Multi-Index Sequential Monte Carlo Methods for partially observed Stochastic Partial Differential Equations

BY YAXIAN XU[1], AJAY JASRA[2] & KODY J. H. LAW[3]

[1]Department of Statistics & Applied Probability, National University of Singapore, Singapore, 117546, SG.

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

[2]Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, 23955-6900, KSA. E-Mail: *ajay.jasra@kaust.edu.sa*

[3]School of Mathematics, University of Manchester, Manchester, M13 9PY, UK.

E-Mail: *kodylaw@gmail.com*

Abstract

In this paper we consider sequential joint state and static parameter estimation given discrete time observations associated to a partially observed stochastic partial differential equation. It is assumed that one can only estimate the hidden state using a discretization of the model. In this context, it is known that the multi-index Monte Carlo (MIMC) method of [12] can be used to improve over direct Monte Carlo from the most precise discretizaton. However, in the context of interest, it cannot be directly applied, but rather must be used within another method such as sequential Monte Carlo (SMC). We show how one can use the MIMC method by renormalizing the standard identity and approximating the resulting identity using the SMC^2 method of [5], which is an exact method that can be used in this context. We prove that our approach can reduce the cost to obtain a given mean square error, relative to just using SMC^2 on the most precise discretization. We demonstrate this with some numerical examples.

Key Words: Stochastic Partial Differential Equations; Multi-Index Monte Carlo, Sequential Monte Carlo

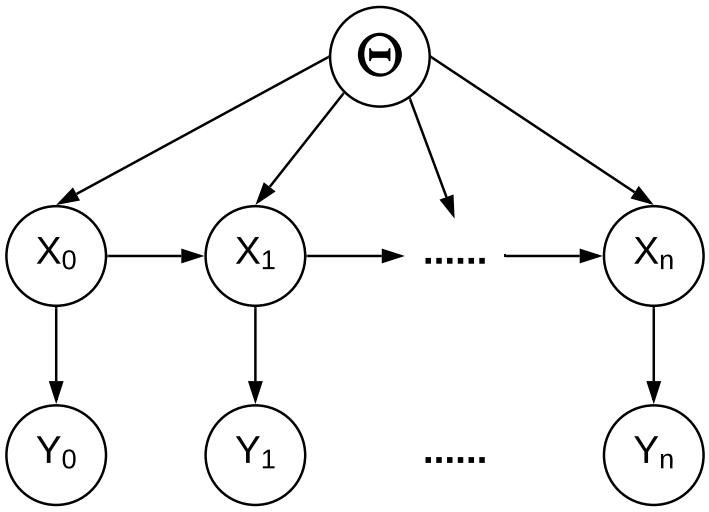

Figure 1: A graphical model of the HMM studied in this paper. fig:hmm

1 Introduction

We consider joint state and static parameter estimation for discrete time observations, associated to a partially observed stochastic partial differential equation (SPDE). Such models can be considered a form of hidden Markov model (HMM), and these have a significant number of practical applications; see e.g. Cappe_2005 [3] for instance. See Figure fig:hmm 1 for a schematic. The objective is to estimate the states $(X_0, X_1, \ldots, X_n) \in \mathsf{X}^{n+1}$ and parameters $\theta \in \Theta$ given the data $(y_0, \ldots, y_n)$, i.e. we want to find

$$\mathbb{P}(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n, \theta \in A_T | y_0, \ldots, y_n)$$

recursively in time n.

In this article we focus upon the scenario where one will have to discretize the time and space element of the SPDE. In this scenario, one is faced with the problem of joint state and static parameter estimation for a HMM with a high-dimensional state; a problem which is notoriously challenging. The main issue is that for any fixed static parameter, one can seldom calculate the joint density (the smoother), given the data, of the hidden states over the observation times. Joint inference on the parameter is even more challenging, due to the dependence of all the hidden states on the parameter (see Figure fig:hmm 1). The state of the art method for consistently solving this problem for a fixed time n is the particle Markov

chain Monte Carlo (PMCMC) method [andrieu1], which involves using sequential Monte Carlo (SMC) within Markov chain Monte Carlo (MCMC). SMC methods can approximate the sequence of joint distributions on $X_{0:n}$ for θ fixed (smoothers). They consist of sampling N samples (also called particles) in parallel, sequentially in time. SMC methods involve the recursion of a mutation step, an importance sampling step, and a resampling step. They provide consistent (as $N \to \infty$) approximations of expectations w.r.t. the smoother. In many contexts, SMC is referred to as a particle filter, hence the name PMCMC. We remark, however, one also seeks to perform statistical inference sequentially in time, which adds yet another complication. There exists a methodology which can extend PMCMC to a dynamic context, called SMC^2 [chopin2].This is an approach which is essentially a class of SMC algorithm which will sequentially sample from a sequence of auxiliary target distributions, which admit the distribution of interest as a marginal. The main reason why one wishes to consider such a level of complexity, is that conventional SMC methodology which is designed for joint state and parameter inference associated to HMMs suffers from the so-called path degeneracy issue (see e.g. [kantas1]), which renders it useless for long time intervals.

In the problem of interest, we are dealing with an expectation w.r.t. a probability measure which is defined on a high-dimensional continuous state-space as a result of the space and time discretization. It is assumed that the computational cost associated to performing any simulation-based numerical method will increase as the precision of the discretization is enhanced. In such a context, the multilevel Monte Carlo (MLMC) method is very effective for reducing the cost to achieve a given level of accuracy [giles,giles1,hein]. Following the success of the MLMC method, the work [mimc] revisited the MLMC identity through the lense of sparse grids for multi-dimensional discretizations. This more general method, which reduces to MLMC for one dimensional (discretizations) problems, is called multi-index Monte Carlo (MIMC). This method can be much more efficient than MLMC for higher dimensional problems. In this approach one rewrites the expectation of interest as a sum of difference of differences (DOD) w.r.t. independent refinements of the discretization levels of different dimensions, for example different spatial dimensions and time. If the number of dimensions

which are discretized is d, then this DOD involves expectations w.r.t. at most 2^d different discretization levels. Then, under appropriate assumptions and given an efficient *coupling* of these 2^d distributions, the cost to achieve a prespecified mean square error (MSE) is reduced with respect to considering the Monte Carlo method using a single discretization level [mimc]. Indeed, under appropriate assumptions, relying strongly on the mixed regularity of the solution of the SPDE, and with an appropriate index set, the MIMC approach can achieve a substantial improvement in cost with respect to MLMC [mimc]. The coupling is essential to ensuring that the *variance* of the estimates of the DODs decay appropriately w.r.t. the discretization indices. This allows one to use fewer samples in approximating higher index residuals, hence balancing the cost in an optimal way.

The sampling of a "good coupling" of the 2^d distributions is the main key to cost reduction in MIMC, as in MLMC. The problem of approximating expectations w.r.t. the joint distribution of interest for a single discretization level/index is already challenging, and coupling 2^d such distributions is naturally much more complex. A method for using MCMC within the MIMC framework was developed in [jklz], based upon an idea in [jklz1] developed originally for PMCMC (see also [franks]). The method involves constructing an approximate coupling of the 2^d targets and using this to approximate a re-normalized multi-index (MI) identity. See [ml_rev] for a pedagogical introduction to this general approximate coupling strategy. In the present context, the aim is to perform inference sequentially as data arrives, where the unobserved process is an SPDE. This is done using the SMC^2 method described above. This involves first extending the MLMC PMCMC method of [jklz1] to the MIMC context, in order to accommodate an SPDE model. Second, this new MIMC PMCMC is deployed within an SMC^2 algorithm, which yields the MI SMC^2 algorithm. Under appropriate assumptions, we prove that our approach can reduce the cost required to obtain a given MSE, relative to just using SMC^2 on the most precise discretization, or even using the MLMC version. This is demonstrated with a numerical example.

This article is structured as follows. In Section [sec:cartoon] 2, we provide a high-level introduction into the underlying idea of the article. In Section [sec:prob_form] 3 we describe the problem and how it

may be solved, if numerical approximation were not required. In Section 4 we show how our approach can be numerically approximated. In Section 5 the theoretical result is given, with the proofs in the appendix. Numerical results are presented in Section 6 .

2 High-Level Discussion of the Approach

sec:cartoon

The method presented in this article is quite complicated. To assist non-experts, we provide a high-level description of the base idea, when one has to discretize in one dimension and minimal notations.

Consider a probability density on state-space X

$$p(x) := \frac{J(x)F(x)}{Z}$$

where J, F are two positive, real-valued functions, $\int_{\mathsf{X}} F(x)dx = 1$, $Z = \int_{\mathsf{X}} J(x)F(x)dx$ (assumed to be finite). It is of interest to compute expectations of manyreal-valued functions $\varphi : \mathsf{X} \to \mathbb{R}$ that are $p-$integrable:

$$\mathbb{E}_p[\varphi(X)] = \int_{\mathsf{X}} \varphi(x)p(x)dx.$$

Suppose that one can only evaluate/has access to a sequence $J_l(x)F_l(x)$, $l \in \{0, 1, \dots\}$ which are positive, real-valued functions such that

$$\lim_{l\to\infty} \mathbb{E}_{p_l}[\varphi(X)] = \mathbb{E}_p[\varphi(X)]$$

where $\mathbb{E}_{p_l}[\varphi(X)] = \int_{\mathsf{X}} \varphi(x)p_l(x)dx$, $p_l(x) = [J_l(x)F_l(x)]/Z_l$, $Z_l = \int_{\mathsf{X}} J_l(x)F_l(x)dx$.

Now, the MLMC identity

$$\mathbb{E}_{p_L}[\varphi(X)] = \sum_{l=1}^{L} \{\mathbb{E}_{p_l}[\varphi(X)] - \mathbb{E}_{p_{l-1}}[\varphi(X)]\} + \mathbb{E}_{p_0}[\varphi(X)]$$

can be very useful to reduce the computational effort in the Monte Carlo approximation of $\mathbb{E}_{p_L}[\varphi(X)]$, to achieve a given error (versus considering only $\mathbb{E}_{p_L}[\varphi(X)]$). The main key to this method, is being able to construct a coupling $\breve{p}_{l,l-1}$ of (p_l, p_{l-1}) for $l \in \{1, 2 \dots\}$. That

is, a probability density function on $\mathsf{X} \times \mathsf{X}$ such that for every $(x, x') \in \mathsf{X} \times \mathsf{X}$

$$\begin{aligned} p_l(x) &= \int_{\mathsf{X}} \check{p}_{l,l-1}(x, x')dx' \\ p_{l-1}(x') &= \int_{\mathsf{X}} \check{p}_{l,l-1}(x, x')dx. \end{aligned}$$

Then, one has

$$\mathbb{E}_{p_l}[\varphi(X)] - \mathbb{E}_{p_{l-1}}[\varphi(X)] = \int_{\mathsf{X}\times\mathsf{X}} \varphi(x)\check{p}_{l,l-1}(x, x')d(x, x') - \int_{\mathsf{X}\times\mathsf{X}} \varphi(x')\check{p}_{l,l-1}(x, x')d(x, x').$$

The idea then, is that if the coupling is sufficiently good, for instance that

$$\int_{\mathsf{X}\times\mathsf{X}} (\varphi(x) - \varphi(x'))^2 \check{p}_{l,l-1}(x, x')d(x, x') \leq h(l)$$

where $\lim_{l\to\infty} h(l) = 0$, h is a positive and real-valued function on $\{0, 1, \dots\}$, then the afore-mentioned benefits are possible; see e.g. [giles,giles1,hein] [10, 11, 13]. The Monte Carlo method would rely on exact sampling from the distribution associated to the coupling $\check{p}_{l,l-1}$.

Let $l \geq 1$ be fixed. In many practical problems of interest, especially the one in this article, deriving a suitable coupling $\check{p}_{l,l-1}$ which is amenable to known simulation methodology can be very challenging. The basic idea used in this paper and as adopted in [jklgz1] [15] is as follows. Suppose one can find a coupling $\check{F}_{l,l-1}$ of (F_l, F_{l-1}). That is, a probability density function on $\mathsf{X} \times \mathsf{X}$ such that for every $(x, x') \in \mathsf{X} \times \mathsf{X}$

$$\begin{aligned} F_l(x) &= \int_{\mathsf{X}} \check{F}_{l,l-1}(x, x')dx' \\ F_{l-1}(x') &= \int_{\mathsf{X}} \check{F}_{l,l-1}(x, x')dx. \end{aligned}$$

It is supposed that $\check{F}_{l,l-1}$ is sufficiently good that

$$\int_{\mathsf{X}\times\mathsf{X}} (\varphi(x) - \varphi(x'))^2 \check{F}_{l,l-1}(x, x')d(x, x') \leq h(l).$$

Now set

$$\tilde{p}_{l,l-1}(x, x') = \frac{\max\{J_l(x), J_{l-1}(x')\}\check{F}_{l,l-1}(x, x')}{\tilde{Z}_{l,l-1}}$$

with $\tilde{Z}_{l,l-1} = \int_{\mathsf{X}\times\mathsf{X}} \max\{J_l(x), J_{l-1}(x')\}\check{F}_{l,l-1}(x, x')d(x, x')$ (assumed to be finite). Now, we

note that

$$\begin{aligned}
\mathbb{E}_{p_l}[\varphi(X)] &= \int_{\mathsf{X}} \varphi(x) p_l(x) dx \\
&= \frac{1}{Z_l} \int_{\mathsf{X}\times\mathsf{X}} \varphi(x) J_l(x) \check{F}_{l,l-1}(x, x') d(x, x') \\
&= \frac{\tilde{Z}_{l,l-1}}{Z_l} \int_{\mathsf{X}\times\mathsf{X}} \varphi(x) \frac{J_l(x)}{\max\{J_l(x), J_{l-1}(x')\}} \tilde{p}_{l,l-1}(x, x') d(x, x') \\
&= \mathbb{E}_{\tilde{p}_{l,l-1}}\left[\varphi(X) \frac{J_l(X)}{\max\{J_l(X), J_{l-1}(X')\}}\right] \Big/ \mathbb{E}_{\tilde{p}_{l,l-1}}\left[\frac{J_l(X)}{\max\{J_l(x), J_{l-1}(X')\}}\right]
\end{aligned}$$

where expectations w.r.t. $\tilde{p}_{l,l-1}$ are denoted $\mathbb{E}_{\tilde{p}_{l,l-1}}$. Thus

$$\begin{aligned}
\mathbb{E}_{p_l}[\varphi(X)] - \mathbb{E}_{p_{l-1}}[\varphi(X)] &= \mathbb{E}_{\tilde{p}_{l,l-1}}\left[\varphi(X) \frac{J_l(X)}{\max\{J_l(X), J_{l-1}(X')\}}\right] \Big/ \mathbb{E}_{\tilde{p}_{l,l-1}}\left[\frac{J_l(X)}{\max\{J_l(X), J_{l-1}(X')\}}\right] - \\
&\quad \mathbb{E}_{\tilde{p}_{l,l-1}}\left[\varphi(X') \frac{J_{l-1}(X')}{\max\{J_l(X), J_{l-1}(X')\}}\right] \Big/ \mathbb{E}_{\tilde{p}_{l,l-1}}\left[\frac{J_{l-1}(X')}{\max\{J_l(X), J_{l-1}(X')\}}\right].
\end{aligned}$$

The main interest of the identity, is the fact that, one can sometimes construct $\check{F}_{l,l-1}$ and very efficient sampling methods for $\tilde{p}_{l,l-1}$, whereas this may not be the case for $\check{p}_{l,l-1}$. It is then possible (e.g. [15]) that the benefits of the MLMC method are gained, even though one does not know how to sample from a good coupling $\check{p}_{l,l-1}$.

AJAY: I removed y here, as it is redundant. Otherwise, this is great. The only issue is that we lost the sketch of PMCMC and SMC^2, which is the most important part, by far. Is this a question of ability (as in my case), or are you morally opposed to concisely explaining at a high level work which is not your own? Without this, I would not hold breath for publication in SISC. KODY: First of all I have explained to you several times that it is easier to present SMC^2 as an SMC method and forget about explaining all the nested issues. These are all given in clarity in the original papers. It is now clear how all the algorithms work (perhaps not why, but that is not our job). Secondly, SMC^2 has no place here, because we explain it in details much later on.

3 Rigorous Problem Formulation

sec:prob_form

3.1 Model

Let $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{X}, \mathcal{X})$ be measurable spaces. We consider a pair of stochastic processes indexed by a time parameter. We are given a sequence of observations $y_0, y_1, \ldots$ which

are realizations of a discrete-time process $\{Y_n\}_{n\in\mathbb{N}_0}$, $Y_n \in \mathsf{Y}$, where the time between observations is one unit. These observations are associated with a continuous-time (Markov) stochastic process $\{X_t\}_{t\geq 0}$, with $X_t \in \mathsf{X}$. The process would typically arise from the finite-time evolution of an SPDE, although we do not make this constraint at this time.

We now present the stochastic model which describes the probabilistic structure of the processes $\{Y_n\}_{n\in\mathbb{N}_0}$ and $\{X_t\}_{t\geq 0}$. In our model, $\theta \in \Theta \subseteq \mathbb{R}^k$ is a static parameter associated to the model. We will define the afore-mentioned structure, conditonal upon θ and then define a prior probability distribution on this static parameter. Let $X_{0:n} = (X_0, \ldots, X_n)$ correspond to a discrete-time skeleton of $\{X_t\}_{t\geq 0}$ on the grid $0 : n$. We are interested in the posterior probability distribution of $(X_{0:n}, \theta)$ conditional on observed data $y_0, \ldots, y_n$, sequentially over discrete unit times (n). It is supposed that for any $n \geq 0$, $A \in \mathcal{Y}$

$$\mathbb{P}(Y_n \in A | y_{0:n-1}, \{x_t\}_{t\in[0,\ldots,n]}, \theta) = \int_A g_\theta(x_n, y) dy$$

where dy is a $\sigma-$finite measure on $(\mathsf{Y}, \mathcal{Y})$ and for each $(\theta, x) \in \Theta \times \mathsf{X}$, $g_\theta(x, \cdot) : \mathsf{Y} \to \mathbb{R}^+$ is a probability density on Y. For each $\theta \in \Theta$, $f_\theta : \mathsf{X}^2 \to \mathbb{R}^+$, (resp. $\mu_\theta : \mathsf{X} \to \mathbb{R}^+$) are the transition density over unit time (resp. initial density) of $\{X_t\}_{t\geq 0}$ (resp. X_0) w.r.t. a dominating $\sigma-$finite measure on $(\mathsf{X}, \mathcal{X})$. Note that for every $(\theta, x) \in \Theta\times\mathsf{X}$, $f_\theta(x, \cdot) : \mathsf{X} \to \mathbb{R}_+$ is a probability density, and for any $n \geq 1$

$$\mathbb{P}(X_n \in A | \{x_r\}_{0\leq r\leq n-1}, \theta) = \int_A f_\theta(x_{n-1}, x) dx\,.$$

Let ν be a probability density w.r.t. Lebesque measure (written $d\theta$) on $(\Theta, \mathcal{B}(\Theta))$ with $\mathcal{B}(\Theta)$ the Borel sets.

For $n \geq 0$, the posterior probability density on $\mathsf{X}^{n+1} \times \Theta$ that is induced by this construction is given by

$$\pi_n(x_{0:n}, \theta) \propto \nu(\theta)\mu_\theta(x_0)g_\theta(x_0, y_0) \prod_{p=1}^{n} f_\theta(x_{p-1}, x_p)g_\theta(x_p, y_p)\,. \tag{1}$$ eq:post_def

In other words for $A \in \vee^{n+1}\mathcal{X} \vee \mathcal{B}(\Theta)$

$$\mathbb{P}((X_{0:n}, \theta) \in A | y_{1:n}) = \int_A \pi_n(x_{0:n}, \theta) d(x_{0:n}, \theta)\,.$$

Henceforth, we will suppress the dependence on $y_{0:n}$ throughout the article. Let $\varphi : \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$ be integrable w.r.t. π_n. Our objective is to compute, recursively in n

$$\mathbb{E}_{\pi_n}[\varphi(X_{0:n}, \theta)] := \int_{\mathsf{X}^{n+1}\times\Theta} \varphi(x_{0:n}, \theta)\pi_n(x_{0:n}, \theta)d(x_{0:n}, \theta)$$

where we use the notation $\mathbb{E}_\pi$ to denote expectations w.r.t. a probability density/measure π. The role of the function φ is as a summary or quantity of interest, relating to the random variables $X_{0:n}, \theta$. That is, our objective is to compute expectations, such as moments, w.r.t. the posterior, with density defined in (1). We note that one often must use Monte Carlo methods to approximate the sequence of expectations.

Before concluding this section we mention a canonical statistical model, the conditionally Gaussian model.

ex:gauss **Example 3.1.** *Let $N(m, C)$ denote a (possibly infinite-dimensional) Gaussian random variable with mean m and covariance operator C, and let $\phi(\cdot; m, C)$ denote its density (with respect to some dominating measure, which may be taken as Lebesgue in finite dimensions). Assume $X_0 \sim N(m_0, \Sigma_0)$. For each $\theta \in \Theta$, let $\Psi_\theta : \mathsf{X} \to \mathsf{X}$ and $h_\theta : \mathsf{X} \to \mathsf{Y}$ be continuous and let $\Sigma_\theta, \Gamma_\theta$ be symmetric positive definite operators. An example of a model is, for $n \geq 0$*

$$\begin{aligned} X_{n+1}|X_n &\sim N(\Psi_\theta(X_n), \Sigma_\theta)\,, \\ Y_n|X_n &\sim N(h_\theta(X_n), \Gamma_\theta)\,. \end{aligned} \tag{2}$$

This model is ubiquitous in the data assimilation literature [20]. Once ν is specified, the model (1) is given by $\mu_\theta(x_0) = \phi(x_0; m_0, \Sigma_0)$, $f_\theta(x, x') = \phi(x'; \Psi_\theta(x), \Sigma_\theta)$, and $g_\theta(x, y) = \phi(y; h_\theta(x), \Gamma_\theta)$.

This model fits into the context of this paper when X is infinite dimensional, e.g. a Hilbert space, and Ψ_θ is the solution of a PDE parametrized by θ.

sec:disc_model

3.2 Discretized Model

The exposition here closely follows that developed in [14]. Here we explicitly assume one must work with a discretized version of the model, that is, there does not (currently) exist

an unbiased and non-negative approximation of $\pi_n(x_{0:n}, \theta)$. We remark that if the latter approximations are available, then the strategy to be outlined is not required.

Set $\alpha \in \mathbb{N}_0^d$, which will refer to a collection of indices which will denote the level of discretization of our model in each of d dimensions. That is, as the components of α increase, so does the accuracy of the approximation. Precise examples are given in Section 6. More explicitly, for any fixed $\alpha \in \mathbb{N}_0^d$, let $(\mathsf{X}_\alpha, \mathcal{X}_\alpha)$ and $(\mathsf{Y}_\alpha, \mathcal{Y}_\alpha)$ be measurable spaces such that for every $n \geq 0$ one can obtain a biased approximation $X_\alpha \in \mathsf{X}_\alpha \subseteq \mathsf{X}$ of X_n, and $Y_\alpha \in \mathsf{Y}_\alpha \subseteq \mathsf{Y}$ of Y_n. That is, one can define the probability density for $n \geq 0$, on $\mathsf{X}_\alpha^{n+1} \times \Theta$:

$$\pi_{n,\alpha}(x_{0:n}, \theta) \propto \nu(\theta)\mu_{\theta,\alpha}(x_0)g_{\theta,\alpha}(x_0, y_0)\prod_{p=1}^{n} f_{\theta,\alpha}(x_{p-1}, x_p)g_{\theta,\alpha}(x_p, y_p) \tag{3}$$

where $y_n \in \mathsf{Y}_\alpha$ for each $n \geq 0$. Here for every $(\alpha, \theta) \in \mathbb{N}_0^d \times \Theta$

- For all $x \in \mathsf{X}_\alpha$, $g_{\theta,\alpha}(x, \cdot)$ is a probability density on Y_α ;
- For all $x \in \mathsf{X}_\alpha$, $f_{\theta,\alpha}(x, \cdot)$ is a probability density on X_α;
- $\mu_{\theta,\alpha}$ is a probability density on X_α.

Consider $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$, where for any $(x_{0:n}, \theta) \in \mathsf{X}^{n+1} \times \Theta$

$$\lim_{\min_{1\leq i\leq d} \alpha_i \to +\infty} \varphi_\alpha(x_{0:n}, \theta) = \varphi(x_{0:n}, \theta).$$

It is assumed that we have

$$\begin{aligned} \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)] &\neq \mathbb{E}_{\pi_n}[\varphi(X_{0:n}, \theta)]. \\ \lim_{\min_{1\leq i\leq d} \alpha_i \to +\infty} |\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)] - \mathbb{E}_{\pi_n}[\varphi(X_{0:n}, \theta)]| &= 0. \end{aligned} \tag{4}$$

As remarked in the introduction, the computational cost associated with X_α, Y_α (sampling, or evaluating the densities $g_{\theta,\alpha}, f_{\theta,\alpha}, \mu_{\theta,\alpha}$) increases as any index of α increases. Our objective is now to compute $\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)]$ recursively for each $n \geq 0$.

3.3 Multi-Index Methods

The approach to be described, provides an approach to approximating $\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)]$ for any fixed $\alpha \in \mathbb{N}_0^d$. Define the difference operator Δ_i, $i \in \{1,\dots,d\}$ as

$$\Delta_i \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)] := \begin{cases} \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)] - \mathbb{E}_{\pi_{n,\alpha-e_i}}[\varphi_{\alpha-e_i}(X_{0:n},\theta)] & \text{if } \alpha_i > 0 \\ \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)] & \text{otherwise} \end{cases}$$

where e_i are the canonical vectors on $\mathbb{R}^d$. Set

$$\Delta \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)] := (\Delta_1 \circ \Delta_2 \circ \cdots \circ \Delta_d)\Big(\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)]\Big) \tag{5}$$ eq:multiincrement

that is, Δ is a composition of $\Delta_1,\dots,\Delta_d$. Note that the order of applying the operators Δ_i in Δ does not matter.

We now consider the identity

$$\mathbb{E}_{\pi_n}[\varphi(X_{0:n},\theta)] = \sum_{\alpha \in \mathbb{N}_0^d} \Delta \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)].$$

The work [12] mimc proposes to leverage this identity by constructing a biased estimator of $\mathbb{E}_{\pi_n}[\varphi(X_{0:n},\theta)]$ as

$$\mathbb{E}_{\pi_n,\mathcal{I}}[\varphi(X_{0:n},\theta)] := \sum_{\alpha' \in \mathcal{I}} \Delta \mathbb{E}_{\pi_{n,\alpha'}}[\varphi_{\alpha'}(X_{0:n},\theta)]\,. \tag{6}$$ eq:mimc

This estimator can (in principle) be approximated by a Monte Carlo method. This can be achieved by (if possible) sampling from a coupling of the (at most) 2^d different probability measures for a given $\alpha \in \mathcal{I}$ (see [12] mimc for details). It is counterintuitive at first to construct a single estimator from a sum of so many other estimators, but in fact if the coupling is strong, and under appropriate assumptions, then this can be much more efficient than a single term estimator. It is remarked that sampling from such a coupling is very challenging, especially in the context of the model considered in Section 3.2. sec:disc_model The residual error is given by

$$\mathbb{E}_{\pi_n}[\varphi(X_{0:n},\theta)] - \mathbb{E}_{\pi_n,\mathcal{I}}[\varphi(X_{0:n},\theta)] = \sum_{\alpha' \notin \mathcal{I}} \Delta \mathbb{E}_{\pi_{n,\alpha'}}[\varphi_{\alpha'}(X_{0:n},\theta)].$$

It is shown in [12, 15] mimc, jklz1 that under appropriate assumptions on the convergence of estimates of the individual terms in (5) eq:multiincrement one can gain significant 'speed-up' relative to single term or even single index MLMC methods. The key point we emphasize here is that *all results of* [12] mimc *, pertaining to all different index sets* $\mathcal{I}$, rely solely on the convergence properties of

estimates of the individual terms (5). Since there are significant other difficulties to deal with in the present work, the finer properties of estimators with various different index sets $\mathcal{I}$ will not be considered here, although we note this is a crucial consideration in practice. Our objective here will rather be to establish a general proof of principle method which provides convergence of estimates of the individual terms in (5) under suitable assumptions. The results will be illustrated in Section 6.

3.4 Renormalized Multi-Index Identity

The following idea builds upon the approaches in [14] and [15]. Consider (6) and in particular consider a single given summand (5) for $\alpha \in \mathcal{I}$, with n fixed. This summand is itself a linear combination of expectations with respect to $1 < k_\alpha \leq 2^d$ probability measures. These k_α probability measures induce k'_α differences in (6); if $k_\alpha = 2^d$, then $k'_\alpha = 2^{d-1}$. We remark that in the case that $k_\alpha = 1$, one does not need to consider how to construct a coupling for an MIMC method as the summand (5) is only an expectation w.r.t. a single probability measure.

For simplicity of notation we will denote the k_α multi-indices by $\alpha(1), \ldots, \alpha(k_\alpha)$, where for $i \in \{1, \ldots, k_\alpha\}$, $\alpha(i) \in \mathcal{I}$. Let $\alpha(i)_j$ denote the $j^{\text{th}}-$element of $\alpha(i)$. The convention of the (non-unique) labelling is such that, $\sum_{j=1}^{d} [\alpha(2i) - \alpha(2i-1)]_j = 1$ for each $i \in \{1, \ldots, k'_\alpha\}$, $\alpha(k_\alpha) = \alpha$ and $\alpha(1) = (\max\{\alpha_1 - 1, 0\}, \ldots, \max\{\alpha_d - 1, 0\})$. This labelling will provide a convenient way to write $\Delta\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)]$ below.

Example 3.2. *Suppose $d = 3$ and $\alpha = (2, 2, 2)$, so $k_\alpha = 8$, $k'_\alpha = 4$. Then a labelling which satisfies the above constraints is*

$$\alpha(1) = (1,1,1), \alpha(2) = (2,1,1), \alpha(3) = (1,1,2), \alpha(4) = (2,1,2),$$

$$\alpha(5) = (1,2,1), \alpha(6) = (2,2,1), \alpha(7) = (1,2,2), \alpha(8) = (2,2,2).$$

Recall the form of the target (3) and the multi-increment summand to be estimated (5). We suppose that there exists a coupling of the discretized dynamics. That is, there exists a Markov density $\check{f}_{\theta,\alpha(1:k_\alpha)}(x(1:k_\alpha), x(1:k_\alpha)')$ such that for any $x(1:k_\alpha) =$

$(x(1),\ldots,x(k_\alpha)) \in \bigotimes_{i=1}^{k_\alpha} \mathsf{X}_{\alpha(i)}$ and any $i \in \{1,\ldots,k_\alpha\}$, $A_i \in \mathcal{X}_{\alpha(i)}$, we have:

$$\int_{\bigotimes_{j=1}^{i-1}\mathsf{X}_{\alpha(j)}\times A_i\times\bigotimes_{j=i+1}^{k_\alpha}\mathsf{X}_{\alpha(j)}} \check{f}_{\theta,\alpha(1:k_\alpha)}(x(1:k_\alpha),x(1:k_\alpha)')dx(1:k_\alpha)' = \int_{A_i} f_{\theta,\alpha(i)}(x(i),x(i)')dx(i)'. \tag{7}$$

eq:coupleddyn

In other words, for a given α, coupled Markov dynamics are performed on the hierarchy of k_α meshes in such a way that the marginal of the coupled dynamics with respect to any of the k_α meshes $\alpha(i)$ corresponds to the exact Markov dynamics on mesh $\alpha(i)$. Importantly, we do not assume that we can evaluate this Markov density. We will only require being able to simulate from it.

Similarly, we suppose that there exists a probability density $\check{\mu}_\theta$ on $\bigotimes_{i=1}^{k_\alpha}\mathsf{X}_{\alpha(i)}$ such that for any $i \in \{1,\ldots,k_\alpha\}$, $A_i \in \mathcal{X}_{\alpha(i)}$, we have:

$$\int_{\bigotimes_{j=1}^{i-1}\mathsf{X}_{\alpha(j)}\times A_i\times\bigotimes_{j=i+1}^{k_\alpha}\mathsf{X}_{\alpha(j)}} \check{\mu}_{\theta,\alpha(1:k_\alpha)}(x(1:k_\alpha))dx(1:k_\alpha) = \int_{A_i} \mu_{\theta,\alpha(i)}(x(i))dx(i). \tag{8}$$

eq:coupledprior

We remark that one can find scenarios for which this is true. We give an example below and another later in Section 6.

sec:numerics

Example 3.3. *As a concrete example, consider the setting of (2) in Example 3.1, where Ψ_θ is the forward solution of a PDE over a unit time interval. One can simulate $X_{0,\alpha(k_\alpha)}$ from $\mu_{\theta,\alpha(k_\alpha)}$, and then coarsen this realization appropriately such that $X_{0,\alpha(i)}$ is a realization of $\mu_{\theta,\alpha(i)}$, for each of the other targets $i < k_\alpha$. For the forward kernel one can similarly simulate a single realization $\chi_{\alpha(k_\alpha)} \sim N(0,\Sigma_{\theta,\alpha(k_\alpha)})$ and then coarsen this to $\chi_{\alpha(i)}$ for driving each of the other dynamics with $i < k_\alpha$. KODY: This example is not precise enough. Some equations would be more helpful.*

Let $\check{g} : \mathbb{N}_0^d \times \bigotimes_{i=1}^{k_\alpha}\mathsf{X}_{\alpha(i)} \times \Theta \times \mathsf{Y} \to (0,\infty)$ be arbitrary for the moment. We consider the following probability density on the space $(\bigotimes_{i=1}^{k_\alpha}\mathsf{X}_{\alpha(i)})^{n+1} \times \Theta$

$$\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta) \propto \nu(\theta)\check{\mu}_{\theta,\alpha(1:k_\alpha)}(x_0(1:k_\alpha))\check{g}_{\theta,\alpha(1:k_\alpha)}(x_0(1:k_\alpha),y_0) \times \prod_{p=1}^{n}\check{f}_{\theta,\alpha(1:k_\alpha)}(x_{p-1}(1:k_\alpha),x_p(1:k_\alpha))\check{g}_{\theta,\alpha(1:k_\alpha)}(x_p(1:k_\alpha),y_p). \tag{9}$$

(5) will be approximated using samples distributed according to this target. In the works [14, 15] the following choice is made

$$\check{g}_{\theta,\alpha(1:k_\alpha)}(x_p(1:k_\alpha), y_p) = \max\{g_{\theta,\alpha(1)}(x_p(1), y_p), \ldots, g_{\theta,\alpha(k_\alpha)}(x_p(k_\alpha), y_p)\}\,,$$

and this is the choice used in this article. It is noted that other choices are possible (see e.g. [9]).

We start by considering a single expectation in (5). Note that (7) and (8), and the form of $\xi_{n,\alpha(1:k_\alpha)}$, immediately imply that for any $\alpha(i)$, $i \in \{1, \ldots, k_\alpha\}$

$$\mathbb{E}_{\pi_{n,\alpha(i)}}[\varphi_{\alpha(i)}(X_{0:n}, \theta)] = \frac{\mathbb{E}_{\xi_{n,\alpha(1:k_\alpha)}}\left[\varphi_{\alpha(i)}(X_{0:n}(i), \theta) \prod_{p=0}^{n} \frac{g_{\theta,\alpha(i)}(X_p(i), y_p)}{\check{g}_{\theta,\alpha(1:k_\alpha)}(X_p(1:k_\alpha), y_p)}\right]}{\mathbb{E}_{\xi_{n,\alpha(1:k_\alpha)}}\left[\prod_{p=0}^{n} \frac{g_{\theta,\alpha(i)}(X_p(i), y_p)}{\check{g}_{\theta,\alpha(1:k_\alpha)}(X_p(1:k_\alpha), y_p)}\right]}\,. \quad (10)$$ eq:ref_2

For ease of notation, we set for any $\alpha(i)$, $i \in \{1, \ldots, k_\alpha\}$

$$H_{i,n,\alpha,\theta}(x_{0:n}(1:k_\alpha)) := \prod_{p=0}^{n} \frac{g_{\theta,\alpha(i)}(x_p(i), y_p)}{\check{g}_{\theta,\alpha(1:k_\alpha)}(x_p(1:k_\alpha), y_p)}.$$

By combining (10) with $H_{i,n,\alpha,\theta}(x_{0:n}(1:k_\alpha))$ and recalling the definitions of Δ and Δ_i given in and above (5), one can then deduce that

$$\Delta\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)] =$$

$$\sum_{i=1}^{k'_\alpha} \tau_{i,\alpha} \left(\frac{\mathbb{E}_{\xi_{n,\alpha(1:k_\alpha)}}[\varphi_{\alpha(2i)}(X_{0:n}(2i), \theta) H_{2i,n,\alpha,\theta}(X_{0:n}(1:k_\alpha))]}{\mathbb{E}_{\xi_{n,\alpha(1:k_\alpha)}}[H_{2i,n,\alpha,\theta}(X_{0:n}(1:k_\alpha))]} - \right.$$

$$\left. \frac{\mathbb{E}_{\xi_{n,\alpha(1:k_\alpha)}}[\varphi_{\alpha(2i-1)}(X_{0:n}(2i-1), \theta) H_{2i-1,n,\alpha,\theta}(X_{0:n}(1:k_\alpha))]}{\mathbb{E}_{\xi_{n,\alpha(1:k_\alpha)}}[H_{2i-1,n,\alpha,\theta}(X_{0:n}(1:k_\alpha))} \right) \quad (11)$$ eq:basic_idea

where $|\alpha| = \sum_{j=1}^{d} \alpha_j$ and $\tau_{i,\alpha} = (-1)^{|\alpha(k_\alpha) - \alpha(2i)|}$. This approach is called approximate coupling; see [16] for further discussion.

Our strategy for approximating $\mathbb{E}_{\pi_{n,\mathcal{I}}}[\varphi(X_{0:n}, \theta)]$ is then the following. Noting (6) and (11), we will approximate each summand in (6), by approximating the r.h.s. of (11). This will be achieved as follows. Independently for each $\alpha \in \mathcal{I}$ (with $k_\alpha > 1$), and serially for each n, we will sample (approximately) from $\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha), \theta)$, and compute a Monte Carlo estimate of $\Delta\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)]$. As noted above, in the case $k_\alpha = 1$, one can simply sample from $\pi_{n,\alpha}$ as no coupling is required.

4 Simulation Strategy

sec:method

The purpose of this Section is to describe a method to approximate $\mathbb{E}_{\pi_n,\mathcal{I}}[\varphi(X_{0:n},\theta)]$. This will be achieved by using the SMC^2 method to approximate expectations w.r.t. $\xi_{n,\alpha(1:k_\alpha)}$ for each $\alpha \in \mathcal{I}$ (with $k_\alpha > 1$), recursively in n. If $k_\alpha = 1$, then one can simply use the standard SMC^2 method considering $\pi_{n,\alpha}$. The SMC^2 approach uses the particle MCMC method andrieu [1], which in turn relies upon particle filters. Therefore, to develop our approach, we first provide a review of particle filters, followed by particle MCMC in the present context.

4.1 Particle Filter

In this section, we focus upon the approximation of the density of the state conditional on fixed parameter θ, given by $\xi_{n,\alpha(1:k_\alpha),\theta}(x_{0:n}(1:k_\alpha)) \propto \xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta)$. It is natural to adopt an SMC approach, in which we sequentially perform importance sampling and then resampling. This procedure is the standard particle filter which is given in Algorithm algo:pf 1.

Algorithm 1 The Particle Filter

- **Initialize**. Set $p = 0$, for $i \in \{1,\ldots,N\}$ sample $x_0^i(1:k_\alpha)$ from $\breve{\mu}_{\theta,\alpha(1:k_\alpha)}$ and evaluate the weight

$$w_{p,\alpha,\theta}^i = \Big(\breve{g}_{\theta,\alpha(1:k_\alpha)}(x_0^i(1:k_\alpha),y_0)\Big)\Big(\sum_{j=1}^N \breve{g}_{\theta,\alpha(1:k_\alpha)}(x_0^j(1:k_\alpha),y_0)\Big)^{-1}$$

- **Iterate**: Set $p = p+1$,
  - Sample $(a_{p-1}^1,\ldots,a_{p-1}^N) \in \{1,\ldots,N\}^N$, where, independently for each $i \in \{1,\ldots,N\}$, $\mathbb{P}(a_{p-1}^i = j) = w_{p-1,\alpha,\theta}^j$.
  - Sample $x_p^i(1:k_\alpha)|x_{p-1}^{a_{p-1}^i}(1:k_\alpha)$ from $\breve{f}_{\theta,\alpha(1:k_\alpha)}(x_{p-1}^{a_{p-1}^i}(1:k_\alpha),\cdot)$, for $i \in \{1,\ldots,N\}$, and evaluate the weight

$$w_{p,\alpha,\theta}^i = \Big(\breve{g}_{\theta,\alpha(1:k_\alpha)}(x_p^i(1:k_\alpha),y_p)\Big)\Big(\sum_{j=1}^N \breve{g}_{\theta,\alpha(1:k_\alpha)}(x_p^j(1:k_\alpha),y_p)\Big)^{-1}.$$

algo:pf

The joint density of all the variables sampled in Algorithm 1, up to time n, is written

$$\psi_{\alpha,\theta}(x_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N})$$

where, for $0 \leq k \leq n$, $x_k^{1:N}(1:k_\alpha) = (x_k^1(1:k_\alpha), \ldots, x_k^N(1:k_\alpha)) \in (\bigotimes_{i=1}^{k_\alpha} \mathsf{X}_{\alpha(i)})^N$, $a_k^{1:N} = (a_k^1, \ldots, a_k^N) \in \{1, \ldots, N\}^N$, $x_{0:n}^{1:N}(1:k_\alpha) = (x_0^{1:N}(1:k_\alpha), \ldots, x_n^{1:N}(1:k_\alpha)) \in (\bigotimes_{i=1}^{k_\alpha} \mathsf{X}_{\alpha(i)})^{(n+1)N}$, $a_{0:n-1}^{1:N} = (a_0^{1:N}, \ldots, a_{n-1}^{1:N}) \in \{1, \ldots, N\}^{nN}$. In particular, a_{n-1}^i is the index at time $n-1$ of the resampled particle which has the index i at time n. For each $i \in \{1, \ldots, N\}$ define the ancestral lineage indices as

$$b_n^i = i \quad \text{and} \quad b_k^i = a_k^{b_{k+1}^i}, \quad k \in \{0, \ldots, n-1\}. \tag{12}$$ eq:lineage

For each $i = 1, \ldots, N$, define

$$\bar{x}_{0:n}^i(1:k_\alpha) := x_{0:n}^{b_{0:n}^i}(1:k_\alpha). \tag{13}$$ eq:smoothingsample

AJAY: This is required to avoid the redundant notation which is used later, and it had somehow gotten lost in the iterations (at some places, but it remained in Algorithm 3). The following empirical measure then provides an approximation of $\xi_{n,\alpha(1:k_\alpha),\theta}(x_{0:n}(1:k_\alpha))$

$$\sum_{i=1}^N w_{n,\alpha,\theta}^i \delta_{\bar{x}_{0:n}^i(1:k_\alpha)}(dx_{0:n}). \tag{14}$$ eq:target_approx

This will prove useful in the next section.

The normalization constant $Z_{n,\alpha,\theta} = \int_{\bigotimes_{i=1}^{k_\alpha} \mathsf{X}_{\alpha(i)}^{n+1}} \xi_{n,\alpha(1:k_\alpha),\theta}(x_{0:n}(1:k_\alpha)) dx_{0:n}(1:k_\alpha)$ can be unbiasedly estimated by [7] delm:04

$$Z_{n,\alpha,\theta}^N = \prod_{p=0}^n \Big(\frac{1}{N} \sum_{i=1}^N \check{g}_{\theta,\alpha(1:k_\alpha)}(x_p^i(1:k_\alpha), y_p) \Big). \tag{15}$$ eq:nc_est

It is noted that particle filters often do not work well in high-dimensions (e.g. [21] snyder). However, in some cases, where the target probability is a high and finite dimensional discretization of an infinite dimensional distribution (as will be the case in the context of this article), the algorithm can work quite well; see e.g. [19] kantas.

4.2 Particle MCMC

In this section, we focus on approximating expectations w.r.t. $\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha), \theta)$, except for fixed n. The notation $\bar{x}_{0:n}^i(1:k_\alpha), \theta^i$, to be used, will refer to the i^{th} sample of

a Markov chain designed to approximate expectations w.r.t. $\xi_{n,\alpha(1:k_\alpha)}$. In the algorithms to be presented, $r(\theta^{i-1},\cdot)$ is a proposal density on Θ which we will assume is a postive probability density w.r.t. $d\theta$ for any θ^{i-1}.

In Algorithm 2 (resp. Algorithm 3) we present an approach to approximate expectations w.r.t. $\xi_{n,\alpha(1:k_\alpha)}(\theta)$ (resp. $\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta)$). The MCMC approach in Algorithm 2 (resp. Algorithm 3)) is called marginal particle MCMC (PMCMC) (resp. PMCMC).

Algorithm 2 Marginal PMCMC Algorithm

- **Initialize**. Set $i=0$ and sample θ^0 from the prior. Given θ^0 run the particle filter in Algorithm 1 and record the estimate of Z^N_{n,α,θ^0} from eq. (15).
- **Iterate**:
  - (I) Set $i=i+1$ and propose θ' given θ^{i-1} from a proposal $r(\theta^{i-1},\cdot)$ (described in the main text).
  - (II) Given θ' run the particle filter in Algorithm 1 and record the estimate $Z^N_{n,\alpha,\theta'}$.
  - (III) Set $\theta^i=\theta'$ with probability

$$\min\left\{1,\frac{Z^N_{n,\alpha,\theta'}\nu(\theta')r(\theta',\theta^{i-1})}{Z^N_{n,\alpha,\theta^{i-1}}\nu(\theta^{i-1})r(\theta^{i-1},\theta')}\right\}$$

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

alg:pmcmcm

The target density associated to the PMCMC kernel described in Algorithm 3, on the state-space $\Theta\times\left(\bigotimes_{i=1}^{k_\alpha}\mathsf{X}_{\alpha(i)}^{n+1}\right)^N\times\{1,\ldots,N\}^{Nn+1}$ is ([1, Theorem 4])

$$\tilde{\xi}_{n,\alpha(1:k_\alpha)}(x^{1:N}_{0:n}(1:k_\alpha),a^{1:N}_{0:n-1},\theta,s):=$$
$$\frac{\xi_{n,\alpha(1:k_\alpha)}(\bar{x}^s_{0:n}(1:k_\alpha),\theta)}{N^{n+1}}\frac{\psi_{\alpha,\theta}(x^{1:N}_{0:n}(1:k_\alpha),a^{1:N}_{0:n-1})}{\check{\mu}_{\theta,\alpha(1:k_\alpha)}(x_0^{b_0^s}(1:k_\alpha))\prod_{p=1}^n w^{b_{p-1}^s}_{p-1,\alpha,\theta}\check{f}_{\theta,\alpha(1:k_\alpha)}(x_{p-1}^{b_{p-1}^s}(1:k_\alpha),x_p^{b_p^s}(1:k_\alpha))},$$

where b_p^s is defined in (12). [1] show that $\bar{x}^s_{0:n}(1:k_\alpha),\theta$ have marginal density $\xi_{n,\alpha(1:k_\alpha)}$. As a result, consider

$$\int_{\bigotimes_{i=1}^{k_\alpha}\mathsf{X}_{\alpha(i)}^{n+1}\times\Theta}\varphi_{\alpha(j)}(x_{0:n}(1:k_\alpha),\theta)\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta)dx_{0:n}(1:k_\alpha)d\theta$$

for integrable and real-valued functions $\varphi_{\alpha(j)}$ and any $j\in\{1,\ldots,k_\alpha\}$. [1] show that, for Algorithm 3, this above quantity is consistently estimated (that is, the estimate converges

Algorithm 3 PMCMC Algorithm

- **Initialize**. As in Algorithm 3. Select a trajectory $x^i_{0:n}(1:k_\alpha)$ from the particle filter just run using (14), denote the stored state $\overline{x}^0_{0:n}(1:k_\alpha)$.

- **Iterate**:
  - (I-II) as in Algorithm 3.
  - (II.b) Select a trajectory $x^{s'}_{0:n}(1:k_\alpha)$ from the particle filter just run using (14).
  - (III) as in Algorithm 3.
  - (III.b) Let $\overline{x}^i_{0:n}(1:k_\alpha) = x^{s'}_{0:n}(1:k_\alpha)$ if $\theta^i = \theta'$, otherwise let $\overline{x}^i_{0:n}(1:k_\alpha) = \overline{x}^{i-1}_{0:n}(1:k_\alpha)$ if $\theta^i = \theta^{i-1}$.

alg:pmcmc

almost surely as $N \to +\infty$) by:

$$\frac{1}{N}\sum_{i=1}^{N} \varphi_{\alpha(j)}(\overline{x}^i_{0:n}(1:k_\alpha), \theta^i)\,.$$

As a result, for n fixed, one can approximate the r.h.s. of (11). We hence refer to using Algorithm 3 in the context of (11) within (6) as MI-PMCMC.

sec:smc2

4.3 SMC^2

To consider the method to be discussed, we start with some definitions. We define the spaces:

$$\begin{aligned}
\mathsf{E}_0 &:= \Big(\bigotimes_{i=1}^{k_\alpha} \mathsf{X}_{\alpha(i)}\Big)^N \times \Theta \\
\mathsf{E}_n &:= \Big(\bigotimes_{i=1}^{k_\alpha} \mathsf{X}^{n+1}_{\alpha(i)}\Big)^N \times \{1,\ldots,N\}^{nN} \times \Theta \quad n \geq 1
\end{aligned}$$

and states

$$\begin{aligned}
U_{0,\alpha} &:= (X_0^{1:N}(1:k_\alpha), \theta) \\
U_{n,\alpha} &:= (X_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N}, \theta) \quad n \geq 1
\end{aligned}$$

and note $U_{0,\alpha} \in \mathsf{E}_0$, $U_{n,\alpha} \in \mathsf{E}_n$, $n \geq 1$.

Algorithm 4 An SMC^2 Algorithm targetting $\tilde{\xi}_{n,\alpha(1:k_\alpha)}(x_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N}, \theta)$

- **Initialize**. Set $p = 0$, for $i \in \{1, \ldots, N_\alpha\}$ sample θ^i from the prior ν, and $X_0^{i,j}(1:k_\alpha)$ from $\check{\mu}_{\theta^i,\alpha(1:k_\alpha)}(\cdot)$, $j \in \{1, \ldots, N\}$. Compute the weight:

$$G_{0,\alpha}(u_{0,\alpha}^i) = \frac{1}{N}\sum_{j=1}^{N} \check{g}_{\theta^i,\alpha(1:k_\alpha)}(x_0^{i,j}(1:k_\alpha), y_0).$$

- **Iterate**:

  - **Select**: Set $p = p+1$, resample $u_{p-1,\alpha}^{1:N_\alpha}$ using the normalized $\{G_{p-1,\alpha}(u_{p-1,\alpha}^i)\}_{i=1}^{N_\alpha}$, denoting the resulting samples $\hat{u}_{p-1,\alpha}^{1:N_\alpha}$.

  - **Mutate**: For $i \in \{1, \ldots, N_\alpha\}$ generate $\tilde{u}_{p-1,\alpha}^i | \hat{u}_{p-1,\alpha}^i$ using one iteration of Algorithm 2.

  - **Extend**: For $i \in \{1, \ldots, N_\alpha\}$ sample $X_p^{i,j}(1:k_\alpha), a_{p-1}^{i,j}$, $j \in \{1, \ldots, N\}$ from

$$\prod_{j=1}^{N} \frac{\check{g}_{\tilde{\theta}^i,\alpha(1:k_\alpha)}(\tilde{x}_{p-1}^{i,a_{p-1}^{i,j}}(1:k_\alpha), y_{p-1})}{\sum_{l=1}^{N} \check{g}_{\tilde{\theta}^i,\alpha(1:k_\alpha)}(\tilde{x}_{p-1}^{i,l}(1:k_\alpha), y_{p-1})} \check{f}_{\tilde{\theta}^i,\alpha(1:k_\alpha)}(\tilde{x}_{p-1}^{i,a_{p-1}^{i,j}}(1:k_\alpha), x_p^{i,j}(1:k_\alpha)).$$

    Set $u_{p,\alpha}^i = (\tilde{u}_{p-1,\alpha}^i, x_p^{i,1:N}(1:k_\alpha), a_{p-1}^{i,1:N})$.

  - **Compute the weight**: For $i \in \{1, \ldots, N_\alpha\}$

$$G_{p,\alpha}(u_{p,\alpha}^i) = \frac{1}{N}\sum_{j=1}^{N} \check{g}_{\theta^i,\alpha(1:k_\alpha)}(x_p^{i,j}(1:k_\alpha), y_p).$$

alg:smc2

We now introduce the SMC^2 method in Algorithm 4. This method is a type of particle filter which targets a sequence of probability distributions $\{\tilde{\xi}_{n,\alpha(1:k_\alpha)}(x_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N}, \theta)\}_{n\geq 0}$ (see [5, Proposition 1] for a justification) which admit $\{\xi_{n,\alpha(1:k_\alpha)}(\theta)\}_{n\geq 0}$ as a particular marginal; we explain how one can estimate expectations w.r.t. $\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha), \theta)$ below. The convergence (as $N_\alpha \to \infty$) of such an algorithm, then follows the theory of particle approximations of Feynman-Kac formulae as in [7] for instance.

4.3.1 Intuition of the Algorithm

To understand the intution of the algorithm, consider approximating expectations w.r.t. $\tilde{\xi}_{0,\alpha(1:k_\alpha)}(x_0^{1:N}(1:k_\alpha),\theta) \propto \Big(\frac{1}{N}\sum_{j=1}^{N}\check{g}_{\theta,\alpha(1:k_\alpha)}(x_0^j(1:k_\alpha),y_0)\Big)\Big(\prod_{j=1}^{N}\check{\mu}_{\theta^i,\alpha(1:k_\alpha)}(x_0^j(1:k_\alpha))\Big)\nu(\theta)$. This can be achieved by (self-normalized) importance sampling, just as in the initialization step of Algorithm 4. This is because the term $G_{0,\alpha}(u_{0,\alpha})$ is an importance weight, which allows one to correct for the discrepancy between the distribution sampled (in the initialization step of Algorithm 4) and the one of interest.

We now want to move our samples, in such a way as to approximate expectations w.r.t. $\tilde{\xi}_{1,\alpha(1:k_\alpha)}(x_{0:1}^{1:N}(1:k_\alpha),a_0^{1:N},\theta)$. This can be achieved in the iterate step of Algorithm 4, as we now explain. In the select step, this is a resampling of the samples, just as in the particle filter. The resulting samples are approximately sampled from $\tilde{\xi}_{0,\alpha(1:k_\alpha)}(x_0^{1:N}(1:k_\alpha),\theta)$. The mutate step will now produce new samples which are still approximately sampled from $\tilde{\xi}_{0,\alpha(1:k_\alpha)}(x_0^{1:N}(1:k_\alpha),\theta)$, as the transition kernel leaves this probability invariant. The extend step now produces the additional random variables needed to approximate expectations w.r.t. $\tilde{\xi}_{1,\alpha(1:k_\alpha)}(x_{0:1}^{1:N}(1:k_\alpha),a_0^{1:N},\theta)$. The strategy employed leads to the convenient weight function $G_{1,\alpha}(u_{1,\alpha})$ in the next step. This corresponds to the ratio, up-to a normalizing constant, of $\tilde{\xi}_{1,\alpha(1:k_\alpha)}(x_{0:1}^{1:N}(1:k_\alpha),a_0^{1:N},\theta)$ to $\tilde{\xi}_{0,\alpha(1:k_\alpha)}(x_0^{1:N}(1:k_\alpha),\theta)$ multipled by the proposal mechanism in the extend step. Expectations w.r.t. $\tilde{\xi}_{1,\alpha(1:k_\alpha)}(x_{0:1}^{1:N}(1:k_\alpha),a_0^{1:N},\theta)$ can now be approximated again by self-normalized importance sampling. The algorithm then just continues for the rest of the sequence $\{\tilde{\xi}_{n,\alpha(1:k_\alpha)}(x_{0:n}^{1:N}(1:k_\alpha),a_{0:n-1}^{1:N},\theta)\}_{n\geq 2}$.

The reason why one considers the sequence $\{\tilde{\xi}_{n,\alpha(1:k_\alpha)}(x_{0:n}^{1:N}(1:k_\alpha),a_{0:n-1}^{1:N},\theta)\}_{n\geq 0}$, instead of the original $\{\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta)\}_{n\geq 0}$, is because the associated algorithm with the former is expected to be more efficient than a related SMC algorithm for the latter. To explain further, one can consider first, the naive algorithm which samples the initial θ from the prior (i.e. N samples) and then runs a particle filter with N associated trajectories $x_{0:n}(1:k_\alpha)$. The main issue here is of course as one never updates the θ, that estimates of expectations associated to θ would be very poor. This is further exacerbated by the path degeneracy problem for particle filters; one does not update the trajectory in the

past and due to the resampling operation the uniqueness of the trajectories in the past will be essentially lost. These latter issues can be circumvented by applying an MCMC kernel of invariant measure $\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta)$ at each time step to each sample. However, as we have remarked, in general PMCMC is though to be the most 'efficient' method to obtain samples from $\xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta)$, hence our use of SMC^2. See [5, 18] for further insights.

4.3.2 Estimating Expectations w.r.t. $\tilde{\xi}_{n,\alpha(1:k_\alpha)}(x_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N}, \theta, s)$

At any time p, sample $S_p^i \in \{1,\dots,N\}$, $i \in \{1,\dots,N_\alpha\}$ with probability

$$\mathbb{P}(S_p^i = j | \tilde{u}_{p,\alpha}^i) = \frac{\check{g}_{\tilde{\theta}^i,\alpha(1:k_\alpha)}(\tilde{x}_p^{i,j}(1:k_\alpha), y_p)}{\sum_{l=1}^N \check{g}_{\tilde{\theta}^i,\alpha(1:k_\alpha)}(\tilde{x}_p^{i,l}(1:k_\alpha), y_p)}. \tag{16}$$ eq:s_def

For $p' \leq p$ define

$$v_{p'}^i(1:k_\alpha) := \tilde{x}_{p'}^{i,b_{p'}^{s_p^i}}(1:k_\alpha)\,,$$

and denote the augmented empirical measure of $(\tilde{u}_{p,\alpha}^{1:N_\alpha}, s_p^{1:N_\alpha})$ as

$$\eta_{p,\alpha}^{N_\alpha} := \frac{1}{N}\sum_{l=1}^{N_\alpha} \delta_{(\tilde{u}_{p,\alpha}^l, s_p^l)}\,.$$

Then for $(i,j) \in \{1,\dots,k_\alpha\}^2$, $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$, $H : \mathbb{N}_0^d \times \bigotimes_{l=1}^{k_\alpha} \mathsf{X}_{\alpha(l)}^{n+1} \times \Theta \to \mathbb{R}$ (measurable and integrable w.r.t. $\xi_{n,\alpha(1:k_\alpha)}$) one can consistently estimate

$$\begin{aligned}
&\int_{\mathsf{E}_p \times \{1,\dots,N\}} \varphi_{\alpha(i)}(\bar{x}_{0:n}^s(i),\theta) H_\alpha(\bar{x}_{0:n}^s(1:k_\alpha),\theta) \tilde{\xi}_{n,\alpha(1:k_\alpha)}(x_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N}, \theta, s) d(x_{0:n}^{1:N}(1:k_\alpha), a_{0:n-1}^{1:N}, \theta, s) \\
&\quad = \int_{\bigotimes_{l=1}^{k_\alpha} \mathsf{X}_{\alpha(l)}^{n+1} \times \Theta} \varphi_{\alpha(i)}(x_{0:n}(i),\theta) H_\alpha(x_{0:n}(1:k_\alpha),\theta) \xi_{n,\alpha(1:k_\alpha)}(x_{0:n}(1:k_\alpha),\theta) d(x_{0:n}(1:k_\alpha),\theta)
\end{aligned}$$

with

$$\eta_{n,\alpha}^{N_\alpha}(\varphi_{\alpha(i)} H_\alpha) := \frac{1}{N_\alpha}\sum_{l=1}^{N_\alpha} \varphi_{\alpha(i)}(v_{0:n}^l(i), \tilde{\theta}^l) H_\alpha(v_{0:n}^l(1:k_\alpha), \tilde{\theta}^l)\,.$$

Hence we have the estimate of (11) as eq:basic_idea

$$\Delta\mathbb{E}_{\pi_{n,\alpha}}^{N_\alpha}[\varphi_\alpha(X_{0:n},\theta)] := \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\left(\frac{\eta_{n,\alpha}^{N_\alpha}(\varphi_{\alpha(2i)} H_{2i,n,\alpha,\theta})}{\eta_{n,\alpha}^{N_\alpha}(H_{2i,n,\alpha,\theta})} - \frac{\eta_{n,\alpha}^{N_\alpha}(\varphi_{\alpha(2i-1)} H_{2i-1,n,\alpha,\theta})}{\eta_{n,\alpha}^{N_\alpha}(H_{2i-1,n,\alpha,\theta})}\right),$$

where we remind the reader that $\tau_{i,\alpha} = (-1)^{|\alpha(k_\alpha)-\alpha(2i)|}$.

sec:theory

5 Theoretical Results

We now consider the SMC^2 procedure in the previous section. We will analyze the variance of our MI method, under the following assumptions.

hyp:A (**A1**) There exist $0 < \underline{C} < \overline{C} < +\infty$ such that for every $\alpha \in \mathcal{I}$, $\theta \in \Theta$, $(x, y) \in \mathsf{X}_\alpha \times \mathsf{Y}_\alpha$

$$\underline{C} \leq g_{\theta,\alpha}(x, y) \leq \overline{C}.$$

hyp:B (**A2**) For every $n \geq 0$, $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$ bounded, every $\alpha \in \mathcal{I}$, there exist a $C(\alpha(1 : k_\alpha))$, with $\lim_{\min_{1\leq i\leq d} \alpha_i \to +\infty} C(\alpha(1 : k_\alpha)) = 0$, such that for any collection of scalar, bounded random variables $\beta(\alpha(1 : k_\alpha), 2i, 2i - 1)$, $i \in \{1, \ldots, k'_\alpha\}$ we have almost surely

$$\sup_{(x_{0:n}(1:k_\alpha),\theta)\in(\bigotimes_{i=1}^{k_\alpha} \mathsf{X}_{\alpha(i)}^{n+1})\times\Theta} \left|\left\{ \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\beta(\alpha(1 : k_\alpha), 2i, 2i - 1)\Big[\varphi_{\alpha(2i)}(x_{0:n}(1 : k_\alpha), \theta)- \right.\right.$$

$$\left.\left. \varphi_{\alpha(2i-1)}(x_{0:n}(1 : k_\alpha), \theta)\Big]\right\}\right| \leq C(\alpha(1 : k_\alpha)) \sum_{i=1}^{k'_\alpha} |\beta(\alpha(1 : k_\alpha), 2i, 2i - 1)|^2.$$

We remind the reader again that $\tau_{i,\alpha} = (-1)^{|\alpha(k_\alpha)-\alpha(2i)|}$.

(A1) is a strong, but standard, assumption that has been used in the HMM literature, particularly in the SMC context; see for instance [7]. (A2) is certainly non-standard and in general one would like to deduce under simpler hypotheses. In our efforts to achieve this, we have not found a suitable technical approach and leave this more involved analysis to future work. The following result is the culmination of our efforts. The expectation below is w.r.t. the randomness in the SMC^2 algorithm.

thm:main **Theorem 5.1.** *Assume (A1-2). Then for every $n \geq 0$, $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$ bounded and every $\alpha \in \mathcal{I}$, there exist a $C(\alpha(1 : k_\alpha))$, with $\lim_{\min_{1\leq i\leq d} \alpha_i \to +\infty} C(\alpha(1 : k_\alpha)) = 0$, such that:*

$$\mathbb{E}\left[\left(\Delta\mathbb{E}^{N_\alpha}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)] - \Delta\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)]\right)^2\right] \leq \frac{C(\alpha(1 : k_\alpha))}{N_\alpha}$$

and

$$\left|\mathbb{E}\left[\Delta\mathbb{E}^{N_\alpha}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)] - \Delta\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)]\right]\right| \leq \frac{C(\alpha(1 : k_\alpha))}{N_\alpha}.$$

Proof. Follows directly from Lemma A.1 and Proposition A.1 in the appendix. □

It is noted that our bound depends upon the time parameter and d and we do not address these aspects in our subsequent discussion.

5.1 MIMC considerations

ssec:mimc

Define a multi-index estimator as

$$\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} := \sum_{\alpha \in \mathcal{I}} \Delta \mathbb{E}_{\pi_{n,\alpha}}^{N_\alpha}[\varphi_\alpha(X_{0:n}, \theta)].$$

Below $\mathrm{Cost}(X_\alpha)$ denotes the cost of sampling the discretized random variable X_α. Recall $C(\alpha(1:k_\alpha))$ appears in Theorem 5.1 and Assumption (A2).

ass:mimc2

Assumption 5.1 (MISMC^2 rates)**.** *For every $n \geq 0$, there exists $C < +\infty$, $w_i, \beta_i, \gamma_i > 0$ for $i = 1, \ldots, d$, such that for every $\alpha \in \mathbb{N}_0^d$:*

(a) $\left|\Delta \mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n}, \theta)]\right| \leq C \prod_{i=1}^d 2^{-w_i \alpha_i}$;

(b) $C(\alpha(1:k_\alpha)) \leq C \prod_{i=1}^d 2^{-\beta_i \alpha_i}$;

(c) $\mathrm{Cost}(X_\alpha) \leq C \prod_{i=1}^d 2^{\gamma_i \alpha_i}$.

Before presenting the main MISMC^2 theorem, we need to introduce some index sets, which relate to Assumption 5.1. The tensor product index set is defined by

$$\mathcal{I}_{\alpha^*} := \{\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d : \alpha_1 \in \{0, \ldots, \alpha_1^*\}, \ldots, \alpha_d \in \{0, \ldots, \alpha_d^*\}\}. \tag{17}$$

eq:tensor

The total degree index set for $L \in \mathbb{R}_+$ and $\zeta \in \mathbb{R}_+^d$ is defined as

$$\mathcal{I}_{\zeta,L}^{\mathrm{TD}} := \left\{\alpha \in \mathbb{N}_0^d : \sum_{i=1}^d \alpha_i \zeta_i \leq L\right\}. \tag{18}$$

eq:tensor

It is suggested in Sec. 2.2 of [12] (and verified in numerical examples) that the optimal index set is given by $\mathcal{I}_L^{\mathrm{TD}} = \mathcal{I}_{\zeta^*,L}^{\mathrm{TD}}$, where $z_i^* \propto \log(2)(w_i + (\gamma_i - \beta_i)/2)$ and $\sum_{i=1}^d z_i^* = 1$. See [12] for a detailed investigation of the various relationships between the rates of convergence and these index sets. The methodology developed is applicable to general index sets, but we will present the theorem for only a simplified set of circumstances in the interest of clarity and simplicity.

pro:mimcmc **Proposition 5.1.** *Assume (A1-2), Assumption 5.1 and that $\beta_i > \gamma_i$, for all $i = 1, \dots, d$, and one of the following cases holds*

[A] *$\mathcal{I} = \mathcal{I}_{\alpha^*}$ and $\sum_{i=1}^d \gamma_i / w_i \leq 2$; or*

[B] *$\mathcal{I} = \mathcal{I}_L^{\mathrm{TD}}$.*

Then there exist $C < +\infty$, and either $\alpha^* = (m_1, \dots, m_d) \in \mathbb{N}_0^d$ in case [A] *or $L \in \mathbb{R}_+$ in case* [B]*, and $\{N_\alpha\}_{\alpha \in \mathcal{I}}$, such that for any $\varepsilon > 0$:*

$$\mathbb{E}\left[\left(\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} - \mathbb{E}[\varphi(X_{0:n}, \theta)]\right)^2\right] \leq C\varepsilon^2 \ ,$$

for a cost of $\mathcal{O}(\varepsilon^{-2})$.

Proof. Under the assumptions above, and following from Theorem 5.1, the proof for case [A] is the same as that of Proposition 3.2 in [14]. Case [B] follows from Theorem 2.2 of [12] (see also Theorem 2 of [11]).

□

This can be readily generalized to different relationship between the coefficients (w_i, β_i, γ_i). There are many different cases in general, but the rules of thumb are that (i) the complexity has a logarithmic penalty if $\beta_j \leq \gamma_j$ for any j, and (ii) there is a smaller exponent on ε as well if $\beta_j < \gamma_j$ for any j. The various conditions can be derived in a similar manner as in [12] (see also [11] and [14] for some discussion). Note that if $\sum_{i=1}^d \gamma_i / w_i > 2$ instead in case [A] then the cost is $\varepsilon^{-\sum_{i=1}^d \gamma_i / w_i}$, corresponding to the cost of a single realization at the finest discretization. In this case, the cost of MLMC will still be at least as large, because a single realization at the finest discretization of the tensor product index set will always be required.

sec:numerics

6 Numerical Results

ssec:set-up

6.1 Modelling

We illustrate the performance of the proposed methods on the Bayesian parameter inference problem of a partially observed stochastic system which is the solution to a SPDE. Comparisons are made with sampling from the most precise discretization of the underlying stochastic system using either PMCMC or SMC^2. The objective here is to illustrate the theory and test the applicability of the method under weaker assumptions than provided by the theory. Therefore, we will restrict attention to the total degree index set $\mathcal{I}_\alpha$, despite its suboptimality in this example.

We consider the stochastic heat equation on a one-dimensional domain $[0,1]$ over the time interval $[0,T]$, i.e.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + au + \theta \dot{W}_t$$

with the Dirichlet boundary condition and initial value $u(x,0) = u_0(x) = \sum_{k=1}^{\infty} u_{k,0}\, e_k(x)$ for $x \in (0,1)$. The eigenfunction $e_k(x) = \sqrt{2}\sin(k\pi x)$ has the corresponding eigenvalue $\lambda_k = k^2\pi^2$ and the noise W_t is the space-time white noise, i.e. the cylindrical Brownian motion given by $W_t = \sum_{k=1}^{\infty} \sqrt{q_k} e_k \beta_t^k$, where β_t^k $(k \geq 1)$ are i.i.d. scalar Brownian motions. The hidden process is assumed to be modelled by the solution to this SPDE with $q_k = 1$ and $u_{k,0} = 1$ for $1 \leq k \leq K_{\max}$ and $u_{k,0} = 0$ otherwise.

Pointwise observations of the process are obtained at times $t(n) = n\delta$ for $n = 1, 2, ..., 100$ and $\delta = 0.001$, and at the locations $x_1 = 1/3$ and $x_2 = 2/3$ under an additive Gaussian noise with mean zero and variance $\tau^2 = 1$. If we denote the observation vector at time $t(n)$ by $y_n = (y_{n,1}, y_{n,2})^T$, the corresponding likelihood function is

$$g(x_n, y_n) \propto \prod_{i=1}^{2} \exp\left(-\frac{1}{2\tau^2}\left(y_{n,i} - u(x_i, t(n))\right)^2\right)$$

where $u(x_i, t(n))$ is the solution of the above SPDE at time $t(n)$ and location x_i and note that $u(x,t) = \sum_{k=1}^{\infty} u_{k,t}\, e_k(x)$. The model parameter θ is assumed to be unknown and is assigned a prior distribution Gamma$(1, \sqrt{0.1})$ where Gamma(a,b) represents the Gamma distribution with shape parameter a and scale parameter b. A fixed sequence of observations

$y_{1:100}$ is simulated with $a = 1/2$ and $\theta^2 = 0.1$.

The problem of interest is the Bayesian static parameter estimation of θ from the above-mentioned model sequentially for each n as the data arrives. Our ultimate goal is to approximate $\mathbb{E}_{\pi_n}[\varphi(\theta)]$, where $\varphi(\theta) = \theta$ and π_n is the posterior density of θ, given $y_{1:n}$, induced by the HMM with no discretization bias. In this case, we are interested in the posterior mean of the model parameter θ.

Given the approximation multi-index $\alpha^* = (m_x, m_t)$, we will estimate $\mathbb{E}_{\pi_{n,\mathcal{I}_{\alpha^*}}}[\varphi(\theta)] = \mathbb{E}_{\pi_{n,\alpha^*}}[\varphi^*_\alpha(\theta)]$ to approximate $\mathbb{E}_{\pi_n}[\varphi(\theta)]$, where π_{n,α^*} is the posterior distribution associated with the multi-index α^* and π_n is the target posterior distribution.

We adopt the exponential Euler scheme developed in [17] for discretizing the underlying hidden process. To be precise, at a multi-index $\alpha = (\alpha_x, \alpha_t)$, the above SPDE is solved with the first $K_\alpha = K_0 \times 2^{\alpha_x}$ eigenfunctions and $M_\alpha = M_0 \times 2^{\alpha_t}$ time steps as follows

$$u_{\alpha,k,i+1} = e^{-\lambda_k h} u_{\alpha,k,i} + \frac{1 - e^{-\lambda_k h}}{\lambda_k} a u_{\alpha,k,i} + r_{k,i} \tag{19}$$

where $r_{k,i} \sim N\left(0, \frac{\theta^2(1-e^{-2\lambda_k h})}{2\lambda_k}\right)$ for $k = 1, ..., K_\alpha$ and $i = 0, 1, ..., M_\alpha - 1$. The time step-size $h = \delta/M_\alpha$ and $u_{\alpha,k,i}$ is the solution for the coefficient associated with the k^{th} eigenfunction, i^{th} time step and the discretization index α.

The coupling of the k_α $(1 \le k_\alpha \le 4)$ discretized probability laws is constructed as follows. We start with the simulation of the most expensive random variable that corresponds to the multi-index α. For simulations involving $\alpha_x - 1$, only the subset of the first $K_{\alpha - e_x}$ components are retained. For simulations involving $\alpha_t - 1$, $r_{k,i}$ in (19) is replaced by $\widehat{r}_{k,i} = e^{-\lambda_k h}\widetilde{r}_{k,2i} + \widetilde{r}_{k,2i+1}$ [4] for $i = 0, 1, ..., M_{\alpha - e_t} - 1$, where $\{\widetilde{r}_{k,i}\}_{i=0}^{M_\alpha - 1}$ are simulated with respect to the multi-index α.

Assumption 5.1 (b) was verified directly by estimating the quantity in Theorem 5.1 using the empirical variance over 20 multi-increment estimators. The values $\beta_x = 1$ and $\beta_t = 2$ were fit, for $\gamma_x = \gamma_t = 1$, which is consistent with the results in [14]. We also assume $w_i = \beta_i/2$, as in [14]. It is noted that assumption (A2) is likely not satisfied in this example, and so the numerical results are testing the applicability of the method under weaker assumptions than provided by the theory.

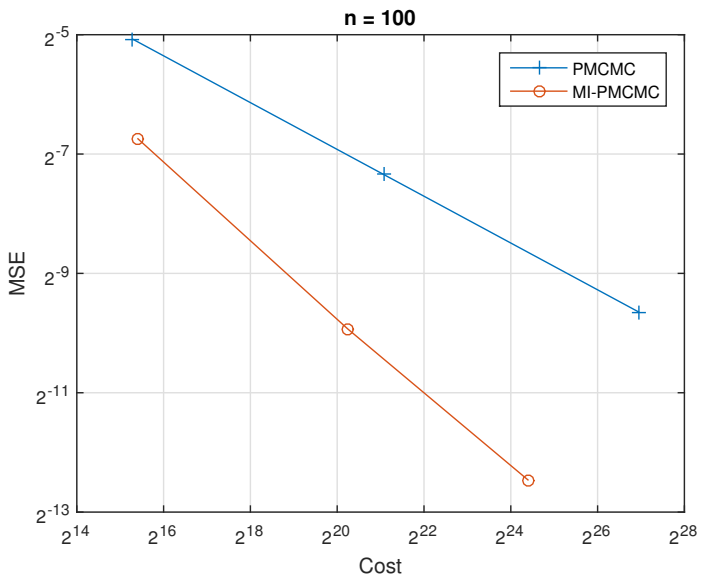

Figure 2: MSE v.s. Cost at time index $n = 100$ for the PMCMC method

fg:PMCMC

6.2 Results using PMCMC

sec:numpmcmc

We consider the estimation of the posterior mean of the model parameter θ in this section with n fixed and $n = 100$. The proposed MI-PMCMC method is implemented, as well as a standard PMCMC at the finest discretization level. As shown by (19), the $u_{\alpha,k,i}$ can be solved by Kalman filter since it fits in the framework of linear Gaussian state-space models. As a result, MCMC with the true likelihoods calculated from the Kalman techniques is implemented to produce the benchmark for computing the MSE of the approximations.

Following the optimal choice of discretization $K = M^2$ as discussed in [17], the cost for a single realization is proportional to M^3. This results in the optimal cost for the ordinary PMCMC of $\mathcal{O}(\varepsilon^{-5})$. For the MI-PMCMC method, we let $m_x = 2m_t \geq 2\log(\varepsilon/2)$ and the optimal $N_\alpha \propto \varepsilon^{-2} m_x 2^{-\alpha_x - 3\alpha_t/2}$, as discussed in [14]. The cost is dominated by $\mathcal{O}(\varepsilon^{-3})$ (where $3 = \sum_{i=x,t} \gamma_i/w_i$ – see discussion following Proposition 5.1). Both algorithms are implemented for 20 runs and with the most precise discretization index $\alpha^* = (2,1),(4,2),(6,3)$. The MSE vs cost plot is illustrated in Figure 2. The cost rates are verified numerically as in Figure 2 and are consistent with [14]. The fitted rate is about -5.1 for the ordinary PMCMC method and -3.1 for MI-PMCMC. It is noted that in the

context of ML-PMCMC for this example, i.e. refining once in both (x, t) at each level, one will find $\gamma = 3$, $\beta = 2$, $\alpha = \beta/2$, and $2 + (\gamma - \beta)/\alpha = 3 (= \gamma/\alpha)$. In other words, the rate is the same as we obtain here for MI-PMCMC [11].

6.3 Results on SMC^2

The proposed SMC^2 method as well as the ordinary SMC^2 method are implemented with the most precise discretization indices $\alpha^* = (2, 1), (4, 2), (6, 3), (8, 4)$ and N is fixed with $N = 500$. The proposed SMC^2 method is run with the optimal choice of $N_\alpha \propto \varepsilon^{-2} m_x 2^{-\alpha_x - 3\alpha_t/2}$ as discussed in [14] and subsection 6.2. The ground truth in this case is calculated by the weighted average of the θ particles from the iterated batch importance sampling algorithm [6] with true likelihood increments derived from the Kalman techniques, which is used for computing the MSE of the approximations. Both algorithms are implemented for 20 runs.

The same rate is expected as in subsection 6.2, under the same choices of (m_x, m_t) and N_α. This is verified numerically, as illustrated in Figure 3, which displays the MSE vs cost plot at different time index $n \in \{50, 65, 80, 100\}$. The fitted rate is about -5.2 for the ordinary SMC^2 and -3 for the multi-index SMC^2 method.

Acknowledgements

We would like to thank Abdul-Lateef Haji-Ali for useful discussions relating to the material in this paper. AJ was supported by an AcRF tier 2 grant: R-155-000-161-112. AJ is affiliated with the Risk Management Institute, the Center for Quantitative Finance and the OR & Analytics cluster at NUS. AJ was supported by a KAUST CRG4 grant ref: 2584. KJHL was supported by the School of Mathematics at the University of Manchester and The Alan Turing Institute. He was also funded in part by Oak Ridge National Laboratory Directed Research and Development Seed funding.

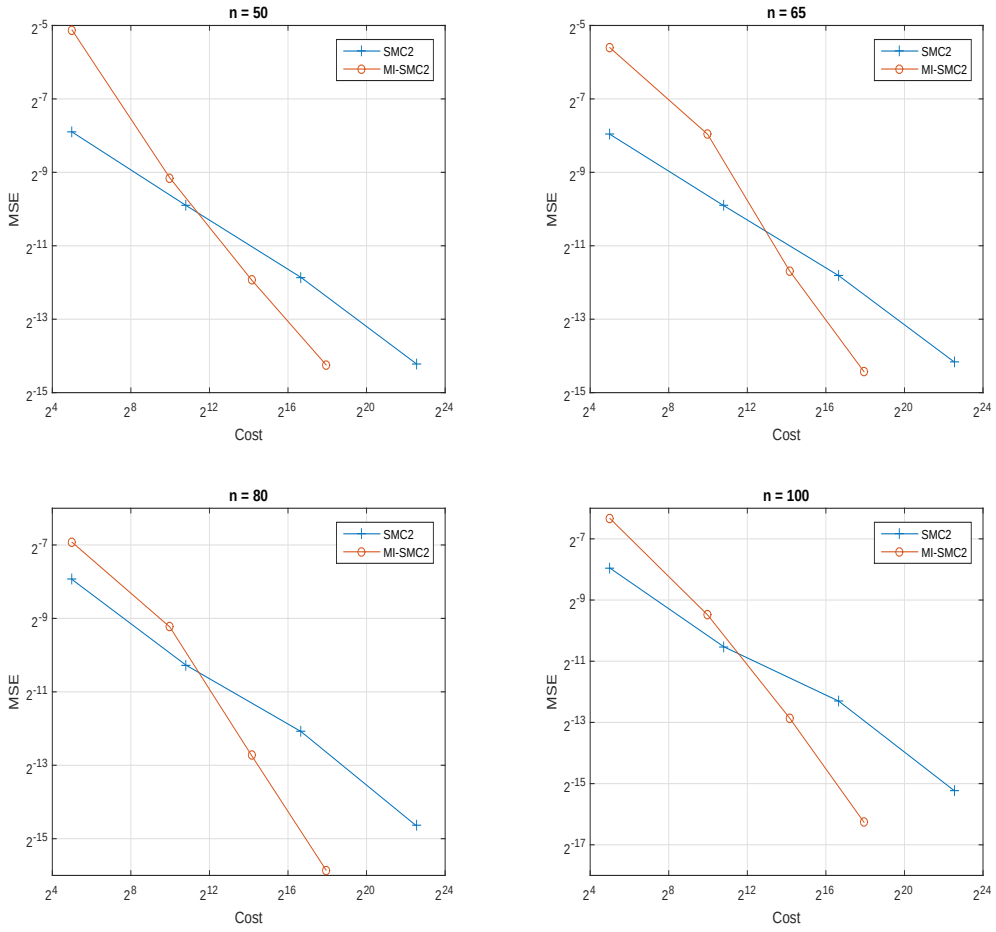

Figure 3: MSE v.s. Cost at different time index $n \in \{50, 65, 80, 100\}$ `fg:cost_error`

A Main Proofs

Let $(E, \mathcal{E})$ be a measurable space. The supremum norm is written as $\|f\| = \sup_{u \in E} |f(u)|$. We will consider a non-negative operator $K : E \times \mathcal{E} \to \mathbb{R}_+$, finite measure μ on $(E, \mathcal{E})$ and a real-valued, measurable $f : E \to \mathbb{R}$ and define the operations:

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

We also write $\mu(f) = \int f(u)\mu(du)$.

Recall the definition of E_p in Section 4.3 and denote by $\mathcal{E}_p$ the associated $\sigma-$algebra. Let $p \geq 1$ and denote by $M_p : \mathsf{E}_{p-1} \times \mathcal{E}_p \to [0, 1]$ the Markov kernel which is a compositon of

1. A marginal PMCMC kernel $\bar{M}_p : \mathsf{E}_{p-1} \times \mathcal{E}_{p-1} \to [0, 1]$, as in Algorithm 2,
2. Followed by the sampling of, for $i \in \{1, \ldots, N_\alpha\}$ $X_p^{i,j}(1 : k_\alpha), a_{p-1}^{i,j}$, $j \in \{1, \ldots, N\}$ in Algorithm 4, the iterate step.

Denote by $\tilde{\eta}_0$ as the initial probability measure (on $(\mathsf{E}_0, \mathcal{E}_0)$) of θ^i and $X_0^{i,j}(1 : k_\alpha)$ $j \in \{1, \ldots, N\}$ in Algorithm 4, the initialization step. Define the probability measure on $(\mathsf{E}_p \times \{1, \ldots, N\}), \mathcal{E}_p \vee 2^{\{1,\ldots,N\}})$, $\eta_{p,\alpha}$:

$$\eta_{p,\alpha}(d(u_{p,\alpha}, s)) := \Big(\int_{\mathsf{E}_0 \times \cdots \times \mathsf{E}_{p-1}} [\prod_{q=0}^{p} G_p(u_{p,\alpha})] \tilde{\eta}_0(du_{0,\alpha}) [\prod_{q=1}^{p-1} M_q(u_{q-1,\alpha}, du_{q,\alpha})] \bar{M}_p(u_{p-1,\alpha}, du_{p,\alpha}) \times$$

$$\mathbb{P}(s|u_{p,\alpha})ds\Big) \Big/ \Big(\int_{\mathsf{E}_0 \times \cdots \times \mathsf{E}_{p-1}} [\prod_{q=0}^{p} G_p(u_{p,\alpha})] \tilde{\eta}_0(du_{0,\alpha}) [\prod_{q=1}^{p-1} M_q(u_{q-1,\alpha}, du_{q,\alpha})] \Big)$$

where ds is counting measure and $\mathbb{P}(s|u_{p,\alpha})$ is as (16).

Note that one can easily show that (11) is equal to

$$\sum_{i=1}^{k'_\alpha} (-1)^{|\alpha(k_\alpha) - \alpha(2i)|} \left(\frac{\eta_{n,\alpha}(\varphi_{\alpha(2i)} H_{2i,n,\alpha,\theta})}{\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})} - \frac{\eta_{n,\alpha}(\varphi_{\alpha(2i-1)} H_{2i-1,n,\alpha,\theta})}{\eta_{n,\alpha}(H_{2i-1,n,\alpha,\theta})} \right).$$

Recall $\tau_{i,\alpha} = (-1)^{|\alpha(k_\alpha) - \alpha(2i)|}$ and set, for each $\varphi, \alpha(i)$,

$$\zeta_{i,n,\varphi}(x_{0:n}(1 : k_\alpha), s, \theta) = \varphi_{\alpha(i)}(x_{0:n}^s(i), \theta) H_{i,n,\alpha,\theta}(x_{0:n}(1 : k_\alpha)).$$

Now set

$$\begin{aligned}
\psi^{N_\alpha}_{n,i,\alpha} &:= \frac{\eta^{N_\alpha}_{n,\alpha}(\zeta_{2i-1,n,\varphi})}{\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})\eta^{N_\alpha}_{n,\alpha}(H_{2i-1,n,\alpha,\theta})} \\
\psi_{n,i,\alpha} &:= \frac{\eta_{n,\alpha}(\zeta_{2i-1,n,\varphi})}{\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})\eta_{n,\alpha}(H_{2i-1,n,\alpha,\theta})} \\
\overline{\psi}^{N_\alpha}_{n,i,\alpha} &:= \psi^{N_\alpha}_{n,i,\alpha} - \psi_{n,i,\alpha}.
\end{aligned}$$

In addition:

$$\begin{aligned}
\Xi^{N_\alpha}_{n,\alpha,1} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\left[\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\right][\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}](\zeta_{2i,n,\varphi} - \zeta_{2i-1,n,\varphi}) \\
\Xi^{N_\alpha}_{n,\alpha,2} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\overline{\psi}^{N_\alpha}_{n,i,\alpha}[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}](H_{2i,n,\alpha,\theta} - H_{2i-1,n,\alpha,\theta}) \\
\Xi^{N_\alpha}_{n,\alpha,3} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}](\zeta_{2i,n,\varphi} - \zeta_{2i-1,n,\varphi}) \\
\Xi^{N_\alpha}_{n,\alpha,4} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\psi_{n,i,\alpha}[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}](H_{2i,n,\alpha,\theta} - H_{2i-1,n,\alpha,\theta}) \\
\Xi^{N_\alpha}_{n,\alpha,5} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\left[\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-2}\right] \times \\
&\quad \eta_{n,\alpha}(\zeta_{2i,n,\varphi} - \zeta_{2i-1,n,\varphi})[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}](H_{2i,n,\alpha,\theta}) \\
\Xi^{N_\alpha}_{n,\alpha,6} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\eta_{n,\alpha}(H_{2i,n,\alpha,\theta} - H_{2i-1,n,\alpha,\theta})\overline{\psi}^{N_\alpha}_{n,i,\alpha} \\
\Xi^{N_\alpha}_{n,\alpha,7} &:= \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-2}\eta_{n,\alpha}(\zeta_{2i,n,\varphi} - \zeta_{2i-1,n,\varphi})[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}](H_{2i,n,\alpha,\theta}).
\end{aligned}$$

lem:1 **Lemma A.1.** *Assume (A1). Then for every $n \geq 0$, $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$ bounded and $\alpha \in \mathcal{I}$ we have that:*

$$\Delta\mathbb{E}^{N_\alpha}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)] - \Delta\mathbb{E}_{\pi_{n,\alpha}}[\varphi_\alpha(X_{0:n},\theta)] = \sum_{j=1}^{7}(-1)^{j+1}\Xi^{N_\alpha}_{n,\alpha,j}.$$

Proof. Follows by standard algebra. (A1) is only used to ensure the existence of all the associated quantities. □

prop:1 **Proposition A.1.** *Assume (A1-2). Then for every $n \geq 0$, $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$ bounded and $\alpha \in \mathcal{I}$, there exist a $C(\alpha(1:k_\alpha))$, with $\lim_{\min_{1\leq i\leq d}\alpha_i \to +\infty} C(\alpha(1:k_\alpha)) = 0$, such that*

for $j \in \{1, \ldots, 7\}$, $N_\alpha \geq 1$*:*

$$\max\{|\mathbb{E}[\Xi^{N_\alpha}_{n,\alpha,j}]|, \mathbb{E}[(\Xi^{N_\alpha}_{n,\alpha,j})^2]\} \leq \frac{C(\alpha(1:k_\alpha))}{N_\alpha}.$$

Proof. We give the proofs in the case $j = 1$ or $j = 3$. All other cases are essentially the same and omitted. Throughout the proof $C(\alpha(1:k_\alpha))$ is a constant that depends on $n \geq 0$, $\varphi : \mathbb{N}_0^d \times \mathsf{X}^{n+1} \times \Theta \to \mathbb{R}$, with $\lim_{\min_{1\leq i\leq d} \alpha_i \to +\infty} C(\alpha(1:k_\alpha)) = 0$. The exact value of $C(\alpha(1:k_\alpha))$ may change from line to line, but the latter property holds.

Set

$$\kappa_{n,\alpha,1}(x_{0:n}(1:k_\alpha), s, \theta) = \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha} \Big[\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\Big] \times$$

$$(\zeta_{2i,n,\varphi}(x_{0:n}(1:k_\alpha), s, \theta) - \zeta_{2i-1,n,\varphi}(x_{0:n}(1:k_\alpha), s, \theta)).$$

Then

$$\mathbb{E}[(\Xi^{N_\alpha}_{n,\alpha,1})^2] = \mathbb{E}\Big[[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}]\Big(\frac{\kappa_{n,\alpha,1}}{\|\kappa_{n,\alpha,1}\|}\Big)^2 \|\kappa_{n,\alpha,1}\|^2\Big].$$

Clearly, applying (A2) to the term $\|\kappa_{n,\alpha,1}\|$ and using that

$$[\eta^{N_\alpha}_{n,\alpha} - \eta_{n,\alpha}]\Big(\frac{\kappa_{n,\alpha,1}}{\|\kappa_{n,\alpha,1}\|}\Big) \leq 2 \tag{20}$$

it follows that

$$\mathbb{E}[(\Xi^{N_\alpha}_{n,\alpha,1})^2] \leq C(\alpha(1:k_\alpha))\mathbb{E}\Big[\Big(\sum_{i=1}^{k'_\alpha} \Big|\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\Big|^2\Big)^2\Big]$$

application of the Minkowski inequality yields

$$\mathbb{E}[(\Xi^{N_\alpha}_{n,\alpha,1})^2] \leq C(\alpha(1:k_\alpha))\Big(\sum_{i=1}^{k'_\alpha} \mathbb{E}\Big[\Big(\Big[\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\Big]\Big)^4\Big]^{1/2}\Big)^2.$$

Note that the summand

$$\mathbb{E}\Big[\Big(\Big[\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\Big]\Big)^4\Big]^{1/2} = \mathbb{E}\Big[\Big(\Big[\frac{\eta_{n,\alpha}(H_{2i,n,\alpha,\theta}) - \eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})}{\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})}\Big]\Big)^4\Big]^{1/2}$$

then applying (A1)

$$\mathbb{E}\Big[\Big(\Big[\eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1} - \eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}\Big]\Big)^4\Big]^{1/2} \leq C\mathbb{E}\Big[\Big(\eta_{n,\alpha}(H_{2i,n,\alpha,\theta}) - \eta^{N_\alpha}_{n,\alpha}(H_{2i,n,\alpha,\theta})\Big)^4\Big]^{1/2}$$

where $C < +\infty$ is a constant that does not depend upon α. Then applying [8, Proposition 2.9] to the term on the r.h.s. of the above equation, yields that

$$\mathbb{E}\Big[\|\kappa_{n,\alpha,1}\|^2\Big] \leq \frac{C(\alpha(1:k_\alpha))}{N_\alpha^2} \tag{21}$$

and hence allows one to derive the upper-bound

$$\mathbb{E}[(\Xi_{n,\alpha,1}^{N_\alpha})^2] \leq \frac{C(\alpha(1:k_\alpha))}{N_\alpha^2}.$$

For the bias, we have

$$|\mathbb{E}[(\Xi_{n,\alpha,1}^{N_\alpha})]| \leq \mathbb{E}\Big[\Big|[\eta_{n,\alpha}^{N_\alpha} - \eta_{n,\alpha}]\Big(\frac{\kappa_{n,\alpha,1}}{\|\kappa_{n,\alpha,1}\|}\Big)\Big|\|\kappa_{n,\alpha,1}\|\Big]$$

it follows by using (20)

$$|\mathbb{E}[(\Xi_{n,\alpha,1}^{N_\alpha})]| \leq 2\mathbb{E}[\|\kappa_{n,\alpha,1}\|]$$

then using Jensen's inequality and (21), we can conclude that

$$|\mathbb{E}[(\Xi_{n,\alpha,1}^{N_\alpha})]| \leq \frac{C(\alpha(1:k_\alpha))}{N_\alpha}.$$

Set

$$\kappa_{n,\alpha,3}(x_{0:n}(1:k_\alpha),s,\theta) = \sum_{i=1}^{k'_\alpha} \tau_{i,\alpha}\eta_{n,\alpha}(H_{2i,n,\alpha,\theta})^{-1}(\zeta_{2i,n,\varphi}(x_{0:n}(1:k_\alpha),s,\theta) -$$

$$\zeta_{2i-1,n,\varphi}(x_{0:n}(1:k_\alpha),s,\theta)).$$

Then

$$\mathbb{E}[(\Xi_{n,\alpha,3}^{N_\alpha})^2] = \mathbb{E}[[\eta_{n,\alpha}^{N_\alpha} - \eta_{n,\alpha}](\kappa_{n,\alpha,3})^2].$$

Applying [8, Proposition 2.9] to the term on the r.h.s. yields

$$\mathbb{E}[(\Xi_{n,\alpha,3}^{N_\alpha})^2] \leq \frac{\|\kappa_{n,\alpha,3}\|^2}{N_\alpha}.$$

Application of (A2) gives

$$\mathbb{E}[(\Xi_{n,\alpha,3}^{N_\alpha})^2] \leq \frac{C(\alpha(1:k_\alpha))}{N_\alpha}.$$

For $|\mathbb{E}[(\Xi_{n,\alpha,3}^{N_\alpha})]|$ using a similar decomposition to [2, eq. (A.2)] one can show that

$$|\mathbb{E}[(\Xi_{n,\alpha,3}^{N_\alpha})]| \leq \frac{C(\alpha(1:k_\alpha))}{N_\alpha}.$$

the proof is omitted as it is standard. □

References

andrieu [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.

beskos [2] Beskos, A., Jasra, A., Law, K. J. H,, Tempone, R., & Zhou, Y. (2017). Multilevel Sequential Monte Carlo samplers. *Stoch. Proc. Appl.*, **127**, 1417-1440.

Cappe_2005 [3] Cappé, O., Ryden, T, & Moulines, É. (2005). *Inference in Hidden Markov Models*. Springer: New York

nov2016multilevel [4] Chernov, A., Hoel, H., Law, K.J.H.,Nobile, F. and Tempone, R. (2017). Multilevel ensemble Kalman filtering for spatio-temporal processes, arXiv preprint arXiv:1710.07282v1.

chopin2 [5] 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

pin2002sequential [6] Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, **89**, 539–552.

delm:04 [7] Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer: New York.

delm:00 [8] Del Moral, P. & Miclo, L. (2000). Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering. In: Azéma J., Ledoux M., Émery M., Yor M. (eds) Séminaire de Probabilités XXXIV. Lecture Notes in Mathematics, **1729**. Springer:Berlin.

franks [9] Franks, J., Jasra, A., Law, K. J. H. & Vihola, M. (2018). Unbiased inference for discretely observed hidden Markov model diffusions. arXiv:1807.10259.

giles [10] Giles, M. B. (2008). Multilevel Monte Carlo path simulation. *Op. Res.*, **56**, 607-617.

giles1 [11] Giles, M. B. (2015) Multilevel Monte Carlo methods. *Acta Numerica* **24**, 259-328.

mimc [12] Haji-Ali, A. L., Nobile, F. & Tempone, R. (2016). Multi-Index Monte Carlo: When sparsity meets sampling. *Numerische Mathematik*, **132**, 767–806.

hein [13] Heinrich, S. (2001). Multilevel Monte Carlo methods. In *Large-Scale Scientific Computing*, (eds. S. Margenov, J. Wasniewski & P. Yalamov), Springer: Berlin.

jklz [14] 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.

jklz1 [15] Jasra, A., Kamatani, K., Law, K. J. H. & Zhou, Y. (2018). Bayesian static parameter estimation via Multilevel Markov chain Monte Carlo. *SIAM J. Sci. Comp.* **40**, A887-A902.

ml_rev [16] Jasra, A., Law, K. J. H., & Suciu, C. (2017). Advanced Multilevel Monte Carlo Methods. arXiv preprint.

spde_disc [17] Jentzen, A., and Kloeden, P. (2009). Overcoming the order barrier in the numerical approximation of stochastic partial differential equations with additive space-time noise, *Proc. Roy. Soc. A*, **465**, pp. 649–667.

kantas1 [18] 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.

kantas [19] 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.

law2015data [20] Law, K., Stuart, A. and Zygalakis, K. (2015). *Data Assimilation*. Springer-Verlag, New York.

snyder [21] Snyder, C., Bengtsson, T., Bickel, P., & Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Month. Weather Rev.*, **136**, 4629–4640.