Surveys*, **1**, 20–71.
- [79] 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.
- [80] ROUSSET, M., & DOUCET, A. (2006). Discussion of Beskos et al. *J. R. Statist. Soc. B*, **68** 374–375.
- [81] SCHÄFER, C. & CHOPIN, N. (2013). Adaptive Monte Carlo on binary sampling spaces. *Statist. Comp.*, **23**, 163–184.
- [82] SCHWEIZER, N. (2012). Non-asymptotic error bounds for sequential MCMC and stability of Feynman-Kac propagators. arXiv preprint, arXiv:1204.2382.
- [83] 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.
- [84] SEN, D., THIERY, A., JASRA, A. (2016). On coupling particle filters. arXiv preprint.
- [85] SINGH, S. S., LINDSTEN, F. & MOULINES, E. (2015). Blocking Strategies and Stability of Particle Gibbs Samplers. arXiv preprint.
- [86] SNYDER, C., BENGTTSSON, T., BICKEL, P., & ANDERSON, J. (2008). Obstacles to high-dimensional particle filtering. *Month. Weather Rev.*, **136**, 4629–4640.
- [87] SPANTINI, A., BIGONI, D. & MARZOUK Y. (2017). Inference via low-dimensional couplings. arXiv preprint.

- [88] 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.
- [89] VILLANI, C. (2008). *Optimal transport: old and new (Vol. 338)*. Springer Science & Business Media.
- [90] WHITELEY, N. P. (2012). Sequential Monte Carlo samplers: Error bounds and insensitivity to initial conditions. *Stoch. Anal.*, **30**, 774–798.