FOUNDATIONS OF
COMPUTATIONAL
MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Multi-index Sequential Monte Carlo Ratio Estimators for Bayesian Inverse problems

Ajay Jasra[2] · Kody J. H. Law[1] · Neil Walton[1] · Shangda Yang[1]

Abstract

We consider the problem of estimating expectations with respect to a target distribution with an unknown normalising constant, and where even the un-normalised target needs to be approximated at finite resolution. This setting is ubiquitous across science and engineering applications, for example in the context of Bayesian inference where a physics-based model governed by an intractable partial differential equation (PDE) appears in the likelihood. A multi-index sequential Monte Carlo (MISMC) method is used to construct ratio estimators which provably enjoy the complexity improvements of multi-index Monte Carlo (MIMC) as well as the efficiency of sequential Monte Carlo (SMC) for inference. In particular, the proposed method provably achieves the canonical complexity of MSE^{-1}, while single-level methods require $MSE^{-\xi}$ for $\xi > 1$. This is illustrated on examples of Bayesian inverse problems with an elliptic PDE forward model in 1 and 2 spatial dimensions, where $\xi = 5/4$ and $\xi = 3/2$, respectively. It is also illustrated on more challenging log-Gaussian process models, where single-level complexity is approximately $\xi = 9/4$ and multilevel Monte Carlo (or MIMC with an inappropriate index set) gives $\xi = 5/4 + \omega$, for any $\omega > 0$, whereas our method is again canonical. We also provide novel theoretical verification

✉ Kody J. H. Law
kody.law@manchester.ac.uk

Ajay Jasra
ajay.jasra@kaust.edu.sa

Neil Walton
neil.walton@manchester.ac.uk

Shangda Yang
shangda.yang@manchester.ac.uk

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

[2] Applied Mathematics and Computational Science Program; Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

Springer

of the product-form convergence results which MIMC requires for Gaussian processes built in spaces of mixed regularity defined in the spectral domain, which facilitates acceleration with fast Fourier transform methods via a cumulant embedding strategy, and may be of independent interest in the context of spatial statistics and machine learning.

Keywords Bayesian inverse problems · Sequential Monte Carlo · Multi-index Monte Carlo

Mathematics subject classification 65C40 · 60J22 · 65C05 · 62F15 · 62M30

1 Introduction

There has been an explosion of work over the past decade involving the enormously successful multilevel Monte Carlo (MLMC) method [27] for estimating expectations with respect to distributions which need to be approximated. The canonical example is the problem for forward uncertainty quantification (UQ), where a single realisation of the random variable of interest requires the solution to a stochastic, ordinary, or partial differential equation (SDE, ODE or PDE) [7, 53, 58]. The MLMC framework formulates this problem in terms of a sum of increments corresponding to approximations at successive resolutions, or *levels*. Under a suitable coupling of the increments, which is typically fairly trivial in the forward context, the variance of the increments decays as the resolution and cost increase, and so progressively fewer samples are required to control the variance at higher levels.

In the context of Bayesian inference, one typically requires expectations with respect to target distributions for which the normalising constant is unknown. As an example, let π denote a probability density on $\mathsf{X} \times \mathsf{Y}$. Assume we know how to evaluate $\pi(x, y) = \pi(y|x)\pi_0(x)$ and $\pi_0(x)$ but not $\pi(y) = \int_{\mathsf{X}} \pi(y|x)\pi_0(x)dx$. Now consider the case where one observes $y \in \mathsf{Y}$ and would like to infer the *posterior distribution* $\pi(x|y)$, given by

$$\pi(x|y) = \frac{\pi(y|x)\pi_0(x)}{\pi(y)} \, . \tag{1}$$

This is referred to as the Bayesian framework, and $\pi(y|x)$ and $\pi_0(x)$ are referred to as the likelihood and the prior, respectively [55]. Note that once the goal of (1) is established, then a method should be capable of efficiently approximating integrals of the form:

$$\frac{1}{Z} \int_{\mathsf{X}} \varphi(x) f(x) dx \, ,$$

where $f(x) \propto \pi(y|x)\pi_0(x)$ and $Z = \int_{\mathsf{X}} f(x)dx$ (i.e. $f(x)$ itself only needs to be proportional to the joint density). Methods which have been designed for exactly this purpose include Markov chain Monte Carlo (MCMC) [26], importance sampling [55],

and combinations thereof such as sequential Monte Carlo (SMC) samplers [15, 21]. The latter methods are particularly powerful, handling elegantly some of the most challenging issues that arise in this context, such as small variance, strong dependence between variables, and multimodality.

Over the past decade the excitement about MLMC has intersected with the Bayesian computation community, in particular relating to the context of Bayesian inverse problems [59], where an intractable PDE often appears inside the likelihood of the posterior distribution of interest. For instance, we will later consider the case where the likelihood takes the from:

$$\pi(y|x) \propto e^{-\frac{1}{2}\|y-\mathcal{G}(x)\|^2},$$

where y is an observed set of real-valued outputs and $\mathcal{G}(x)$ is a solution to the outputs from the intractable PDE for a given set of input parameters x. This context appears to be much more subtle, due to the complications of combining these technologies. Early work is related to MCMC [22, 34] and SMC samplers [5, 6, 48]. More recently, the methodology has also been applied to the context of partially observed diffusions [35, 39], for parameter inference [37], online state inference [2, 13, 28, 30, 39, 42], or both [19]. A notable recent body of work relates to continuous-time observations in this context [4, 45, 56]. Another notable trend is the application of randomised MLMC methods [11, 33, 40, 41, 44] in this context. Typically, these methods require unbiased estimators of increments, which is particularly challenging in the inference context. The first work to use randomised MLMC in the context of inference was [11], and unbiased increment estimators were available in the context of that work. Other more recent instalments have utilised double randomisation strategies in order to remove the bias of increment estimators [33, 40, 41].

The benefits of MLMC are somewhat hampered by the dimension of the underlying problem. This is an important issue, particularly in the context of a PDE or a SPDE. For example, the error associated with a finite element method (FEM) approximation of a PDE typically depends upon the mesh diameter, h, while the number of degrees of freedom typically scales like h^{-d}, where d is the dimension of the associated PDE. The multi-index Monte Carlo (MIMC) method was introduced to gracefully handle the dimension dependence of this problem [29] following, in spirit, from the seminal work on sparse grids [10]. Instead of an estimator based on a sum of increments, the MIMC method constructs an estimator based on a sum over an index set of $d-$fold composition of increments. Under suitable regularity conditions, this approach is able to leverage convergence in each dimension independently and thereby mitigate the curse of dimensionality.

The MIMC method has very recently been applied to the inference context [19, 38, 43]; however, the estimates required for increments of increments have proven challenging from a theoretical perspective, and this has severely limited progress thus far. In particular, an MIMC method for inference with provable convergence guarantees does not currently exist, except a ratio estimator using simple importance sampling, as considered for MLMC and QMC in this work [57]. Such estimators are not expected to be practical for complex target distributions due to a large constant associated with importance sampling [1, 12].

The current work breaks down this theoretical barrier and unveils the MISMC sampler ratio estimator for posterior inference. By employing a ratio estimator, we introduce a theoretically tractable method which provably achieves the benefits of both SMC samplers for inference and MIMC for multi-dimensional discretisation. In particular, rather than dealing with self-normalised increments of increments, as previous methods have done, the innovation is to construct instead a ratio of MIMC estimators of an un-normalised integral and its normalising constant, both of which can be unbiasedly estimated with SMC sampler. This seemingly minor difference of formulation substantially simplifies the analysis and enables us to establish a theory for the convergence of an MIMC method for inference problems—a theory which until this point had been elusive.

This article is structured as follows. In Sect. 2, we provide a class of motivating problems for the methodology that is developed. In Sect. 3, we provide a review of the relevant computational methodology that is used in our approach. In Sect. 4, we present our method and theoretical results. In Sect. 6, we present numerical results. Finally, in the appendix several technical results are given, necessary for the theory that is presented in Sect. 4.

2 Motivating Problems

We consider the setting of Bayesian inference for an elliptic partial differential equation and for the log-Gaussian Cox model, where we must also perform numerical estimation.

2.1 Elliptic Partial Differential Equation

We consider the following elliptic PDE. Consider a convex domain $\Omega \subset \mathbb{R}^D$ with boundary $\partial \Omega \in C^0$, a function (force vector field) $\mathsf{f} : \Omega \to \mathbb{R}$ and a function (permeability) $a(x) : \Omega \to \mathbb{R}_+$ which is parameterized by $x \in \mathsf{X}$. For each $x \in \mathsf{X}$, we define the (pressure field) $u(x) : \Omega \to \mathbb{R}$ as the solution to the following PDE on Ω

$$-\nabla \cdot (a(x)\nabla u(x)) = \mathsf{f}, \quad \text{on } \Omega , \tag{2}$$

$$u(x) = 0, \quad \text{on } \partial \Omega . \tag{3}$$

In the above PDE, we assume the force vector field is known, e.g. $\mathsf{f} = 1$. However, we assume the permeability $a(x)$ depends upon a parameter x which is a random variable, specifically $x \sim \pi_0$. The dependence of a on x induces a dependence of the solution u on x. Hence, the solution itself, $u(x)(z)$, is a random variable for each $z \in \Omega$.

For concreteness, assume that $D = 2$ and $\Omega = [0, 1]^2$. Assume a uniform prior,

$$x \sim U(-1, 1)^d =: \pi_0. \tag{4}$$

For $x \sim \pi_0$, and $z \in \Omega$, the permeability will take the form

$$a(x)(z) = a_0 + \sum_{i=1}^{d} x_i \psi_i(z), \tag{5}$$

where ψ_i are smooth functions with $\|\psi_i\|_{L^\infty(\Omega)} \le 1$ for $i = 1, \ldots, d$, and $a_0 > \sum_{i=1}^{d} x_i$. In particular, for simplicity and concreteness, let $d = 2$ and

$$a(x)(z) = 3 + x_1 \cos(3\pi z_1) \sin(3\pi z_2) + x_2 \cos(\pi z_1) \sin(\pi z_2).$$

2.1.1 Finite Element Approximation and Error Estimates

Consider the 1D piecewise linear nodal basis functions ϕ_j^K defined as follows, for mesh $\{z_i^K = i/(K+1)\}_{i=0}^{K+1}$, and for $j = 1, \ldots, K$,

$$\phi_j^K(z) = \begin{cases} \frac{z - z_{j-1}^K}{z_j^K - z_{j-1}^K}, & z \in [z_{j-1}^K, z_j^K] \\ 1 - \frac{z - z_j^K}{z_{j+1}^K - z_j^K}, & z \in [z_j^K, z_{j+1}^K] \\ 0, & \text{else.} \end{cases}$$

Now, for $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}^2$, consider the tensor product grid over $\Omega = [0,1]^2$ formed by

$$\{(z_{i_1}^{K_{1,\alpha}}, z_{i_2}^{K_{2,\alpha}})\}_{i_1=0, i_2=0}^{K_{1,\alpha}+1, K_{2,\alpha}+1},$$

where $K_{1,\alpha} = 2^{\alpha_1}$ and $K_{2,\alpha} = 2^{\alpha_2}$ (and the mesh-width in each direction is bounded by $2^{-\alpha_k}, k = 1, 2$). Let $i = i_1 + K_{1,\alpha} i_2$ for $i_1 = 1, \ldots, K_{1,\alpha}$ and $i_2 = 1, \ldots, K_{2,\alpha}$ and $K_\alpha = K_{1,\alpha} K_{2,\alpha}$, and let $\phi_i^\alpha(z) = \phi_{i_1,i_2}^\alpha(z_1, z_2) = \phi_{i_1}^{K_{1,\alpha}}(z_1) \phi_{i_2}^{K_{2,\alpha}}(z_2)$ be piecewise bilinear functions. The weak solution of the PDE (2) and (3) will be approximated by $u_\alpha(x) = \sum_{i=1}^{K_\alpha} u_\alpha^i(x) \phi_i^\alpha \in V$. Given x, the values of $u_\alpha^i(x)$ are defined by substituting the expansion into (2) and taking inner product with ϕ_j^α for $j = 1, \ldots, K_\alpha$. In particular, observe that we have

$$\left\langle -\nabla \cdot \left(a(x) \nabla \sum_{i=1}^{K_\alpha} u_\alpha^i(x) \phi_i^\alpha \right), \phi_j^\alpha \right\rangle = \langle f, \phi_j^\alpha \rangle \quad j = 1, \ldots, K_\alpha$$

Using integration by parts and observing that $\phi_i^\alpha|_{\partial\Omega} \equiv 0$, then

$$\sum_{i=1}^{K_\alpha} \langle a(x) u_\alpha^i(x) \nabla \phi_i^\alpha, \nabla \phi_j^\alpha \rangle = \langle f, \phi_j^\alpha \rangle \quad j = 1, \ldots, K_\alpha$$

We can represent the solution as a vector $\mathbf{u}_\alpha(x) = [u_\alpha^i(x) : i = 1, \ldots, K_\alpha]$, and define $\mathbf{f}_{\alpha,j} = \langle f, \phi_j^\alpha \rangle$ and

$$\mathbf{A}_{\alpha,ij}(x) := \int_{z_{1,j_1-1}}^{z_{1,j_1+1}} \int_{z_{2,j_2-1}}^{z_{2,j_2+1}} a(x)(z)\nabla\phi_i^\alpha(z) \cdot \nabla\phi_j^\alpha(z)dz,$$

where we introduce the notation $j := j_1 + j_2 K_{1,\alpha}$ (for $j_1 = 1, \ldots, K_{1,\alpha}$ and $j_2 = 1, \ldots, K_{2,\alpha}$).

Observe that if $i = i_1 + i_2 K_{1,\alpha}$, then the integral is zero for all i such that $i_k < j_k - 1$ or $i_k > j_k + 1$, for $k \in \{1, 2\}$. So the above matrix $\mathbf{A}_\alpha(x)$ is sparse, and it is straightforward to verify that it is symmetric positive definite.

The approximate weak solution to Eqs. (2), (3) is given by the system

$$\mathbf{A}_\alpha(x)\mathbf{u}_\alpha(x) = \mathbf{f}_\alpha.$$

Due to the sparsity of $\mathbf{A}_\alpha(x)$, for $D \leq 2$ the solution can be obtained for a cost of roughly $\mathcal{O}(K_\alpha)$ using an iterative solver based on Krylov subspaces, such as conjugate gradients [52]. For $D \geq 3$ it may no longer be possible to achieve a linear cost—see, e.g. [29]. See the references [9, 16] for further description and much more.

The weak solution u of (2) and (3) is said to be $W^{2,2}$ regular if there exists a $C > 0$, such that

$$\|\nabla^2 u\| \leq C\|\mathbf{f}\|$$

for every $\mathbf{f} \in L^2(\Omega)$, where $\|\cdot\|$ denotes the $L^2(\Omega)$ norm. For the purposes of the present work, it suffices to observe the following proposition [8, 24].

Proposition 2.1 *For $a(x)$ given by (5) and uniformly over $x \in [-1, 1]^d$, $\mathbf{f} \in L^2$ and Ω convex, the weak solution of (2) and (3) is $W^{2,2}$ regular, and there exists a $C > 0$ such that*

$$\|\nabla(u_\alpha(x) - u(x))\| \leq C2^{-\min\{\alpha_1,\alpha_2\}}.$$

Furthermore,

$$\|u_\alpha(x) - u(x)\| \leq C2^{-2\min\{\alpha_1,\alpha_2\}}.$$

2.1.2 A Bayesian Inverse Problem

In the PDE (2) and (3), the parameter x is unknown. Here, we infer estimates about the true value x from noisy observations of the solution to the PDE, $u(x)$. A further confounding factor is that the closed form solution to $u(x)$ is, in general, not known in closed form and instead we must numerically approximate $u(x)$ with $u_\alpha(x)$ as described above.

Now observations y will be introduced and we will consider the inverse problem, given by

$$\pi(dx) := \pi(dx|y) \propto L(x)\pi_0(dx) \tag{6}$$

where $L(x) \propto \pi(y|x)$ and the dependence upon y is suppressed in the notation. We will use the notations $d\pi(x) = \pi(dx) = \pi(x)dx$ to mean the same thing, i.e. probability under π of an infinitesimal volume element dx (Lebesgue measure by default) centred at x, and the argument may be omitted from $d\pi$ where the meaning is understood.

Define the following vector-valued function

$$\mathcal{G}(u(x)) = [v_1(u(x)), \ldots, v_n(u(x))]^\top, \tag{7}$$

where $v_i \in L^2$ and $v_i(u(x)) = \int v_i(z)u(x)(z)dz$ for $i = 1, \ldots, n$, for some $n \geq 1$. It is assumed that the data take the form

$$y = \mathcal{G}(u(x)) + v, \qquad v \sim N(0, \Xi), \qquad v \perp x, {}^{1} \tag{8}$$

and we define

$$L(x) := \exp\left(-\frac{1}{2}|y - \mathcal{G}(u(x))|_\Xi^2\right).$$

Here y is suppressed from the notation. Also we apply the convention that $|w|_\Xi := (w^\top \Xi^{-1} w)^{1/2}$.

In particular, $u(x)$ is the (weak) solution map of (2) and (3), for given input x. Denote its weak approximation at resolution multi-index α by $u_\alpha(x)$. The approximated likelihood is given by

$$L_\alpha(x) := \exp(-\frac{1}{2}|y - \mathcal{G}(u_\alpha(x))|_\Xi^2),$$

and the associated target is

$$\pi_\alpha(dx) \propto L_\alpha(x)\pi_0(dx). \tag{9}$$

The following proposition summarises the key result.

Proposition 2.2 *In the present context, there is a $C > 0$ such that $u, u_\alpha \leq C$, hence a $c > 0$ such that $L, L_\alpha \geq c > 0$, and so (6) and (9) are well defined. Furthermore, following Proposition 2.1 and the continuity of L as a function of u, the following rate estimate holds uniformly in x*

$$|L_\alpha(x) - L(x)| \leq C2^{-2\min\{\alpha_1,\alpha_2\}}.$$

For the concrete example of $D = 2$, let the observations be given by $v_i(u) := u(z_i)$, for $i = 1, \ldots, 4$, where $z_i \in \{(0.25, 0.25), (0.25, 0.75), (0.75, 0.75), (0.75, 0.25)\}$, and let $\Xi = \xi^2 I$. This example has been considered in the context of an MLSMC sampler method in [6]. It is noted that this example extends the theory described, since $v_i \notin L^2$.

---

[1] Here we use $\perp$ to denote pairwise independence of random variables.

2.2 Log-Gaussian Process Models

Another model problem which will be considered is the log-Gaussian process (LGP), and the related log-Gaussian Cox (LGC) process, which are commonly used in spatial statistics. In this example, the dimension of the state space grows with level.

Specifically, we aim to model a data set comprised of the location of $n = 126$ Scots pine saplings in a natural forest in Finland [47], denoted $z_1, \dots, z_n \in [0, 1]^2$. The LGC version of our model is based on the one presented in [31]. The process of interest is defined as $\Lambda = \exp(x)$ where x is a Gaussian process, a priori distributed in terms of a KL-expansion as follows, for $z \in [0, 2]^2$,

$$x(z) = \theta_1 + \sum_{k \in \mathbb{Z} \times \mathbb{Z}_+ \cup \mathbb{Z}_+ \times 0} \rho_k(\theta)(\xi_k \phi_k(z) + \xi_k^* \phi_{-k}(z)), \quad \xi_k \sim \mathcal{CN}(0, 1) \text{ i.i.d.},$$

$$(10)$$

where $\mathcal{CN}(0, 1)$ denotes a standard complex Normal distribution, ξ_k^* is the complex conjugate of ξ_k, and $\phi_k(z) \propto \exp[\pi i z \cdot k]$ are Fourier series basis functions (with $i = \sqrt{-1}$), and

$$\rho_k^2(\theta) = \theta_2 / ((\theta_3 + k_1^2)(\theta_3 + k_2^2))^{\frac{\beta+1}{2}}. \tag{11}$$

The coefficient β controls the smoothness of the Gaussian process. The parameters θ will be assumed known in the present work, but these can also be fit within a hierarchical modelling framework. The associated prior measure is denoted by μ_0. Following the formulation from [31], the likelihoods are defined by

$$\text{(LGC)} \quad \frac{d\pi}{d\pi_0}(x) \propto \exp\left[\sum_{j=1}^{n} x(z_j) - \int_{[0,1]^2} \exp(x(z))dz\right], \tag{12}$$

$$\text{(LGP)} \quad \frac{d\pi}{d\pi_0}(x) \propto \exp\left[\sum_{j=1}^{n} x(z_j) - n \log \int_{[0,1]^2} \exp(x(z))dz\right]. \tag{13}$$

See, e.g. [61] for a description of the LGP version, which is given second above. Note that only $z \in [0, 1]^2$ is required. The periodic prior measure is defined on $[0, 2]^2$ so that no boundary conditions are imposed on the sub-domain $[0, 1]^2$ and the fast Fourier transform (FFT) can be used for approximation, as described below.

The finite approximation is constructed as follows. First the KL expansion (10) is truncated

$$x_\alpha(z) = \theta_1 + \sum_{k \in \mathcal{A}_\alpha} \rho_k^2(\theta)(\xi_k \phi_k(z) + \xi_k^* \phi_{-k}(z)), \quad \xi_k \sim \mathcal{CN}(0, 1) \text{ i.i.d.}, \tag{14}$$

where $\mathcal{A}_\alpha := \{-2^{\alpha_1/2}, \dots, 2^{\alpha_1/2}\} \times \{1, \dots, 2^{\alpha_2/2}\} \cup \{1, \dots, 2^{\alpha_2/2}\} \times 0$. Note that $x_\alpha(z)$ can be approximated on a grid $\{0, 2^{-\alpha_1}, \dots 1 - 2^{-\alpha_1}\} \times \{0, 2^{-\alpha_2}, \dots 1 - 2^{-\alpha_2}\}$

using the FFT with a cost $\mathcal{O}((\alpha_1 + \alpha_2)2^{\alpha_1 + \alpha_2})$. Now $\hat{x}_\alpha(z)$ is defined as an interpolant (for example linear) over the grid output from FFT. The finite approximation of the likelihood is then defined by

$$\text{(LGC)} \qquad \frac{d\pi_\alpha}{d\pi_0}(x_\alpha) \propto \exp\left[\sum_{j=1}^n \hat{x}_\alpha(z_j) - Q(\exp(x_\alpha))\right], \qquad (15)$$

$$\text{(LGP)} \qquad \frac{d\pi_\alpha}{d\pi_0}(x_\alpha) \propto \exp\left[\sum_{j=1}^n \hat{x}_\alpha(z_j) - n\log Q(\exp(x_\alpha))\right], \qquad (16)$$

where Q denotes a quadrature rule, which may for example be given by $Q(\exp(x_\alpha)) = 2^{-(\alpha_1+\alpha_2)} \sum_{h \in \prod_{i=1}^2 \{0, 2^{-\alpha_i}, ..., 1-2^{-\alpha_i}\}} \exp(x_\alpha(h))$ or $Q(\exp(x_\alpha)) = \int \hat{x}_\alpha(z)dz$.

If one uses the prior with isotropic spectrum $\rho_k^2(\theta) = \theta_2/(\theta_3 + k_1^2 + k_2^2)^{\frac{3}{2}}$, then our target measure coincides with the standard prior of [31] in the limit as $\min_i \alpha_i \to \infty$. One can understand the connection in this context via the circulant embedding method based on FFT [46]. However, previous work has employed the (dense) kernel representation of the covariance function instead of diagonalising it with FFT. For our product-form spectrum, the regularity would be matched for $\beta = 1$, corresponding to a product of Ornstein–Uhlenbeck processes. Instead, we will choose $\beta = 1.6$ for convenience, which means that our prior is slightly smoother.

2.2.1 LGP and LGC Theoretical Results

First, we state a simple convergence result for Gaussian process of the form (10) with spectral decay corresponding to (11).

It will be useful to define the following operator A on the space of functions in $L^2(\Omega)$:

$$A = \sum_{k \in \mathbb{Z}^2} a_k \phi_k \otimes \phi_k, \qquad a_k = (1 + k_1^2)(1 + k_2^2), \qquad (17)$$

the mixed Sobolev-like norms

$$\|x\|_q := \|A^{q/2}x\|, \qquad (18)$$

where $A^{q/2} = \sum_{k \in \mathbb{Z}^2} a_k^{q/2} \phi_k \otimes \phi_k$ and we recall $\|\cdot\|$ is the $L^2(\Omega)$ norm, and the spaces

$$H_q^{\mathsf{m}} := \{x \in L^2(\Omega); \|x\|_q < \infty\}. \qquad (19)$$

We note that these spaces ensure *mixed regularity* (hence superscript m), rather than the typical regularity associated with standard Sobolev spaces. It is precisely this property which multi-index methods are designed to exploit. Figure 1 shows the contours of functions in $H_{1/2}^{\mathsf{m}}$ (red) and $W^{1/2,2}$ (blue), along with the regions associated with
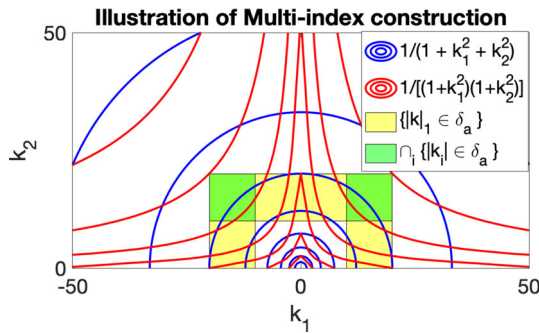
Fig. 1 A cartoon of variance contours associated with a function in $H^m_{1/2}$ (red) and a function in $W^{1/2,2}$ (blue). Letting $\delta_a = [2^{a/2}, 2^a]$, the spectral region associated with an increment of approximations on the index sets defined in (14), $\{|k|_1 \in \delta_a\} = \{k \in \mathcal{A}_\alpha\} \cap \{k \notin \mathcal{A}_{\alpha/2}\}$ (with $\alpha = (a, a)$), is depicted in yellow. Its intersection with the region associated with an *increment of increments*, $\cap_{i=1}^2 \{|k_i| \in \delta_a\}$, is depicted in green (Color figure online)

an increment (yellow) and increment of an increment at approximation level $\alpha = (a, a)$. From inspection, it is clear that increments of increments are of higher order in comparison with increments for functions in the mixed space, but not for functions in the standard space.

The following proposition is proven in "Appendix B".

Proposition 2.3 *Let $x \sim \pi_0$, where π_0 is a Gaussian process of the form (10) with spectral decay corresponding to (11), and let x_α correspond to truncation on the index set $\mathcal{A}_\alpha$ as in (14). Then $x \in H^m_q$ for all $q < \beta/2$, and for $r \in [0, q)$ there is a $C > 0$ such that*

$$\|x_\alpha - x\|^2_r \le C\|x\|^2_q 2^{-2(q-r)\min_i \alpha_i} .$$

For $\beta = 1$, (11) looks like a product of OU processes, with the regularity of Wiener measure [54]. Hence, following from Proposition 2.3, x is almost surely continuous for $\beta \ge 1$, and (12) and (13) are well defined. It is worth noting that (12) and (13) are not well defined for the standard prior, e.g. from [31], since the Sobolev embedding theorem (see, e.g. [59]) does not guarantee that the solution is almost surely continuous. However, non-infinitesimal representations of (12), i.e. for finite partitions of the domain, can still be computed as long as Λ is integrable.

The following proposition ensures our LGC and LGP posterior measures are well defined on function space and has a density with respect to the prior. It is proven in "Appendix B".

Proposition 2.4 *Given $x : \Omega \to \mathbb{R}$ is a Gaussian process, with probability measure denoted π_0, defined on compact finite dimensional space Ω, that is almost surely continuous and has a finite mean and covariance. If we define π by*

$$\text{(LGC)} \quad \frac{d\pi}{d\pi_0}(x) \propto \exp\left[\sum_{j=1}^n x(z_j) - \int_\Omega \exp(x(z))dz\right],$$

$$\text{(LGP)} \quad \frac{d\pi}{d\pi_0}(x) \propto \exp\left[\sum_{j=1}^{n} x(z_j) - n \log \int_{\Omega} \exp(x(z))dz\right],$$

for $n \in \mathbb{N}$ then $\pi(dx)$ is a well-defined probability measure, and can be represented in terms of its density with respect to π_0:

$$\pi(dx) = \frac{d\pi}{d\pi_0}\pi_0(dx).$$

The analogue to Proposition 2.2 takes the following form. The proposition is again reproduced, and proven, in "Appendix B".

Proposition 2.5 *For both LGP and LGC, there is a $C > 0$ such that for $x \sim \pi_0$ and $x_\alpha = P_{\mathcal{A}_\alpha} x$, where $P_{\mathcal{A}_\alpha}$ denotes the orthogonal projection onto the index set $\mathcal{A}_\alpha$ defined in (14), the following rate estimate holds for all $q < (\beta - 1)/2$*

$$\mathbb{E}|L_\alpha(x_\alpha) - L(x)|^2 \leq C2^{-2\min\{q,1\}\min\{\alpha_1,\alpha_2\}}.$$

3 Computational Methodology

3.1 Approximate Monte Carlo

For concreteness, in this subsection we will consider the case of the PDE example from Sect. 2.1. The case of Sect. 2.2 follows similarly. Let $\mathsf{X} := [-1, 1]^d$ be the domain of x.

3.1.1 Monte Carlo

The forward uncertainty quantification (UQ) problem is the following. Given a quantity of interest $\varphi_\alpha = \varphi \circ u_\alpha : \mathsf{X} \to \mathbb{R}$, compute its expectation

$$\mathbb{E}\varphi_\alpha(x) = \int_{\mathsf{X}} \varphi(u_\alpha(x))\pi_0(dx).$$

The typical strategy is to independently sample $x^i \sim \pi_0$, for $i = 1, \ldots, N$, and then approximate

$$\mathbb{E}\varphi_\alpha(x) \approx \frac{1}{N}\sum_{i=1}^{N} \varphi(u_\alpha(x^i)).$$

For example, we can let

$$\varphi_\alpha(x) = \|u_\alpha(x)\|^2 = \int_{\Omega} |u_\alpha(x)(z)|^2 dz \approx \sum_{i=1}^{K_{\alpha_1}}\sum_{j=1}^{K_{\alpha_2}} u_\alpha^i(x)u_\alpha^j(x) \int_{\Omega} \phi_i^\alpha(z)\phi_j^\alpha(z)dz,$$

where the latter can be written as $\mathbf{u}_\alpha^T \mathbf{K}_\alpha \mathbf{u}_\alpha$, where $\mathbf{K}_{\alpha,ij} := \langle \phi_i^\alpha, \phi_j^\alpha \rangle$.

3.1.2 Multi-index Monte Carlo

With MIMC [29], we apply the approximation

$$\mathbb{E}\varphi(x) \approx \sum_{\alpha \in \mathcal{I}} \mathbb{E}[\Delta\varphi_\alpha(x)] \tag{20}$$

where the difference of differences operator is defined as $\Delta\varphi_\alpha := \Delta_D \circ \cdots \circ \Delta_1 \varphi_\alpha$ with $\Delta_i \varphi_\alpha := \varphi_\alpha - \varphi_{\alpha - e_i}$, e_i is the ith standard basis vector in $\mathbb{R}^D$, and $\varphi_\alpha \equiv 0$ if $\alpha_i < 0$ for any i. The task is then to approximate the expectation of the increment of increments for each $\alpha \in \mathcal{I} \subset \mathbb{Z}^D$. For example, for $D = 2$, one must approximate

$$\mathbb{E}[\Delta\varphi_\alpha(x)] = \int_{[-1,1]^2} \Big([\varphi(u_\alpha(x)) - \varphi(u_{\alpha_1, \alpha_2 - 1}(x))]$$
$$- [\varphi(u_{\alpha_1 - 1, \alpha_2}(x)) - \varphi(u_{\alpha_1 - 1, \alpha_2 - 1}(x))] \Big) \pi_0(dx)$$

To do this we sample $x_\alpha^i \sim \pi_0$, i.i.d. for $i = 1, \ldots, N_\alpha$, and then approximate

$$\mathbb{E}[\Delta\varphi_\alpha(x)] \approx \mathbb{E}_\alpha^{N_\alpha}[\Delta\varphi_\alpha(x)] := \frac{1}{N_\alpha} \sum_{i=1}^{N_\alpha} \Delta\varphi_\alpha(x_\alpha^i)$$

Observe that $\mathbb{E}[\mathbb{E}_\alpha^{N_\alpha}[\Delta\varphi_\alpha(x)]] = \mathbb{E}[\Delta\varphi_\alpha(x)]$. Furthermore, based on approximation properties of u_α, one expects a $C > 0$ such that

$$\mathbb{E}\left[(\mathbb{E}_\alpha^{N_\alpha}[\Delta\varphi_\alpha(X)] - \mathbb{E}[\Delta\varphi_\alpha(X)])^2 \right] \leq \frac{C}{N_\alpha} \prod_{i=1}^{D} 2^{-\beta_i \alpha_i}. \tag{21}$$

For the example in Sect. 2.1.1, we have $\beta_1 = \beta_2 = 4$ [29].

In particular, as we will now describe, we know how to choose the index set $\mathcal{I}$ and schedule of $\{N_\alpha\}_{\alpha \in \mathcal{I}}$ such that the following estimator delivers the same MSE for significantly smaller cost than the standard method of Sect. 3.1

$$\mathbb{E}\varphi(x) \approx \mathbb{E}_\mathcal{I}^{\text{MI}}\varphi(x) := \sum_{\alpha \in \mathcal{I}} \mathbb{E}_\alpha^{N_\alpha}[\Delta\varphi_\alpha(x)]$$

where $\mathbb{E}_\alpha^{N_\alpha}$ indicates that N_α independent samples are used at each level α. A concise, but not comprehensive, summary of the approach is given in the review [27]. For a detailed treatment see [29]. MLMC corresponds to the case in which there is one index. The MLMC methodology is more generally applicable to problems in which the target distribution—in this case the pushforward of π_0 through u, $(u)_\# \pi_0$, i.e. the distribution of $u(x)$ for $x \sim \pi_0$—needs to be approximated first to finite resolution, α, before Monte Carlo can be applied.

Assumption 3.1 There exist positive constants s_i, β_i, γ_i and C for $i = 1, 2, \ldots, D$, such that the following holds

(a) $|\mathbb{E}[\Delta\varphi_\alpha(x)]| \leq C2^{-\langle\alpha,s\rangle}$;
(b) $\mathbb{E}\left[\left(\mathbb{E}_\alpha^{N_\alpha}[\Delta\varphi_\alpha(X)] - \mathbb{E}[\Delta\varphi_\alpha(X)]\right)^2\right] \leq CN_\alpha^{-1}2^{-\langle\alpha,\beta\rangle}$;
(c) $\mathrm{COST}(\Delta\varphi_\alpha(x)) \leq C2^{\langle\alpha,\gamma\rangle}$.

For a random variable X, the cost function $\mathrm{COST}(X)$ denotes the computational complexity of a single sample of X. The following two propositions are standard results for MIMC and are proven in [29].

Proposition 3.1 *Assume Assumption 3.1, with $\beta_i > \gamma_i$, for all $i = 1, \ldots, D$. Then for the total degree index set $\mathcal{I}_L := \{\alpha \in \mathbb{N}^D : \sum_{i=1}^D \delta_i\alpha_i \leq L, \sum_{i=1}^D \delta_i = 1\}$, there are values of $L \in \mathbb{N}$, $\delta_i \in (0, 1]$ and $\{N_\alpha\}_{\alpha\in\mathcal{I}_L}$ such that*

$$\mathbb{E}\left[\left(\sum_{\alpha\in\mathcal{I}_L} \mathbb{E}_\alpha^{N_\alpha}[\Delta\varphi_\alpha(X)] - \mathbb{E}[\varphi(X)]\right)^2\right] < C\varepsilon^2, \tag{22}$$

with a computational complexity of $\mathcal{O}(\varepsilon^{-2})$ for any small $\varepsilon > 0$.

Remark 3.1 Under the same assumptions as in Proposition 3.1, if the index set is replaced with the tensor product index set $\mathcal{I}_{L_1:L_d} := \{\alpha \in \mathbb{N}^D : \alpha_1 \in \{0, \ldots, L_1\}, \ldots, \alpha_D \in \{0, \ldots, L_D\}\}$, then the same complexity result can be obtained only with an **additional constraint** that $\sum_{j=1}^D \gamma_j/s_j \leq 2$.

3.2 Monte Carlo for Inference

For simplicity, in this subsection, we define the algorithm for the target π, although we note that in practice this cannot be implemented for finite cost for our targets, and it must be replaced with π_α. This sets the stage for our method, which combines the inference approach with the approximation approach described in Sect. 3.1.

3.2.1 Markov Chain Monte Carlo and Importance Sampling

In the context of Bayesian inference, the objective is ultimately to compute expectations with respect a probability distribution π proportional to $f > 0$, where one can evaluate f but not its integral, denoted by $Z = \int f(dx)$, so $\pi(dx) = f(dx)/Z$. In particular, we define $f(dx) := L(x)\pi_0(dx)$ as the target, in the limit $\alpha \to \infty$ of (9). That is, for arbitrary $\varphi : \mathsf{X} \to \mathbb{R}$, we want to compute integrals

$$\pi(\varphi) := \int_\mathsf{X} \varphi(x)\pi(dx) = \frac{1}{Z}\int_\mathsf{X} \varphi(x)f(dx) = \frac{f(\varphi)}{Z}. \tag{23}$$

If we could simulate from π, we would approximate this by

$$\mathbb{E}^N(\varphi) := \frac{1}{N} \sum_{i=1}^{N} \varphi(x^{(i)}), \quad x^{(i)} \sim \pi. \tag{24}$$

However, in the present context this is not possible because the normalising constant Z is typically unknown and must be calculated numerically. Markov chain Monte Carlo (MCMC) and (self-normalised) importance sampling are the standard methods to solve such problems [55]. Both methods provide estimators $\widehat{\varphi}^N$ with a dimension-independent convergence rate analogous to $\mathbb{E}^N(\varphi)$, for some $C_\varphi > 0$:

$$\|\widehat{\varphi}^N - \pi(\varphi)\|_p^2 \le \frac{C_\varphi}{N}$$

For MCMC, C_φ typically depends at worst polynomially on d, and can sometimes be made independent [17, 50]. However, due to its intrinsic locality, MCMC is doomed to fail for distributions which are concentrated around several modes with low probability in between. In the case of importance sampling, the latter case is handled gracefully; however, one must be careful since often $C_\varphi = \mathcal{O}(e^d)$ [1, 3, 12]. To be precise, estimating $\int \varphi d\pi$ using samples from $\overline{\pi}$ results in $C_\varphi = \mathcal{O}(\exp(D_{KL}(\pi_\varphi \| \overline{\pi})))$, where

$$\pi_\varphi = \frac{1}{\int \varphi d\pi} \varphi \pi,$$

and $D_{KL}(\nu \| \mu)$ is the Kullback–Leibler divergence from μ to ν [12].

If one can simulate from some $\overline{\pi}$ such that $\overline{\pi}(dx) = \frac{1}{\overline{Z}} \overline{f}(dx)$ with $\overline{Z} = \int \overline{f}(dx)$, $\overline{f}(dx) = \overline{L}(x)\pi_0(dx)$, and $L/\overline{L} \le C$, then importance sampling consists of replacing the above unbiased approximation by the following biased but consistent one

$$\frac{\sum_{i=1}^{N_s} \varphi(x^{(i)}) \frac{L(x^{(i)})}{\overline{L}(x^{(i)})}}{\sum_{i=1}^{N_s} \frac{L(x^{(i)})}{\overline{L}(x^{(i)})}}, \quad x^{(i)} \sim \overline{\pi}. \tag{25}$$

MCMC methods instead proceed by constructing a Markov chain $\mathcal{M} : \mathsf{X} \times \sigma(\mathsf{X}) \to [0, 1]$, where $\sigma(\mathsf{X})$ is the sigma algebra of measurable sets, such that for all $A \in \sigma(\mathsf{X})$

$$(\pi \mathcal{M})(A) := \int_\mathsf{X} \pi(dx') \mathcal{M}(x', A) = \pi(A),$$

i.e. the Markov chain is $\pi-$invariant. Provided the Markov chain is also ergodic then one may simulate a trajectory and approximate (23) by

$$\frac{1}{N_s} \sum_{i=1+N_b}^{N_s+N_b} \varphi(x^{(i)}), \quad x^{(i)} \sim \mathcal{M}^{(i)}(x^{(0)}, \cdot). \tag{26}$$

Here, as above, N_s is the number of samples used, while N_b is the number of initial samples that are unused because we must first allow our Markov chain to approach stationarity.

The most popular MCMC method is Metropolis–Hastings (MH), which proceeds by designing a proposal Markov kernel $\mathcal{Q}$ such that the following composition Markov kernel is ergodic. First, sample $x' \sim \mathcal{Q}(x^{(i)}, dx') = q(x^{(i)}, x')\pi_0(dx')$, then let $x^{(i+1)} = x'$ with probability

$$\min\left\{1, \frac{L(x')q(x', x^{(i)})}{L(x^{(i)})q(x^{(i)}, x')}\right\}. \tag{27}$$

Otherwise, let $x^{(i+1)} = x^{(i)}$. Notice that again, as in (25), only the un-normalised target density L is required. Note that in order to customise the presentation to the context at hand, we presented a particular category of MH methods, designed for probability measures on general state spaces, which have densities with respect to π_0. Such methods are justified by the framework of [60], and a particularly convenient instantiation arises for Gaussian process priors π_0, where it is easy to define $\mathcal{Q}$ such that $q(x, x') = q(x', x)$ for all $x, x' \in X$. See [17, 50], and the more recent slice sampler variant [49].

Sequential Monte Carlo samplers combine these 2 fundamental algorithms – importance sampling, and propagation by MCMC kernels – along a sequence of intermediate targets, and are able to achieve some very impressive results. The next subsection introduces this technology.

3.2.2 Sequential Monte Carlo Samplers

Sequential Monte Carlo (SMC) samplers are able to merge the best of both worlds, by repeatedly leveraging importance sampling on a sequence of target distributions which are close. In particular, define $h_1, \ldots, h_{J-1}$ by $h_j = L_{j+1}/L_j$, where $L_1 = \overline{L}$, $L_J = L$, $\overline{L}$ appears in (25) (and may likely be π_0), $f(dx) = L(x)\pi_0(dx)$ is the un-normalised target, and for $j = 2, \ldots, J-1$, L_i interpolates in between.

Let $\pi_j = f_j/Z_j$, where $Z_j = \int f_j(dx)$ and $f_j(dx) = L_j(x)\pi_0(dx)$. Note that $\overline{f}(dx)\prod_{i=1}^{J-1} h_j(x) = f(dx)$. The idea of SMC is to simulate from $\overline{\pi} = \pi_1$ and use these samples to construct a self-normalised importance sampling estimator of f_2 with weights h_1 as in (25), and iterate for $j = 1, \ldots, J-1$, resulting in a self-normalised importance sampling estimator of π. There is, however, an obvious issue with this idea. In particular, the locations of the sampled points remain unchanged over each stage of the algorithm for this sequential importance sampling estimator. This leads to degeneracy that is no better than the original (one step) importance sampling estimator (25).

The key idea introduced in [14, 21, 36, 51] is to use Markov transition kernels between successive target distributions π_j and π_{j+1} in order to decorrelate (or "jitter") the particles, while preserving the intermediate target. The standard approach is to let $\mathcal{M}_j$ for $j = 2, \ldots, J$ be such that $(\pi_j \mathcal{M}_j)(dx) = \int \pi_j(dx')\mathcal{M}_j(x', dx) = \pi_j(dx)$, e.g. an MCMC kernel of the type introduced in the previous subsection, (27).

The resulting SMC algorithm is given in Algorithm 1. Define

$$\pi_j^N(\varphi) := \frac{1}{N} \sum_{i=1}^N \varphi(x_j^{(i)}). \tag{28}$$

Algorithm 1 SMC sampler

Let $x_1^{(i)} \sim \pi_1$ for $i = 1, \ldots, N$, and $Z_1^N = 1$. For $j = 2, \ldots, J$, repeat the following steps:

(0) Store $Z_j^N = Z_{j-1}^N \frac{1}{N} \sum_{k=1}^N h_{j-1}(x_{j-1}^{(k)})$.

(i) Define $w_j^i = h_{j-1}(x_{j-1}^{(i)}) / \sum_{k=1}^N h_{j-1}(x_{j-1}^{(k)})$, for $i = 1, \ldots, N$.

(ii) Resample. Select $I_j^i \sim \{w_j^1, \ldots, w_j^N\}$, and let $\hat{x}_j^{(i)} = x_{j-1}^{(I_j^i)}$, for $i = 1, \ldots, N$.

(iii) Mutate. Draw $x_j^{(i)} \sim M_j(\hat{x}_j^{(i)}, \cdot)$, for $i = 1, \ldots, N$.

In the resampling step of Algorithm 1, the samples are resampled according to their weights, so that some "unfit" (low weight) particles will "die" while other "fit" (high weight) ones will "multiply". As such, it can be viewed as a sort of genetic selection mechanism [20]. One can understand this operation as preserving the distribution of particles as well as the degeneracy of the sample, while exchanging variance of weights for redundancy of particles. Therefore, at a given instance, there is no net gain; however, future generations will have replenished diversity. As an example, one can use multinomial resampling. See [15] for details.

3.2.3 Estimating the Normalising Constant with SMC

Recall $Z_j = \int f_j(dx)$, and observe that

$$\pi_j(h_j) = \frac{1}{Z_j} \int \frac{L_{j+1}(x)}{L_j(x)} f_j(dx) = \frac{Z_{j+1}}{Z_j}.$$

It follows that the ratio of normalising constants of $\pi_J = \pi$ to $\pi_1 = \overline{\pi}$, $Z/\overline{Z}$, is given by

$$\frac{Z_J}{Z_1} = \prod_{j=1}^{J-1} \pi_j(h_j).$$

If $Z_1 = 1$ is known, then this is simply equal to Z, the normalising constant of π.

Observe that using Algorithm 1 we can construct an estimator of each factor by

$$\pi_j^N(h_j) = \frac{1}{N} \sum_{i=1}^N h_j(x_j^{(i)}).$$

Recall that for any $\varphi : \mathsf{X} \to \mathbb{R}$ we have $f_J(\varphi) := \int \varphi(x) f_J(dx) = f_J(1)\pi_J(\varphi)$, by definition. Now define the following estimator

$$f_J^N(\varphi) := \prod_{j=1}^{J-1} \pi_j^N(h_j)\pi_J^N(\varphi) = Z_J^N \pi_J^N(\varphi) . \qquad (29)$$

where $\pi_j^N(\varphi)$ and Z_J^N are as defined in Algorithm 1.

Note that by definition

$$\pi_J^N(\varphi) = \frac{f_J^N(\varphi)}{Z_J^N} = \frac{f_J^N(\varphi)}{f_J^N(1)} .$$

4 Multi-index Sequential Monte Carlo

With the necessary notation and concepts defined in the previous section, we now establish our theoretical results for Multi-Index Sequential Monte Carlo. Through this we can provide theoretical guarantees for the Bayesian inverse problems, such as those defined in Sect. 2.1.2 and we develop methods which apply the MIMC methodology of Sect. 3.1.2 to that problem.

The main result is an estimator which retains the well-known efficiency of SMC samplers while provably achieving the complexity benefits of MIMC. This problem has been considered before in [19, 38, 43], but the present work is the first to establish convergence guarantees under reasonable verifiable assumptions. To this end, our objective is to apply SMC samplers to estimate (23) while utilising a multi-index decomposition of the form (20).

After formulating our problem and introducing some additional notation, we present and prove our main convergence result Theorem 4.1.

4.1 Formulation

For convenience we denote the vector of multi-indices

$$\boldsymbol{\alpha}(\alpha) := (\boldsymbol{\alpha}_1(\alpha), \dots, \boldsymbol{\alpha}_{2^D}(\alpha)) \in \mathbb{Z}_+^{D \times 2^D} ,$$

where $\boldsymbol{\alpha}_1(\alpha) = \alpha$, $\boldsymbol{\alpha}_{2^D}(\alpha) = \alpha - \sum_{i=1}^{D} e_i$, and $\boldsymbol{\alpha}_i(\alpha)$ for $1 < i < 2^D$ correspond to the intermediate multi-indices involved in computing $\Delta\varphi_\alpha$, as described above (23). We note that when α is on the boundary of $\mathbb{Z}_+^D$ then several of the terms involved in $\Delta\varphi_\alpha$ are 0, but we find this notation more expedient than letting $\boldsymbol{\alpha}(\alpha) \in \mathbb{Z}_+^{D \times k_\alpha}$ where $k_\alpha = 2^{\#\{i : \alpha_i \neq 0\}} \in \{0, 2, \dots, 2^D\}$ adjusts the dimension k_α when α is on the boundary of the index set.

Define $f_\alpha(dx) := L_\alpha(x)\pi_0(dx)$, $Z_\alpha := \int_{\mathsf{X}} f_\alpha(dx)$ and $\pi_\alpha(dx) = f_\alpha(dx)/Z_\alpha$, following from (9). There are two fundamental strategies one may adopt for estimating $\pi(\varphi) = f(\varphi)/f(1)$ using a multi-telescoping identity as in (20). The first considers

the following representation

$$\pi(\varphi) = \sum_{\alpha \in \mathbb{Z}_+^D} \Delta(\pi_\alpha(\varphi_\alpha)) = \sum_{\alpha \in \mathbb{Z}_+^D} \Delta\left(\frac{1}{Z_\alpha} f_\alpha(\varphi_\alpha)\right). \tag{30}$$

Note we allow φ_α to depend on α^*—for example it could involve the solution to the PDE.

Directly estimating $\Delta(\pi_\alpha(\varphi_\alpha))$ would be quite natural if we were able to sample from a coupling of $(\pi_{\boldsymbol{\alpha}_1(\alpha)}, \ldots, \pi_{\boldsymbol{\alpha}_{2D}(\alpha)})$, i.e. a distribution $\Pi_\alpha : \sigma(X^{2^D}) \to [0, 1]$ such that

$$\int_{\boldsymbol{x}_{-j} \in X^{2^D-1}} \Pi_\alpha(d\boldsymbol{x}) = \pi_{\boldsymbol{\alpha}_j(\alpha)}(d\boldsymbol{x}_j), \qquad \text{for } j = 1, \ldots, 2^D.^2$$

In practice, however, this is non-trivial to achieve. One successful strategy for MLMC methods is to construct instead an approximate coupling Π_α such that $\pi_{\boldsymbol{\alpha}_i(\alpha)}/\Pi_\alpha$ is bounded for all $i = 1, \ldots, 2^D$, then simulate from this and construct self-normalised importance sampling estimators of the type (25) for each of the individual summands of $\Delta\left(\frac{1}{Z_\alpha} f_\alpha(\varphi_\alpha)\right)$ appearing in (30). This strategy was introduced for MLMCMC in [37] and has subsequently been applied to MIMC in the contexts of MCMC [38] and SMC [19, 43]. These MIMC works lack rigorous convergence results, due to the challenge of achieving rigorous rates for the individual summands, as well as the effect of cumbersome off-diagonal terms in the MSE estimates arising from bias of the summands (which are higher-order for MLMC). Both of these issues are handled elegantly with the present method.

In the present work, we adopt the second fundamental strategy, which is to use the ratio decomposition

$$\pi(\varphi) = \frac{f(\varphi)}{f(1)} = \frac{\sum_{\alpha \in \mathbb{Z}_+^D} \Delta(f_\alpha(\varphi_\alpha))}{\sum_{\alpha \in \mathbb{Z}_+^D} \Delta(f_\alpha(1))}. \tag{31}$$

In their limiting forms in (30) and (31), the expressions are equivalent; however, from an approximation perspective they are fundamentally different. In the context of SMC, there are advantages to the latter. In particular, this alleviates both of the issues with arising from bias of the summands in the aforementioned approach, which have prevented rigorous convergence results until now. A similar strategy was used to construct randomised MLMC estimators for Bayesian parameter estimation with particle MCMC in [11]. This method comprises the main result of this work, and its development is the topic of the following subsection.

---

^{2} Here, $\boldsymbol{x}_{-j}$ omits the jth coordinate from $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{2^D}) \in X^{2^D}$

4.2 Main Result

In order to make use of (31), we need to construct estimators of $\Delta(f_\alpha(\zeta_\alpha))$, both for our quantity of interest $\zeta_\alpha = \varphi_\alpha$ and for $\zeta_\alpha = 1$. To that end, we shall construct a coupling which approximates Π_α, and has well-behaved importance weights with respect to Π_α. Let

$$\Pi_0(d\boldsymbol{x}) = \pi_0(d\boldsymbol{x}_1) \prod_{i=2}^{2^D} \delta_{\boldsymbol{x}_1}(d\boldsymbol{x}_i), \tag{32}$$

where $\delta_{\boldsymbol{x}_1}$ denotes the Dirac delta function at $\boldsymbol{x}_1$. Note that this is an exact coupling of the prior in the sense that for any $j \in \{1, \dots, 2^D\}$

$$\int_{\boldsymbol{x}_{-j} \in \mathsf{X}^{2^D-1}} \Pi_0(d\boldsymbol{x}) = \pi_0(d\boldsymbol{x}_j). \tag{33}$$

Indeed, it is the same coupling used in Sect. 3.1.2. It is hoped that this coupling of the prior will carry over to provide error estimates analogous to (21) for the estimator (31), when computed using SMC. We note that one can estimate (31) directly by importance sampling with respect to the prior, as described in Sect. 3.1.2; however, this is not expected to be as efficient as using SMC. We hence adapt Algorithm 1 to an extended target which is an approximate coupling of the actual target as in [11, 19, 37, 38, 43], and utilise a ratio of estimates analogous to (29), similar to what was done in [11]. To this end, we define a likelihood on the coupled space as

$$\mathbf{L}_\alpha(\boldsymbol{x}) = \max\{L_{\boldsymbol{\alpha}_1(\alpha)}(\boldsymbol{x}_1), \dots, L_{\boldsymbol{\alpha}_{k_\alpha}(\alpha)}(\boldsymbol{x}_{k_\alpha})\}. \tag{34}$$

Note that $k_\alpha = 2^{\#\{i; \alpha_i \neq 0\}} \in \{0, 2, \dots, 2^D\} \leq 2^D$ adjusts the effective dimension of the target when α is on the boundary of the index set. The approximate coupling is defined by

$$F_\alpha(d\boldsymbol{x}) = \mathbf{L}_\alpha(\boldsymbol{x})\Pi_0(d\boldsymbol{x}), \qquad \Pi_\alpha(d\boldsymbol{x}) = \frac{1}{F_\alpha(1)} F_\alpha(d\boldsymbol{x}). \tag{35}$$

Example 4.1 (*Approximate Coupling*) As an example of the approximate coupling constructed in equations (32), (34), and (35), let $D = 2$, $d = 1$, and $\alpha = (1, 1)$. Then, we have

$$\Pi_{(1,1)}(x_1, x_2, x_3, x_4) \propto \max\{L_{11}(x_4), L_{10}(x_3), L_{01}(x_2), L_{00}(x_1)\}$$
$$\pi_0(x_1)\delta_{x_1}(x_2)\delta_{x_1}(x_3)\delta_{x_1}(x_4).$$

Note that for our choice of prior coupling (32), we effectively have a single distribution

$$\Pi_{(1,1)}(x) \propto \max\{L_{11}(x), L_{10}(x), L_{01}(x), L_{00}(x)\}\pi_0(x),$$

but any coupling of the prior which preserves the marginals as in (33) is admissible, so we prefer to consider this as a target on the "diagonal hyperplane" $x_1 = x_2 = x_3 = x_4$, as above.

Let $H_{\alpha,j} = \mathbf{L}_{\alpha,j+1}/\mathbf{L}_{\alpha,j}$ for some intermediate distributions $F_{\alpha,1}, \ldots, F_{\alpha,J} = F_\alpha$. In our case, we use the natural intermediate targets $F_{\alpha,j}(d\mathbf{x}) = \mathbf{L}_\alpha(\mathbf{x})^{\tau_j}\Pi_0(d\mathbf{x})$, where $\tau_1 = 0$, $\tau_j < \tau_{j+1}$, and $\tau_J = 1$. For $j = 1, \ldots, J$, we define

$$\Pi_{\alpha,j}(d\mathbf{x}) = \frac{1}{F_{\alpha,j}(1)} F_{\alpha,j}(d\mathbf{x})$$

and we let $\boldsymbol{\mathcal{M}}_{\alpha,j}$ be a Markov transition kernel such that $(\Pi_{\alpha,j}\boldsymbol{\mathcal{M}}_{\alpha,j})(d\mathbf{x}) = \Pi_{\alpha,j}(d\mathbf{x})$, analogous to $\mathcal{M}$ in Sect. 3.2.2. Any MCMC kernel as described in Sect. 3.2.1 with target distribution $\Pi_{\alpha,j}$ is suitable for this purpose. An example is the Metropolis–Hastings kernel described above and in (27).

Remark 4.1 (*Tempering*) Tempering accurately is crucial, because if the effective sample size drops too low, then the population will lack sufficient diversity to survive. The purpose of the sequential resampling and mutation is precisely to preserve diversity in the sample. Sometimes a fixed tempering schedule is suitable for this purpose, for example $\tau_j = (j-1)/(J-1)$. An alternative is to use an adaptive tempering strategy.

Given a (possibly un-normalised) weighted sample $\{w^{(k)}, \mathbf{x}^{(k)}\}_{k=1}^N$, the *effective sample size* (ESS) is defined as follows

$$\text{ESS} = \frac{\left(\sum_{k=1}^N w^{(k)}\right)^2}{\sum_{k=1}^N (w^{(k)})^2}$$

This quantity serves as a proxy for the variance of the weighted sample. To understand the name, note that if $w^{(k)} \propto 1$ for all k, then ESS $= N$, while if $w^{(k^*)} \propto 1$ for some k^* and $w^{(k)} = 0$ for $k \neq k^*$ then ESS $= 1$. If $\tau_j = \tau_{j-1} + h$, for $h > 0$, then the intermediate weights will be $w^{(k)} = \mathbf{L}_\alpha(\mathbf{x}^{(k)})^h$, and the corresponding ESS(h) is a scalar function of h which quantifies the sample attrition that results from the importance sampling step; precisely what we are aiming to control. The adaptive tempering parameter τ_j is computed by firstly solving ESS(h) = ESS$_{\text{min}}$ with a pre-specified value of ESS$_{\text{min}}$, and then letting $\tau_j \leftarrow \tau_{j-1} + h$. In this way, the effective sample size is ESS$_{\text{min}}$ each time importance sampling is carried out. The tempering procedure is carried until $\tau_j = 1$.

Remark 4.2 (*Role of Dimension D*) Note that in high dimensions one would select an index set in which there are few (or no) terms on the interior. In the present work, we do not explicitly consider the dependence on D (which is reasonable for small $D \leq 5$ say); however, the methodology is applicable for high-dimensional targets and that is the subject of future work. The cost of simulating the approximate coupling at level α will feature a constant 2^D multiplying Assumption 4.2(C), because that is how many likelihood evaluations are required to compute (34), and hence corresponding multi-increment. The constant can be large, but this will not alter the complexity estimates.

Algorithm 2 SMC sampler for coupled estimation of $\Delta(f_\alpha(\zeta_\alpha))$

Let $x_1^{(i)} \sim \pi_1$ for $i = 1, \ldots, N$, $Z_1^N = 1$, and $\omega_{1,k} = 1$ for $k = 1, \ldots, 2^D$. For $j = 2, \ldots, J$, $k = 1, \ldots, 2^D$, repeat the following steps for $i = 1, \ldots, N$:

(0) Store $Z_j^N = Z_{j-1}^N \frac{1}{N} \sum_{k=1}^N H_{\alpha,j-1}(x_{j-1}^{(k)})$.

(i) Define $w_j^i = H_{\alpha,j-1}(x_{j-1}^{(i)}) / \sum_{k=1}^N H_{\alpha,j-1}(x_{j-1}^{(k)})$.

(ii) Resample. Select $I_j^i \sim \{w_j^1, \ldots, w_j^N\}$, and let $\hat{x}_j^{(i)} = x_{j-1}^{(I_j^i)}$.

(iii) Mutate. Draw $x_j^{(i)} \sim \mathcal{M}_{\alpha,j}(\hat{x}_j^{(i)}, \cdot)$.

For $j = 1, \ldots, J$, and for random variables $x_j^{(i)}$, $i = 1, \ldots, N$ (which will be sampled from the Markov chain $\mathcal{M}_{\alpha,j}$) we define

$$\Pi_{\alpha,j}^N(dx) := \frac{1}{N} \sum_{i=1}^N \delta_{x_j^{(i)}}(dx), \tag{36}$$

and then define

$$Z_\alpha^N := \prod_{j=1}^{J-1} \Pi_{\alpha,j}^N(H_{\alpha,j}), \quad F_\alpha^N(dx) := Z_\alpha^N \Pi_{\alpha,J}^N(dx). \tag{37}$$

We require the following assumption

Assumption 4.1 Let $J \in \mathbb{N}$ be given, and let X be a Banach space. For each $j \in \{1, \ldots, J\}$ there exists some $C > 0$ such that for all $(\alpha, x) \in \mathbb{Z}_+^D \times \mathsf{X}^{2^D}$

$$C^{-1} < Z, H_{\alpha,j}(x), \mathbf{L}_\alpha(x) \le C.$$

The following proposition can easily be deduced from [Theorem 7.4.2, [20]].

Proposition 4.1 *Under Assumption* 4.1, *we have* $\mathbb{E}[F_\alpha^N(\psi)] = F_\alpha(\psi)$.

Define

$$\psi_{\zeta_\alpha}(x) := \sum_{k=1}^{2^D} \iota_k \omega_k(x) \zeta_{\alpha_k(\alpha)}(x_k), \quad \omega_k(x) := \frac{\mathbf{L}_{\alpha_k(\alpha)}(x_k)}{\mathbf{L}_\alpha(x)}, \tag{38}$$

where $\iota_k \in \{-1, 1\}$ is the sign of the kth term in Δf_α and $\zeta_\alpha : \mathsf{X} \to \mathbb{R}$. Following from Proposition 4.1, we have that

$$\mathbb{E}[F_\alpha^N(\psi_{\zeta_\alpha})] = F_\alpha(\psi_{\zeta_\alpha}) = \Delta f_\alpha(\zeta_\alpha). \tag{39}$$

Now given $\mathcal{I} \subseteq \mathbb{Z}_+^D$, $\{N_\alpha\}_{\alpha \in \mathcal{I}}$, and $\varphi : \mathsf{X} \to \mathbb{R}$, for each $\alpha \in \mathcal{I}$, run an independent SMC sampler as in Algorithm 2 with N_α samples, define $Z_\alpha^N = Z_J^N$, and define the

MIMC estimator as

$$\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} = \frac{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\varphi_\alpha})}{\max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\}}, \tag{40}$$

where $Z_{\min}$ is a lower bound on Z as given in Assumption 4.1, and $F_\alpha^{N_\alpha}$ is defined in (37).

4.2.1 Theoretical Results

Throughout this subsection, $C > 0$ is a constant whose value may change from line to line. The following theorem is the main theoretical result which underpins the results to follow.

Theorem 4.1 *Assume Assumption 4.1. Then, for any $J \in \mathbb{N}$ there exists a $C > 0$ such that for any $N \in \mathbb{N}$, $\psi : \mathsf{X}^{2^D} \to \mathbb{R}$ bounded and measurable and $\alpha \in \mathbb{Z}_+^D$*

$$\mathbb{E}\left[|F_\alpha^N(\psi) - F_\alpha(\psi)|^2\right] \leq \frac{C}{N} F_\alpha(\psi^2).$$

Furthermore,

$$F_\alpha(\psi_{\zeta_\alpha}^2) \leq C \int_{\mathsf{X}} (\Delta(L_\alpha(x)\zeta_\alpha(x)))^2 \pi_0(dx),$$

where $\psi_{\zeta_\alpha}(x)$ is as (38).

Proof The first result follows from Lemmas 5.1, 5.2 and 5.3, given in Sect. 5. The second result is derived as follows:

$$F_\alpha(\psi_{\zeta_\alpha}^2) = \int_{\mathsf{X}^{2^D}} \left(\sum_{k=1}^{2^D} \iota_k \frac{L_{\boldsymbol{\alpha}_k(\alpha)}(x_k)}{L_\alpha(x)} \zeta_{\boldsymbol{\alpha}_k(\alpha)}(x_k)\right)^2 L_\alpha(x) \Pi_0(dx)$$

$$= \int_{\mathsf{X}^{2^D}} \frac{1}{L_\alpha(x)} \left(\sum_{k=1}^{2^D} \iota_k L_{\boldsymbol{\alpha}_k(\alpha)}(x_k) \zeta_{\boldsymbol{\alpha}_k(\alpha)}(x_k)\right)^2 \Pi_0(dx)$$

$$\leq C \int_{\mathsf{X}} (\Delta(L_\alpha(x)\zeta_\alpha(x)))^2 \pi_0(dx).$$

The first 2 lines are direct substitution and the inequality follows by defining $C^{-1} = \inf_{x \in \mathsf{X}^{2^D}} L_\alpha(x)$ and using the definition of Π_0 in (32). □

Following from above, the assumptions below will be made.

Assumption 4.2 For any $\zeta : \mathsf{X} \to \mathbb{R}$ bounded and Lipschitz, there exist $C, \beta_i, s_i, \gamma_i > 0$ for $i = 1, \ldots, D$ such that for resolution vector $(2^{-\alpha_1}, \ldots, 2^{-\alpha_D})$, i.e. resolution $2^{-\alpha_i}$ in the ith direction, the following holds:

(B) $|\Delta f_\alpha(\zeta)| =: B_\alpha \le C2^{-\langle\alpha,s\rangle}$;

(V) $\int_X (\Delta(L_\alpha(x)\zeta_\alpha(x)))^2 \pi_0(dx) =: V_\alpha \le C2^{-\langle\alpha,\beta\rangle}$;

(C) $\mathrm{COST}(F_\alpha(\psi_\varphi)) =: C_\alpha \propto 2^{\langle\alpha,\gamma\rangle}$.

The proofs of the main theorems will rely on one more result, Lemma 4.1, given immediately afterwards.

The next theorem comprises the main result of the paper.

Theorem 4.2 *Assume Assumptions* 4.1 *and* 4.2, *with* $\beta_i > \gamma_i$ *for* $i = 1, \ldots, D$. *Then, for any* $\varepsilon > 0$ *and suitable* $\varphi : X \to \mathbb{R}$, *it is possible to choose a total degree index set* $\mathcal{I}_L := \{\alpha \in \mathbb{N}^D : \sum_{i=1}^{D} \delta_i \alpha_i \le L, \sum_{i=1}^{D} \delta_i = 1\}$, $\delta_i \in (0, 1]$ *and* $\{N_\alpha\}_{\alpha \in \mathcal{I}_L}$, *such that for some* $C > 0$

$$\mathbb{E}[(\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} - \pi(\varphi))^2] \le C\varepsilon^2 \,,$$

and $\mathrm{COST}(\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}}) \le C\varepsilon^{-2}$, *the canonical rate. The estimator* $\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}}$ *is defined in equation* (40).

Proof Starting from Lemma 4.1 and given Theorem 4.1, and the Assumptions 4.2, the result follows in a similar fashion to standard MIMC theory [19, 27, 29, 38, 43]. The proof is given in "Appendix A" for completeness. □

Remark 4.3 Under the same assumptions as in Theorem 4.2, and similar to Proposition 3.1, if the index set is replaced with the tensor product index set $\mathcal{I}_{L_1:L_d} := \{\alpha \in \mathbb{N}^D : \alpha_1 \in \{0, \ldots, L_1\}, \ldots, \alpha_D \in \{0, \ldots, L_D\}\}$, then the same complexity result can be obtained only with an **additional constraint** that $\sum_{j=1}^{D} \gamma_j/s_j \le 2$.

Lemma 4.1 *For the estimator* (40) $\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} = \dfrac{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\varphi_\alpha})}{\max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\}}$, *the following inequality holds*

$$\mathbb{E}[(\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} - \pi(\varphi))^2] \le C \max_{\zeta \in \{\varphi, 1\}} \left(\sum_{\alpha \in \mathcal{I}} \mathbb{E}\left[\left(F_\alpha^{N_\alpha}(\psi_{\zeta_\alpha}) - F_\alpha(\psi_{\zeta_\alpha})\right)^2\right] + \left(\sum_{\alpha \notin \mathcal{I}} F_\alpha(\psi_{\zeta_\alpha})\right)^2 \right).$$

Proof Recall that from (40) we have $\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} = \dfrac{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\varphi_\alpha})}{\max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\}}$. So

$$\mathbb{E}[(\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} - \pi(\varphi))^2] = \mathbb{E}\left[\left(\frac{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\varphi_\alpha})}{\max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\}} - \frac{f(\varphi)}{f(1)}\right)^2\right]$$

$$= \mathbb{E}\left[\left(\frac{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\varphi_\alpha})}{\max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\}f(1)}\left(f(1) - \max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\}\right)\right.\right.$$

$$\left.\left. + \frac{1}{f(1)}\left(\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\varphi_\alpha}) - f(\varphi)\right)\right)^2\right].$$

Since $f(1) \geq Z_{\min}$ and $|\max\{A, Z\} - \max\{B, Z\}| \leq |A - B|$, we have

$$\mathbb{E}\left[\left(\max\{\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1), Z_{\min}\} - f(1)\right)^2\right] \leq \mathbb{E}\left[\left(\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1) - f(1)\right)^2\right]. \tag{41}$$

Then, we have

$$\mathbb{E}[(\widehat{\varphi}_{\mathcal{I}}^{\mathrm{MI}} - \pi(\varphi))^2] \leq C \max_{\zeta \in \{\varphi, 1\}} \mathbb{E}\left[\left(\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\zeta_\alpha}) - f(\zeta)\right)^2\right]$$

$$\leq C \max_{\zeta \in \{\varphi, 1\}} \mathbb{E}\left[\left(\sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_{\zeta_\alpha}) - \sum_{\alpha \in \mathcal{I}} F_\alpha(\psi_{\zeta_\alpha})\right)^2\right.$$

$$\left. + \left(\sum_{\alpha \in \mathcal{I}} F_\alpha(\psi_{\zeta_\alpha}) - f(\zeta)\right)^2\right]$$

$$= C \max_{\zeta \in \{\varphi, 1\}} \left(\sum_{\alpha \in \mathcal{I}} \mathbb{E}\left[\left(F_\alpha^{N_\alpha}(\psi_{\zeta_\alpha}) - F_\alpha(\psi_{\zeta_\alpha})\right)^2\right] + \left(\sum_{\alpha \notin \mathcal{I}} F_\alpha(\psi_{\zeta_\alpha})\right)^2\right).$$

$\square$

Remark 4.4 It is remarked that one **always has** $Z > 0$; therefore, given a target error level ε, one can always **replace** $Z_{\min} \leftarrow \varepsilon$, and achieve the same result. To see this, denoting

$$\hat{Z}^{\mathrm{MI}} = \sum_{\alpha \in \mathcal{I}} F_\alpha^{N_\alpha}(\psi_1),$$

observe that line (41) can be replaced with

$$|\max\{\hat{Z}^{\mathrm{MI}}, \varepsilon\} - Z| \leq |\hat{Z}^{\mathrm{MI}} - Z| + |\varepsilon| + |\max\{Z - \varepsilon, 0\} - \max\{Z, 0\}|$$

$$\leq |\hat{Z}^{\mathrm{MI}} - Z| + 2|\varepsilon|.$$

Theorem 4.2 formulates the total degree index set with general δ satisfying some loose conditions. In the paper [29], optimal δ is constructed according to a profit indicator. The focus of the present work is on the canonical case, where the complexity is dominated by low levels, so we simply choose $\delta \propto s$. The proof of Theorem 4.2 is based on the general δ, and it is easy to see that this choice suffices.

To achieve the canonical rate of complexity, Theorem 4.3 with the tensor product index set relies on the essential assumption that $\sum_{j=1}^{D} \frac{\gamma_j}{s_j} \leq 2$, which ensures that the samples at the finest index do not dominate the cost. If the assumption is violated, then only the subcanonical complexity $\sum_{j=1}^{D} \frac{\gamma_j}{s_j}$ can be achieved. This rate may often

be $D-$dependent, resulting in a so-called *curse of dimensionality*. In comparison, Theorem 4.2 with the total degree index set releases this constraint, and improves the computational complexity for many problems from subcanonical to canonical, as illustrated in the numerical examples.

4.3 Verification of Assumptions

Here we briefly discuss the models considered before in connection with the required Assumptions 4.2. Note that both posteriors have the form $\exp(\Phi(x))$ for some $\Phi : X \to \mathbb{R}$, and are approximated by $\Phi_\alpha : X \to \mathbb{R}$.

Proposition 4.2 *Let X be a Banach space with $D = 2$ s.t. $\pi_0(X) = 1$, with norm $\|\cdot\|_X$. For all $\epsilon > 0$, there exists a $C(\epsilon) > 0$ such that the following holds for Φ, Φ_α given by (13), (16), or $\log(L)$, $\log(L_\alpha)$ from (9), respectively:*

$$\Delta \exp(\Phi_\alpha(x_\alpha)) \leq C(\epsilon) \exp(\epsilon \|x\|_X^2)\Big(|\Delta \Phi_\alpha(x_\alpha)| + |\Delta_1 \Phi_{\alpha-e_2}(x_{\alpha-e_2})|$$

$$\times |\Delta_2 \Phi_{\alpha-e_1}(x_{\alpha-e_1})|\Big).$$

Proof Let us introduce the shorthand notation $A_{11} = \Phi_\alpha(x_\alpha)$, $A_{10} = \Phi_{\alpha-e_2}(x_{\alpha-e_2})$, $A_{01} = \Phi_{\alpha-e_1}(x_{\alpha-e_1})$, $A_{00} = \Phi_{\alpha-e_1-e_2}(x_{\alpha-e_1-e_2})$. We have

$$\begin{aligned}
\Delta \exp(\Phi_\alpha(x_\alpha)) &= \exp(A_{11}) - \exp(A_{10}) - (\exp(A_{01}) - \exp(A_{00})) \\
&= \exp(A_{10})(\exp(A_{11} - A_{10}) - 1) - \exp(A_{00})(\exp(A_{01} - A_{00}) - 1) \\
&= \exp(A_{10})(\exp(A_{11} - A_{10}) - \exp(A_{01} - A_{00})) \\
&\quad + (\exp(A_{01}) - \exp(A_{00}))(\exp(A_{10} - A_{00}) - 1) \\
&\leq C(\epsilon) \exp(\epsilon \|x\|_X^2)(|A_{11} - A_{10} - (A_{01} - A_{00})| \\
&\quad + |A_{01} - A_{00}||A_{10} - A_{00}|)\,,
\end{aligned}$$

where we have added and subtracted $\exp(A_{10})(\exp(A_{01} - A_{00}) - 1)$ in going from the second to the third line. The final line follows from the mean value theorem and equations (59) and (72) with $X = H_r^m$, $r > 1/2$. These trivially hold for (9).

The issue which prevented us from achieving above for LGC is that terms like $\exp(-A_{10}) \propto \exp(-\Phi_\alpha(x_\alpha))$ appear in the constant, which involve a factor like $\exp(\int \exp(x(z))dz)$. Fernique theorem does not guarantee that such double exponentials are finite. However, for LGP, we instead have

$$\exp(-A_{10}) \propto (\int_\Omega \exp(x(z))dz)^n \leq |\Omega|^n \exp(n\|x\|_{L^\infty(\Omega)}) \leq |\Omega|^n \exp(n\|x\|_r).$$

$\square$

PDE. The following proposition updates Proposition 2.1 and is proven in the literature on mixed regularity of the solution of elliptic PDE, as mentioned already in Sect. 3.1.2. See, e.g. [29] and references therein.

Proposition 4.3 *Let u_α be the solution of (2) and (3) at resolution α, as described in Sect. 2.1.1, for $a(x)$ given by (5) and uniformly over $x \in [-1, 1]^d$, and f suitably smooth. Then, there exists a $C > 0$ such that*

$$\|\Delta u_\alpha(x)\|_V \le C 2^{-\alpha_1 - \alpha_2} .$$

Furthermore,

$$\|\Delta u_\alpha(x)\| \le C 2^{-2(\alpha_1 + \alpha_2)} .$$

Note that since $L_\alpha(x) \le C < \infty$ by Assumption 4.1, the constant in Proposition 4.2 can be made uniform over x, and hence the required rate in Assumption 4.2 is established immediately.

LGP. Will restrict consideration to LGP here, since LGC features double exponentials which are difficult to handle theoretically in this context. The following proposition updates Proposition 2.3 as required for differences of differences.

Proposition 4.4 *Let $x \sim \pi_0$, where π_0 is a Gaussian process of the form (10) with spectral decay corresponding to (11), and let x_α correspond to truncation on the index set $\mathcal{A}_\alpha = \cap_{i=1}^2 \{|k_i| \le 2^{\alpha_i}\}$ as in (14). Then there is a $C > 0$ such that for all $q < (\beta - 1)/2$*

$$\|\Delta x_\alpha\|^2 \le C \|x\|_q^2 2^{-2q \sum_{i=1}^2 \alpha_i} .$$

Proof The proof follows along the same lines as that of Proposition 2.3 (B.1), except instead of projection onto $\cup_{i=1}^2 \{|k_i| > 2^{\alpha_i}\}$, the projection here is onto the set of indices $\cap_{i=1}^2 \{2^{\alpha_i - 1} \le |k_i| \le 2^{\alpha_i}\}$, i.e.

$$\|A^{-q/2} P_{\cap_{i=1}^2 \{2^{\alpha_i - 1} \le |k_i| \le 2^{\alpha_i}\}}\|_{\mathcal{L}(L^2, L^2)} \le C 2^{-q \sum_{i=1}^2 \alpha_i} .$$

$\square$

The key phenomenon that takes place is that the difference of difference Δx_α leaves a remainder which is an intersection $\cap_{i=1}^2 \{2^{\alpha_i - 1} \le |k_i| \le 2^{\alpha_i}\}$, rather than the union $\cup_{i=1}^2 \{2^{\alpha_i - 1} \le |k_i| \le 2^{\alpha_i}\}$, associated with the truncation error in Proposition 2.3, which one would achieve with a single difference from $x_\alpha - x_{\alpha-1}$. This eliminates all indices in which $k_i^{-1} = \mathcal{O}(1)$ for some i, and provides the required product-form rates.

Proposition 4.5 *The rate from Proposition 4.4 is inherited by the likelihood, resulting in verification of Assumption 4.2(V) with $\beta_i = \beta$.*

Proof The proof follows along the lines of Proposition 2.5. In this case, following from Proposition 4.2, for all $\epsilon > 0$ and $q < \beta/2$, we have

$$\mathbb{E}_{\pi_0} (\Delta L_\alpha(x_\alpha))^2 \le \mathbb{E}_{\pi_0} C(\epsilon) \exp(\epsilon \|x\|_r^2) \Big(\Delta \Phi_\alpha(x_\alpha) + \Delta_1$$

$$\Phi_{\alpha-e_2}(x_{\alpha-e_2})\Delta_2\Phi_{\alpha-e_1}(x_{\alpha-e_1})\Big)^2$$

$$\leq C2^{-2q\sum_{i=1}^{2}\alpha_i},$$

where the second line is computed with estimates similar to (72), and Fernique Theorem to conclude, as in the proof of Proposition 2.5. □

5 Proofs Relating Theorem 4.1

In this section, we prove Lemmas 5.1, 5.2 and 5.3 from which Theorem 4.1 is an immediate consequence. We fix $\alpha \in \mathcal{I}$ throughout this section and thus, to avoid notational overload, we henceforth suppress it from our notation.

For $j = 2, \ldots, J$, we define

$$\Phi_j(\Pi) := \frac{\Pi(H_{j-1}\mathcal{M}_j)}{\Pi(H_{j-1})}, \tag{42}$$

and observe that the iterates of the algorithm of Sect. 3.2.2 can be rewritten in the concise form

$$x_j^i \sim \Phi_j(\Pi_{j-1}^N), \quad \text{for } i = 1, \ldots, N, \tag{43}$$

where we recall the definition (28) for the empirical measure Π_{j-1}^N. Let $\Phi_1(\Pi) := \Pi$. A finer error analysis beyond Proposition 4.1 requires keeping track of the effect of errors $(\Pi_j^N - \Phi_j(\Pi_{j-1}^N))$ and accounting for the cumulative error at time J. To this end, and for any $\psi : X^{2^D} \to \mathbb{R}$ bounded, we define the following partial propagation operator for $p = 1, \ldots, J$

$$Q_{n,J}(\psi)(x_n) = \int_{X^{2^D \times (J-n)}} \psi(x_J) \prod_{j=n+1}^{J} H_{j-1}(x_{j-1})\mathcal{M}_j(x_{j-1}, x_j)dx_{n+1:J}, \tag{44}$$

where $Q_{J,J} = I_{2^D}$, i.e. $Q_{J,J}(\psi)(x_J) = \psi(x_J)$. We will assume for simplicity that $F_1 := \Pi_0$, the prior, so that $F_1(1) = 1$. Note that then $\Pi_0(Q_1(\psi)(x_1)) = F_J(\psi) = F(\psi)$.

We present now the following well-known representation of the error as a martingale w.r.t. the natural filtration of the particle system (see [11, 20])

$$F_J^N(\psi) - F(\psi) = \sum_{n=1}^{J} \underbrace{F_n^N(1)\left[\Pi_n^N - \Phi_n(\Pi_{n-1}^N)\right](Q_{n,J}(\psi))}_{S_{n,J}^N(\psi)}, \tag{45}$$

where we denote the summands as $S_{n,J}^N(\psi)$. This clearly shows the unbiasedness property presented in Proposition 4.1. In particular, $\mathbb{E}F_J^N(\psi) = F(\psi)$, as can be seen by backwards induction conditioning first on $\{x_{J-1}^i\}_{i=1}^N$ and recalling the form of $F_n^N(1)$ given in (37). This brings us to our first supporting lemma. Throughout these calculations, C is a finite constant whose value may change on each appearance. The dependencies of this constant on the various algorithmic parameters are made clear from the statement.

Lemma 5.1 *Assume Assumption 4.1. Then, for any $J \in \mathbb{N}$ there exists $C > 0$ such that for any $N \in \mathbb{N}$ and any $\psi : \mathsf{X}^{2^D} \to \mathbb{R}$ bounded and measurable*

$$\mathbb{E}[(F_J^N(\psi) - F(\psi))^2] \leq \frac{C}{N} \sum_{n=1}^{J} \mathbb{E}[(Q_{n,J}(\psi)(x_n^1))^2].$$

Proof Following from (45), we have

$$\mathbb{E}(F_J^N(\psi) - F(\psi))^2 \leq C \sum_{n=1}^{J} \mathbb{E}[(S_{n,J}^N(\psi))^2]$$

$$\leq \frac{C}{N} \mathbb{E}[(Q_{n,J}(\psi)(x_n^1))^2].$$

The first inequality results from application of the Burkholder–Gundy–Davis inequality. The second inequality follows via an application of the conditional Marcinkiewicz–Zygmund inequality and the fact that $F_n^N(1)$ is upper-bounded by a constant via Assumption 4.1. $\qquad\square$

Lemma 5.2 *Assume Assumption 4.1. Then, for any $J \in \mathbb{N}$ and $n \in \{1, \ldots, J\}$ there exists a $C > 0$ such that for any $N \in \mathbb{N}$ and $\psi : \mathsf{X}^{2^D} \to \mathbb{R}$ bounded and measurable*

$$(Q_{n,J}(\psi)(x_n))^2 \leq C Q_{n,J}(\psi^2)(x_n).$$

Proof Observe that for any $x_n \in \mathsf{X}^{2^D}$, $Q_{n,J}(\psi)(x_n)/Q_{n,J}(1)(x_n)$ is a probability distribution. Therefore, Jensen's inequality provides

$$(Q_{n,J}(\psi)(x_n))^2 = (Q_{n,J}(1)(x_n))^2 \left(\frac{Q_{n,J}(\psi)(x_n)}{Q_{n,J}(1)(x_n)}\right)^2$$

$$\leq Q_{n,J}(1)(x_n) Q_{n,J}(\psi^2)(x_n).$$

The result follows with $C = \sup_{x_n \in \mathsf{X}^{2^D}} Q_{n,J}(1)(x_n)$. $\qquad\square$

Lemma 5.3 *Assume Assumption 4.1. Then for any $J \in \mathbb{N}$ and $n \in \{1, \ldots, J\}$ there exists a $C > 0$ such that for any $N \in \mathbb{N}$ and any $\psi : \mathsf{X}^{2^D} \to \mathbb{R}$ bounded and measurable*

$$\mathbb{E}[Q_{n,J}(\psi^2)(x_n^1)] \leq C F(\psi^2).$$

Proof We proceed by induction. The result for $n = 1$ follows immediately from Lemma 5.2 and the fact that we defined $\Phi_1(\Pi_1) = \Pi_1 = F_1 = \Pi_0$:

$$\Pi_0(Q_{1,J}(\psi^2)(\boldsymbol{x}_1)) = \int_{\mathsf{X}^{2D \times J}} \psi^2(\boldsymbol{x}_J)\Pi_0(\boldsymbol{x}_1) \prod_{j=1}^{J-1} H_j(\boldsymbol{x}_j)\mathcal{M}_{j+1}(\boldsymbol{x}_j, \boldsymbol{x}_{j+1})d\boldsymbol{x}_{1:J}$$

$$= \int_{\mathsf{X}^{2D \times J}} \psi^2(\boldsymbol{x}_J)F(\boldsymbol{x}_J)d\boldsymbol{x}_J .$$

Now, assume the result holds for $n - 1$:

$$\mathbb{E}[Q_{n-1,J}(\psi^2)(\boldsymbol{x}_{n-1}^1)] = \mathbb{E}(\Phi_{n-1}(\Pi_{n-2}^N)[Q_{n-1,J}(\psi^2)]) \leq CF(\psi^2), \qquad (46)$$

and we will show that this implies it holds for n.

We have that

$$\Phi_n(\Pi_{n-1}^N)[Q_{n,J}(\psi^2)] = \underbrace{\Phi_n(\Phi_{n-1}(\Pi_{n-2}^N))[Q_{n,J}(\psi^2)]}_{T_1}$$

$$+ \underbrace{\{\Phi_n(\Pi_{n-1}^N) - \Phi_n(\Phi_{n-1}(\Pi_{n-2}^N))\}[Q_{n,J}(\psi^2)]}_{T_2} .$$

We consider bounding the expectations of T_1 and T_2 in turn.
T_1. We have

$$\Phi_n(\Phi_{n-1}(\Pi_{n-2}^N))[Q_{n,J}(\psi^2)] = \frac{1}{\Phi_{n-1}(\Pi_{n-2}^N)(H_{n-1})}$$

$$\times \int_{\mathsf{X}^{2D \times 2}} \Phi_{n-1}(\Pi_{n-2}^N)(d\boldsymbol{x}_{n-1})H_{n-1}(\boldsymbol{x}_{n-1})\mathcal{M}_n(\boldsymbol{x}_{n-1}, d\boldsymbol{x}_n)Q_{n,J}(\psi^2)(\boldsymbol{x}_n) .$$

By Assumption 4.1 $\inf_{\boldsymbol{x}} H_{n-1}(\boldsymbol{x}) \geq C^{-1}$ and

$$Q_{n-1,J}(\psi^2)(\boldsymbol{x}_{n-1}) = \int_{\mathsf{X}^{2D}} H_{n-1}(\boldsymbol{x}_{n-1})\mathcal{M}_n(\boldsymbol{x}_{n-1}, d\boldsymbol{x}_n)Q_{n,J}(\psi^2)(\boldsymbol{x}_n) .$$

Therefore, by the inductive hypothesis

$$\mathbb{E}\left(\Phi_n(\Phi_{n-1}(\Pi_{n-2}^N))[Q_{n,J}(\psi^2)]\right) \leq C\mathbb{E}\left(\Phi_{n-1}(\Pi_{n-2}^N)[Q_{n-1,J}(\psi^2)]\right)$$

$$\leq CF(\psi^2) .$$

T_2. For the second term, we have

$$\left| \{\Phi_n(\Pi_{n-1}^N) - \Phi_n(\Phi_{n-1}(\Pi_{n-2}^N))\}[Q_{n,J}(\psi^2)] \right|$$

$$= \left| \frac{\Pi_{n-1}^N(H_{n-1}M_n Q_{n,J}(\psi^2))}{\Pi_{n-1}^N(H_{n-1})} - \frac{\Phi_{n-1}(\Pi_{n-2}^N)(H_{n-1}M_n Q_{n,J}(\psi^2))}{\Phi_{n-1}(\Pi_{n-2}^N)(H_{n-1})} \right|$$

$$\leq \underbrace{\left| \frac{1}{\Pi_{n-1}^N(H_{n-1})}(\Pi_{n-1}^N - \Phi_{n-1}(\Pi_{n-2}^N))(Q_{n-1,J}(\psi^2)) \right|}_{T_{2,1}}$$

$$+ \underbrace{\left| \frac{(\Phi_{n-1}(\Pi_{n-2}^N) - \Pi_{n-1}^N)(H_{n-1})}{\Pi_{n-1}^N(H_{n-1})\Phi_{n-1}(\Pi_{n-2}^N)(H_{n-1})} \Phi_{n-1}(\Pi_{n-2}^N)(Q_{n-1,J}(\psi^2)) \right|}_{T_{2,2}} .$$

These two terms are now considered.

$T_{2,1}$. The expected value of $T_{2,1}$ can be bounded as follows:

$$\mathbb{E}\left[\left| \left(\Pi_{n-1}^N - \Phi_{n-1}(\Pi_{n-2}^N) \right)(Q_{n-1,J})(\psi^2) \right| \right] \leq \mathbb{E}\left[\left| \Pi_{n-1}^N(Q_{n-1,J})(\psi^2) \right| \right]$$

$$+ \mathbb{E}\left[\left| \Phi_{n-1}(\Pi_{n-1}^N)(Q_{n-1,J}(\psi^2)) \right| \right]$$

$$= \mathbb{E}\left[\Pi_{n-1}^N(Q_{n-1,J})(\psi^2) \right]$$

$$+ \mathbb{E}\left[\Phi_{n-1}(\Pi_{n-1}^N)(Q_{n-1,J}(\psi^2)) \right]$$

$$\leq 2\mathbb{E}\left[\Phi_{n-1}(\Pi_{n-1}^N)(Q_{n-1,J}(\psi^2)) \right]$$

$$\leq 2CF(\psi^2) .$$

Where the triangle inequality is used in the first line, positivity is used in the second, the Martingale property is used in the third, and the induction hypothesis is used to conclude. Thus, after appropriately redefining the constant C, we have that

$$\mathbb{E}[T_{2,1}] \leq CF(\psi^2).$$

$T_{2,2}$. Finally, for the second term, note that by Assumption 4.1 there is a $C < \infty$ such that

$$\left| \frac{(\Phi_{n-1}(\Pi_{n-2}^N) - \Pi_{n-1}^N)(H_{n-1})}{\Pi_{n-1}^N(H_{n-1})\Phi_{n-1}(\Pi_{n-2}^N)(H_{n-1})} \right| \leq C .$$

Thus, by again applying the induction hypothesis (46) one has that

$$\mathbb{E}[T_{2,2}] \leq CF(\psi^2)$$

and this suffices to complete the proof. $\qquad\square$
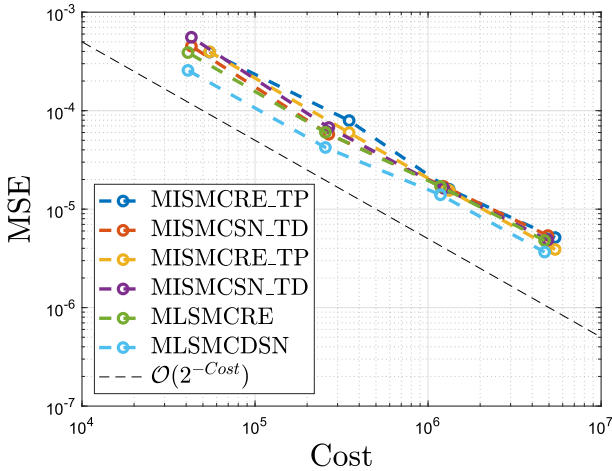
Fig. 2 2D Elliptic PDE with random diffusion coefficient rate of convergence for MLSMC and MISMC with the self-normalised increments estimator and the ratio estimator, where MISMC is applied with the tensor product index set and the total degree index set. Each MSE is computed with 200 realisations. Rates of regression: (1) MISMCSN_TP: -1.007 (2) MISMCSN_TD: -0.996 (3) MISMCRE_TP: -0.964 (4) MISMCRE_TD: -0.925 (5) MLSMCSN: -0.880 (6) MLSMCRE: -0.918

6 Numerical Results

First, we considered the toy example of a 1D DE simplification of the PDE introduced in Sect. 2.1. Since the method reduces to a multilevel method in this case, the results are provided in "Appendix C.1".

6.1 2D Elliptic PDE with Random Diffusion Coefficient

In this subsection, we look at the 2D elliptic PDE with random diffusion coefficient from Sect. 2.1. The problem is defined in (2)–(3). The domain of interest is $\Omega = [0, 1]^2$, the forcing term is $f = 100$, $a(x)(z) = 3 + x_1 \cos(3z_1) \sin(3z_2) + x_2 \cos(z_1) \sin(z_2)$, and the prior is $x \sim U[-1, 1]^2$. The observation operator and observation take the form of (7) and (8), respectively.

Let the observations be given at a set of four points - {(0.25,0.25), (0.25,0.75), (0.75,0.25), (0.75,0.75)}. Corresponding observations are generated by $y = u_\alpha(x^*) + \nu$, where $u_\alpha(x^*)$ is the approximate solution of the PDE at $\alpha = [10, 10]$ with $x^* = [-0.4836, -0.5806]$ drawn from $U[-1, 1]^2$, and $\nu \sim N(0, 0.5^2)$. Due to the zero Dirichlet boundary condition, the solution is zero when $\alpha_i = 0$ and $\alpha_i = 1$ for $i = 1, 2$. So we set $\alpha_i \leftarrow \alpha_i + 2$ for $i = 1, 2$ as the starting indices. The 2D PDE solver applied here is modified based on a MATLAB toolbox called IFISS [23] such that the solver can accept a random coefficient and solve the problem of interest. The algorithm is applied with Metropolis–Hastings method and a fixed tempering schedule for all α, where $J = 3$ and $\tau_j = (j - 1)/2$.
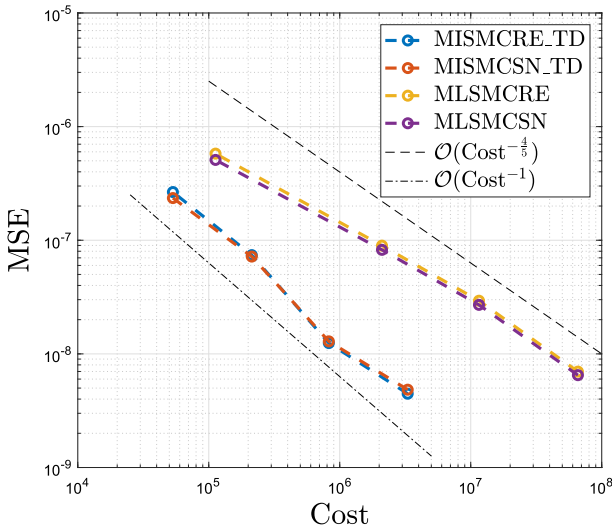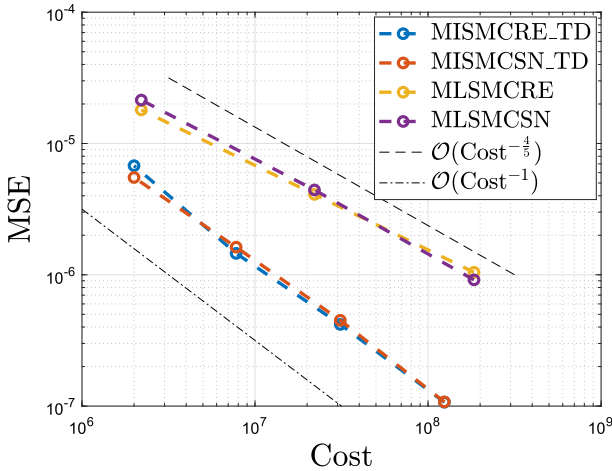
Fig. 3 LGC model with MLSMC and MISMC with the self-normalised increments estimator and the ratio estimator, where MISMC corresponds to the total degree index set. Each MSE is computed with 100 realisations. Rate of regression: (1) MISMCRE_TD: -1.022 (2) MISMCSN_TD: -0.973 (3) MLSMCRE: -0.686 (4) MLSMCSN: -0.677

For this example, we have $s_1 = s_2 = 2$ and $\beta_1 = \beta_2 = 4$ for the mixed rates corresponding to Assumptions 4.2, which implies that along the diagonal $\alpha_1 = \alpha_2$ the rates for ΔF_α are $s_1 + s_2 = 4$ and $\beta_1 + \beta_2 = 8$. This is shown in Figs. 7 and 8. The contour plot 9 performs a more general illustration. For the multilevel formulation, $s = 2$ and $\beta = 4$, which can be observed from Fig. 10.

Considering the quantity of interest $x_1^2 + x_2^2$, MSE in Fig. 2 is calculated with 200 realisations. Total computational costs are computed with the same idea as the previous questions. The reference solution is computed by MLSMC instead of MISMC to avoid errors in algorithm. MISMC algorithm is carried on with the two different index sets mentioned above—tensor product and total degree index set. According to the rates of regression in the caption of Fig. 2, both MLSMC and MISMC with the self-normalised increment estimator and the ratio estimator have the rate of convergence close to -1 falling into the canonical case, which is as expected.

The advantage of MISMC can be shown in the following examples, where we can only achieve subcanonical rates with MLSMC and MISMC with the tensor product index set but the canonical rate with MISMC with the total degree index set. It is worth to note that this advantage is because the total degree index set which abandons the most expensive estimation.

6.2 Log-Gaussian Process Models

After considering the PDE examples in previous subsections, we show the numerical results of the LGC model introduced in Sect. 2.2. The parameters are chosen as $\theta = (\theta_1, \theta_2, \theta_3) = (0, 1, 110.339)$. For this particular example, the increment rates

Fig. 4 LGP model with MLSMC and MISMC with the self-normalised increments estimator and the ratio estimator, where MISMC corresponds to the total degree index set. Each MSE is computed with 100 realisations. Rate of regression: (1) MISMCRE_TD: −0.994 (2) MISMCSN_TD: −0.950 (3) MLSMCRE: −0.643 (4) MLSMCSN: -712

associated with MLSMC are $s = 0.8$ and $\beta = 1.6$, while the mixed rates associated with MISMC are $s_i = 0.8$ and $\beta_i = 1.6$ for $i = 1, 2$. The rates for s and β can be observed from Fig. 14 and mixed rates for s_i and β_i for $i = 1, 2$ can be observed from Figs. 11, 12 and 13. This forward simulation method has a cost rate of $\gamma = 2 + \omega$, for any $\omega > 0$, while the traditional full factorisation method used in [32] (and references therein) has $\gamma = 6$. However, one has $\gamma_i = 1 + \omega < \beta_i < \gamma$. This means circulant embedding will deliver a single level complexity of approximately $\mathrm{MSE}^{-9/4}$, while the traditional grid-based approach has complexity $\mathrm{MSE}^{-19/4}$. An implementation of MLMC delivers $\mathrm{MSE}^{-5/4-\omega}$. Finally, MIMC with TD index set ($\delta_i = 0.5$ for $i = 1, 2$) delivers *canonical complexity of* MSE^{-1}. Note that, because $\sum_{j=1}^{2} \frac{\gamma_j}{s_j} = 5/2 > 2$, the important assumption $\sum_{j=1}^{2} \frac{\gamma_j}{s_j} \le 2$ for MISMC with TP index set is violated, the cost of the finest level samples dominates the total cost, and MISMC TP therefore has the same subcanonical complexity as MLMC.

SMC sampler is applied with the pre-conditioned Crank-Nicolson (pCN) MCMC [17, 50] as the mutation kernel and adaptive tempering described in Remark 4.1. The quantity of interest is taken as $\varphi(x) = \int_{[0,1]^2} \exp(x(z)) dz$ and $\alpha_i \leftarrow \alpha_i + 5$ for $i = 1, 2$ are the starting indices. Figure 3, and the rate of regression in the caption, show the above claims that MISMC TD is canonical with rate of convergence close to -1 and MLSMC is subcanonical. MISMC TP is not included here since the computational complexity of this method is the same as that of MLSMC for this example. The only difference between the two methods is the constant. Compared with MLSMC, MISMC TP has extra indices and two extra terms at all internal indices. This means that MISMC TP has a larger constant than MLSMC. MISMC TD turns the computational complexity from subcanonical, which all that is achievable with MISMC TP, MLSMC, and SMC, to canonical, indicating the benefits of MISMC TD.

6.3 Log-Gaussian Process Model

In this subsection, we consider the LGP model introduced in Sect. 2.2. By changing the likelihood and parameters accordingly as $\theta = (\theta_1, \theta_2, \theta_3) = (0, 1, 27.585)$, the LGP model follows the same analysis as the LGC model in the previous subsection and gives the same numerical results for regularity and complexity. More precisely, the increment rates associated with MLSMC are $s = 0.8$ and $\beta = 1.6$, and the mixed rates associated with MISMC are $s_i = 0.8$ and $\beta_i = 1.6$ for $i = 1, 2$, which are the same as LGC. Figure 18 shows the increment rates for s and β, and Figs. 15, 16 and 17 show the mixed rates for s_i and β_i for $i = 1, 2$. The rates corresponding to the computational costs of MLSMC and MISMC are $\gamma = 2 + \omega$ and $\gamma_i = 1 + \omega < \gamma$, for any $\omega > 0$, respectively. Being the same as that of LGC, the complexity of LGP is $\text{MSE}^{-5/4-\omega}$ with MLSMC and MSE^{-1} with MISMC.

As above, SMC sampler is applied with the pre-conditioned Crank-Nicolson (pCN) MCMC [17, 50] as the mutation kernel and adaptive tempering described in Remark 4.1. Considering the quantity of interest $\varphi(x) = \int_{[0,1]^2} \exp(x(z))dz$ and letting the starting indices $\alpha_i \leftarrow \alpha_i + 5$ for $i = 1, 2$, Fig. 4, and the rate of regression in the caption, show the same claims that MISMC TD is canonical with rate of convergence close to -1 and MLSMC is subcanonical.

A Proofs relating to Theorem 4.2 and Remark 4.3

Let $B_\alpha = F_\alpha(\psi_{\zeta_\alpha})$, and recall from Assumption 4.2 (V) that

$$\mathbb{E}\left[\left(F_\alpha^{N_\alpha}(\psi_{\zeta_\alpha}) - F_\alpha(\psi_{\zeta_\alpha})\right)^2\right] \leq V_\alpha/N_\alpha \,,$$

and from Assumption 4.2 (C) that the total computational cost is $\sum_{\alpha \in \mathcal{I}} N_\alpha C_\alpha$. Following from Theorem 4.1, we have

$$\mathbb{E}[(\widehat{\varphi}_{\mathcal{I}}^{\text{MI}} - \pi(\varphi))^2] \leq C \max_{\zeta \in \{\varphi, 1\}} \left(\sum_{\alpha \in \mathcal{I}} \frac{V_\alpha}{N_\alpha} + \left(\sum_{\alpha \notin \mathcal{I}} B_\alpha\right)^2\right). \qquad (47)$$

Then, $\mathbb{E}[(\widehat{\varphi}_{\mathcal{I}}^{\text{MI}} - \pi(\varphi))^2]$ is less than $C\varepsilon^2$ as long as both $\max_{\zeta \in \{\varphi, 1\}} \sum_{\alpha \in \mathcal{I}} V_\alpha$ and $\max_{\zeta \in \{\varphi, 1\}} \left(\sum_{\alpha \notin \mathcal{I}} B_\alpha\right)^2$ are of $\mathcal{O}(\varepsilon^2)$. We can now prove the Remark 4.3 as follows.

Proof We start from inequality (47) and replace the general index set $\mathcal{I}$ by the tensor product index set $\mathcal{I}_{L_1:L_D} := \{\alpha \in \mathbb{N}^d : \alpha_1 \in \{0, \ldots, L_1\}, \ldots, \alpha_D \in \{0, \ldots, L_D\}\}$. Let $L_i = \lceil \log_2(D/\varepsilon)/s_i \rceil$, for $i = 1, \ldots, D$, where $\lceil \cdot \rceil$ denotes ceiling a noninteger to an integer. The bias term is bounded as follows

$$\sum_{\alpha \notin \mathcal{I}_{L_1:L_D}} B_\alpha = \sum_{\alpha \notin \mathcal{I}_{L_1:L_D}} F_\alpha(\psi_{\zeta_\alpha}) \leq C \sum_{\alpha \notin \mathcal{I}_{L_1:L_D}} \prod_{i=1}^{D} 2^{-\alpha_i s_i} \leq C \sum_{i=1}^{D} 2^{-L_i s_i},$$

where the inequality above follows from Assumption 4.2(B). Substituting L_i in the inequality, the bias term is of $\mathcal{O}(\varepsilon)$. By Lemma A.4, $\sum_{\alpha \in \mathcal{I}_{L_1:L_D}} V_\alpha/N_\alpha$ is minimised and equals ε^2 by choosing

$$N_\alpha = \varepsilon^{-2} \left(\sum_{\alpha' \in \mathcal{I}_{L_1:L_D}} \sqrt{V_{\alpha'} C_{\alpha'}} \right) \sqrt{\frac{V_\alpha}{C_\alpha}}.$$

The sample size can only be treated as an integer and there should be at least one sample in each multi-index of resolution. So let the upper bound of N_α be

$$N_\alpha \leq 1 + \varepsilon^{-2} \left(\sum_{\alpha' \in \mathcal{I}_{L_1:L_D}} \sqrt{V_{\alpha'} C_{\alpha'}} \right) \sqrt{\frac{V_\alpha}{C_\alpha}}.$$

Then, the total computational cost $C_{\mathcal{I}_{L_1:L_D}}$ is given by

$$C_{\mathcal{I}_{L_1:L_D}} = \sum_{\alpha \in \mathcal{I}_{L_1:L_D}} N_\alpha C_\alpha = \mathcal{O}\left(\varepsilon^{-2} \left(\sum_{\alpha \in \mathcal{I}_{L_1:L_D}} \sqrt{V_\alpha C_\alpha} \right)^2 + \sum_{\alpha \in \mathcal{I}_{L_1:L_D}} C_\alpha \right).$$

By Assumption 4.2, we have

$$\sum_{\alpha \in \mathcal{I}_{L_1:L_D}} \sqrt{V_\alpha C_\alpha} \leq \sum_{\alpha \in \mathcal{I}_{L_1:L_D}} C \prod_{i=1}^{D} 2^{\alpha_i(\gamma_i - \beta_i)/2} = \left(\prod_{i=1}^{D} C \sum_{\alpha_i=0}^{L_i} 2^{\alpha_i(\gamma_i - \beta_i)/2} \right).$$

Since $\beta_i > \gamma_i$, $\sum_{\alpha \in \mathcal{I}_{L_1:L_D}} \sqrt{V_\alpha C_\alpha} = \mathcal{O}(1)$. In addition, $\sum_{\alpha \in \mathcal{I}_{L_1:L_D}} C_\alpha = \mathcal{O}(\varepsilon^{-\sum_{i=1}^{D} D\gamma_i/s_i})$ and this is bounded by $\mathcal{O}(\varepsilon^{-2})$ due to the assumption that $\sum_{i=1}^{D} \gamma_i/s_i \leq 2$. Thus, the total computational cost is dominated by $\mathcal{O}(\varepsilon^{-2})$. $\square$

The proof of Theorem 4.2 is similar as that of Remark 4.3. The details are as follows.

Proof We start from inequality (47) and replace the general index set with the total degree index set $\mathcal{I}_L := \{\alpha \in \mathbb{N}^d : \sum_{i=1}^{D} \delta_i \alpha_i \leq L, \sum_{i=1}^{D} \delta_i = 1\}$.

Let $L = \log\left(\varepsilon^{-1}(\log\varepsilon^{-1})^{2(n_1-1)}\right)/A_1$, where $A_1 = \min_{i=1,\dots,D}\log(2)\delta_i^{-1}s_i$ and $n_1 = \#\{i = 1,\dots,D : \log(2)\delta_i^{-1}s_i = A_1\}$. Using Lemma A.1, the bias term can be bounded as follows

$$
\begin{aligned}
\sum_{\alpha\notin\mathcal{I}_L} B_\alpha &= \sum_{\alpha\notin\mathcal{I}_L} F_\alpha(\psi_{\zeta_\alpha}) \\
&\leq C \sum_{\alpha\notin\mathcal{I}_L}\prod_{i=1}^{D} 2^{-\alpha_i s_i} \\
&\leq C \int_{\{\mathbf{x}\in\mathbb{R}_+^D:\ \sum_{i=1}^{D}\delta_i x_i > L,\ \sum_{i=1}^{D}\delta_i = 1\}}\prod_{i=1}^{D} e^{-\log(2)x_i s_i}\,d\mathbf{x} \\
&= C \int_{\{\mathbf{x}\in\mathbb{R}_+^D:\ \sum_{i=1}^{D} x_i > L\}} e^{-\sum_{i=1}^{D}\log(2)\delta_i^{-1}x_i s_i}\,d\mathbf{x} \\
&\leq C e^{-A_1 L} L^{n_1-1}
\end{aligned}
$$

where A_1 and n_1 are defined above. Substituting in the L and applying Lemma A.3, the bias term is of $\mathcal{O}(\varepsilon)$.

Following the similar steps as the proof for Remark 4.3 and replacing the tensor product index set with the total degree index set, the total computational cost $C_{\mathcal{I}_L}$ can be formulated as

$$
C_{\mathcal{I}_L} = \sum_{\alpha\in\mathcal{I}_L} N_\alpha C_\alpha = \mathcal{O}\left(\varepsilon^{-2}\left(\sum_{\alpha\in\mathcal{I}_L}\sqrt{V_\alpha C_\alpha}\right)^2 + \sum_{\alpha\in\mathcal{I}_L} C_\alpha\right).
$$

Starting from the first term, since $\beta_i > \gamma_i$, we have

$$
\begin{aligned}
\sum_{\alpha\in\mathcal{I}_L}\sqrt{V_\alpha C_\alpha} &\leq \sum_{\alpha\in\mathcal{I}_L} 2^{\sum_{i=1}^{D}\alpha_i(\gamma_i-\beta_i)/2} \\
&\leq \frac{1}{\prod_{i=1}^{D}(1 - 2^{(\gamma_i-\beta_i)/2})}.
\end{aligned}
$$

Considering the second term $\sum_{\alpha\in\mathcal{I}_L} C_\alpha$ and using Lemma A.2, we have

$$
\begin{aligned}
\sum_{\alpha\in\mathcal{I}_L} C_\alpha &= \sum_{\alpha\in\mathcal{I}_L} 2^{\sum_{i=1}^{D}\alpha_i\gamma_i} \\
&\leq C \int_{\{\mathbf{x}\in\mathbb{R}_+^D:\ \sum_{i=1}^{D} x_i \leq L\}} e^{\sum_{i=1}^{D}\log(2)\delta_i^{-1}\gamma_i x_i}\,d\mathbf{x} \\
&\leq C e^{A_2 L} L^{n_2-1},
\end{aligned}
$$

where $A_2 = \max_{i=1,\dots,D}\log(2)\delta_i^{-1}\gamma_i$ and $n_2 = \#\{i = 1,\dots,D : \log(2)\delta_i^{-1}\gamma_i = A_2\}$.

Substituting L into the upper bound, and since $2s_i \geq \beta_i > \gamma_i$, we have $\gamma_i/s_i \leq 2$ which gives $\sum_{\alpha \in \mathcal{I}_L} C_\alpha \leq \mathcal{O}(\varepsilon^{-2})$. Then, the summation of the two terms is of $\mathcal{O}(\varepsilon^{-2})$. $\qquad \square$

Lemmas A.1 and A.2 are from Lemma 6.3 and 6.2 of [29].

Lemma A.1 *For $L \geq 1$ and $\mathbf{a} \in \mathbb{R}_+^D$, there exists a $C(\mathbf{a}) > 0$ such that the following inequality holds*

$$\int_{\{\mathbf{x} \in \mathbb{R}_+^D: \, \sum_{i=1}^D x_i > L\}} e^{-\sum_{i=1}^D a_i x_i} \, d\mathbf{x} \leq C e^{-A_1 L} L^{n_1 - 1},$$

where

$$A_1 = \min_{i=1,\dots,D} a_i, \quad n_1 = \#\{i = 1, \dots, D : a_i = A_1\}.$$

Lemma A.2 *For $L \geq 1$ and $\mathbf{a} \in \mathbb{R}_+^D$, there exists a $C(\mathbf{a}) > 0$ such that the following inequality holds*

$$\int_{\{\mathbf{x} \in \mathbb{R}_+^D: \, \sum_{i=1}^D x_i \leq L\}} e^{\sum_{i=1}^D a_i x_i} \, d\mathbf{x} \leq C e^{A_2 L} L^{n_2 - 1},$$

where

$$A_2 = \max_{i=1,\dots,D} a_i, \quad n_2 = \#\{i = 1, \dots, D : a_i = A_2\}.$$

Lemma A.3 *For $L = \log\left(\epsilon^{-1}(\log \epsilon^{-1})^{2(n-1)}\right)/A$*

$$e^{-AL} L^{n-1} \leq C\epsilon$$

where $C = (2(n-1)/A)^{n-1}$.

Proof The argument follows by the following sequence of equalities. The final inequality, below, follows since $[\log x - x] \leq 0$ for $x = \log \log \epsilon^{-1}$.

$$\log\left(e^{-AL} L^{n-1}\right)$$

$$= -AL + (n-1) \log L$$

$$= -\log \epsilon^{-1} - 2(n-1) \log \log \epsilon^{-1} + (n-1) \log \left(\frac{1}{A} \log\left(\epsilon^{-1}(\log \epsilon^{-1})^{2(n-1)}\right)\right)$$

$$= \log \epsilon - 2(n-1) \log \log \epsilon^{-1} + (n-1) \log \frac{1}{A} + (n-1) \log \log \epsilon^{-1}$$

$$\quad + (n-1) \log \left(2(n-1) \log \log \epsilon^{-1}\right)$$

$$= \log \epsilon + (n-1) \log(2(n-1)/A) + (n-1)\left[\log \log \log \epsilon^{-1} - \log \log \epsilon^{-1}\right]$$

$$\leq \log(\epsilon (2(n-1)/A)^{n-1}),$$

as required. □

Lemma A.4 *For a fixed* $\varepsilon^2 = \sum_{\alpha \in \mathcal{I}} \mathbb{E}\left[\left(F_\alpha^{N_\alpha}(\psi_{\zeta_\alpha}) - F_\alpha(\psi_{\zeta_\alpha})\right)^2\right] = \sum_{\alpha \in \mathcal{I}} \frac{V_\alpha}{N_\alpha}$
(from Lemma 4.1), the cost is minimised by choosing N_α such that

$$N_\alpha = \varepsilon^{-2} \left(\sum_{\alpha' \in \mathcal{I}} \sqrt{V_{\alpha'} C_{\alpha'}}\right) \sqrt{\frac{V_\alpha}{C_\alpha}}.$$

Proof Given fixed $\varepsilon^2 = \sum_{\alpha \in \mathcal{I}} \frac{V_\alpha}{N_\alpha}$, the cost can be minimised as a function of $\{N_\alpha\}_{\alpha \in \mathcal{I}}$ by applying the Lagrange multiplier method. For some Lagrange multiplier λ, we solve the minimisation problem

$$\min_{N_\alpha} \sum_{\alpha \in \mathcal{I}} N_\alpha C_\alpha + \lambda^2 \left(\sum_{\alpha \in \mathcal{I}} \frac{V_\alpha}{N_\alpha} - \varepsilon^2\right).$$

This gives the optimal value of $N_\alpha = \lambda \sqrt{V_\alpha / C_\alpha}$ for each $\alpha \in \mathcal{I}$. Plugging the solution to $\{N_\alpha\}_{\alpha \in \mathcal{I}}$ into the constraint equation $\varepsilon^2 = \sum_{\alpha \in \mathcal{I}} \frac{V_\alpha}{N_\alpha}$ gives $\lambda = \varepsilon^{-2} \sum_{\alpha \in \mathcal{I}} \sqrt{V_\alpha C_\alpha}$ and therefore

$$N_\alpha = \varepsilon^{-2} \left(\sum_{\alpha' \in \mathcal{I}} \sqrt{V_{\alpha'} C_{\alpha'}}\right) \sqrt{\frac{V_\alpha}{C_\alpha}}.$$

□

B LGC Results

We restate and prove Proposition 2.3.

Proposition B.1 *Let $x \sim \pi_0$, where π_0 is a Gaussian process of the form (10) with spectral decay corresponding to (11), and let x_α correspond to truncation on the index set $\mathcal{A}_\alpha$ as in (14). Then $x \in H_q^m$ for all $q < \beta/2$, and for $r \in [0, q)$ there is a $C > 0$ such that*

$$\|x_\alpha - x\|_r^2 \leq C \|x\|_q^2 2^{-2(q-r) \min_i \alpha_i}.$$

Proof Since $x \sim \pi_0$ is a Gaussian process, in order to prove $x \in H_q^m$ for all $q < \beta/2$, it suffices to prove that $\mathbb{E}\|x\|_q^2 < \infty$. Indeed there is a $C > 0$ such that

$$\mathbb{E}\|x\|_q^2 \leq C \sum_{k \in \mathbb{Z}^2} ((1 + k_1^2)(1 + k_2^2))^{q - \frac{\beta+1}{2}},$$

from which it is clear that $2q < \beta$ provides a sufficient condition for summability. To see this, define $x_k = \langle \phi_k, x \rangle \equiv \int \phi_k(z) x^*(z) dz$, and note that $\mathbb{E}|x_k|^2 = \rho_k(\theta)$ and $\{\phi_k\}_{k \in \mathbb{Z}}$ are orthonormal. In more detail,

$$
\begin{aligned}
\mathbb{E}[\|x\|_q^2] &= \mathbb{E}\left[\|A^{q/2}x\|^2\right] \\
&= \mathbb{E}\left\|\sum_{k \in \mathbb{Z}} a_k^{q/2} \phi_k \underbrace{\langle \phi_k, x \rangle}_{x_k}\right\|^2 \\
&= \sum_{k,k' \in \mathbb{Z}^2} a_k^{q/2} a_{k'}^{q/2} \underbrace{\langle \phi_k, \phi_{k'} \rangle}_{\delta_{k,k'}} \mathbb{E} x_k x_{k'} \\
&= \sum_{k \in \mathbb{Z}^2} (1+k_1^2)^q (1+k_2^2)^q \mathbb{E}|x_k|^2 \\
&= \sum_{k \in \mathbb{Z}^2} (1+k_1^2)^q (1+k_2^2)^q \rho_k(\theta)^2 \\
&\leq C \sum_{k \in \mathbb{Z}^2} (1+k_1^2)^q (1+k_2^2)^q \frac{1}{((1+k_1^2)(1+k_2^2))^{\frac{\beta+1}{2}}} \\
&= C \sum_{k \in \mathbb{Z}^2} ((1+k_1^2)(1+k_2^2))^{q-\frac{\beta+1}{2}}
\end{aligned}
$$

Now let $P_{\mathcal{A}_\alpha}$ denote the projection onto the index set $\mathcal{A}_\alpha$. Observe that there is a $C > 0$ such that

$$
\|A^{-q/2} - A^{-q/2} P_{\mathcal{A}_\alpha}\|_{\mathcal{L}(L^2,L^2)}^2 = \sup_{\|x\|=1} \sum_{k \notin \mathcal{A}_\alpha} a_k^{-q} x_k^2
$$
$$
\leq C(2^{-2q\alpha_1} + 2^{-2q\alpha_2}), \tag{48}
$$

where $\mathcal{L}(L^2, L^2)$ denotes the space of linear operators from L^2 to L^2.

For $r \in [0, q)$, we have

$$
\begin{aligned}
\|x_\alpha - x\|_r^2 &= \|A^{-q/2} A^{(q+r)/2}(x - P_{\mathcal{A}_\alpha}x)\|^2 \\
&\leq \|(A^{-(q-r)/2} - A^{-(q-r)/2} P_{\mathcal{A}_\alpha}) A^{q/2}x\|^2 \\
&\leq C \|x\|_q^2 2^{-2(q-r)\min_i \alpha_i},
\end{aligned}
$$

where the first line follows from the definition, the second follows from commutativity of $P_{\mathcal{A}_\alpha}$ and A, and the final line follows from the definition of operator norm and (48). □

We restate and prove Proposition 2.4.

Proposition B.2 *Given $x : \Omega \to \mathbb{R}$ is a Gaussian process, with probability measure denoted π_0, defined on compact finite dimensional space Ω, that is almost surely*

continuous and has a finite mean and covariance. If we define π by

$$\text{(LGC)} \quad \frac{d\pi}{d\pi_0}(x) \propto \exp\left[\sum_{j=1}^{n} x(z_j) - \int_{\Omega} \exp(x(z))dz\right],$$

$$\text{(LGP)} \quad \frac{d\pi}{d\pi_0}(x) \propto \exp\left[\sum_{j=1}^{n} x(z_j) - n\log\int_{\Omega} \exp(x(z))dz\right].$$

for $n \in \mathbb{N}$ then $\pi(dx)$ is a well-defined probability measure, and can be represented in terms of its density with respect to π_0:

$$\pi(dx) = \frac{d\pi}{d\pi_0}\pi_0(dx).$$

We first proof the proposition for LGC.

Proof For π with LGC to be well defined, the first exponential above must be integrable. Specifically, we require that

$$0 < Z := \mathbb{E}_{\pi_0}\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \int_{\Omega} \exp(x(s))ds\right\}\right] < \infty.$$

To upper-bound Z notice that $\sum_{j=1}^{n} x(s_j)$ is a real-valued Gaussian random variable with finite mean and variance, which we denote by μ and σ^2, respectively. Also, $\exp(x(s))$ is non-negative thus

$$
\begin{aligned}
Z := \mathbb{E}_{\pi_0}&\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \int_{\Omega} \exp(x(s))ds\right\}\right] \\
&\leq \mathbb{E}_{\pi_0}\left[\exp\left\{\sum_{j=1}^{n} x(s_j)\right\}\right] = \mathbb{E}[e^{\mu+\sigma^2/2}] < \infty.
\end{aligned}
$$

This gives the required upper-bound.

For the lower-bound, we note that since $x(s)$ is almost surely continuous and Ω is compact, then $\sup_{s\in\Omega} x(s)$ is almost surely finite. Thus $\int_{\Omega} \exp(x(s))ds$ is almost surely finite, because $\int_{\Omega} \exp(x(s))ds < |\Omega| \exp(\sup_{s\in\Omega} x(s))$. Thus, by Monotone convergence, there exists a value of K_1 such that

$$\mathbb{P}\left(\int_{\Omega} \exp(x(s))ds > K_1\right) \leq 1/4$$

Similarly since $\sum_{i=1}^{n} x_i(s)$ is Gaussian, there exists a value of K_2 such that

$$\mathbb{P}\left(\sum_{j=1}^{n} x_j(s) \leq -K_2\right) \leq 1/4$$

Taking, $K = 2(K_1 \vee K_2)$

$$\mathbb{P}\left(\sum_{j=1}^{n} x(s_j) - \int_{\Omega} \exp(x(s))ds > -K\right)$$

$$\geq \mathbb{P}\left(\sum_{j=1}^{n} x(s_j) > -K/2, -\int_{[0,1]^2} \exp(x(s))ds > -K/2\right)$$

$$\geq 1 - \mathbb{P}\left(\sum_{j=1}^{n} x(s_j) \leq -K/2\right) - \mathbb{P}\left(K/2 \leq \int_{\Omega} \exp(x(s))ds\right)$$

$$\geq 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}.$$

Thus Markov's inequality for corresponding value of K gives

$$Z := \mathbb{E}_{\pi_0}\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \int_{\Omega} \exp(x(s))ds\right\}\right] \geq \frac{1}{2}e^{-K} > 0,$$

as required. □

We now prove the proposition for LGP.

Proof For π with LGP to be well defined, the second exponential above must be integrable. Specifically, we require that

$$0 < Z := \mathbb{E}_{\pi_0}\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - n \log\left(\int_{\Omega} \exp(x(s))ds\right)\right\}\right] < \infty.$$

First, we show Z is lower bounded by 0. Notice that since the process $x(s)$ is almost surely continuous in s and the domain Ω is compact. Thus almost surely it holds that

$$\tilde{Z} := \exp\left\{\sum_{j=1}^{n} x(s_j) - n \log\left(\int_{\Omega} \exp(x(s))ds\right)\right\} > 0$$

As an immediate consequence $Z = \mathbb{E}_{\pi_0}[\tilde{Z}] > 0$, as required.

We now show Z is upper-bounded. Notice that, if we let $u = (u_i : i = 1, .., n)$ be independent uniformly distributed on Ω (wlog we assume $|\Omega| = 1$) then we can bound as follows

$$
\begin{aligned}
\tilde{Z} &= \exp\left\{\sum_{j=1}^{n} x(s_j)\right\} \left(\int_{\Omega} \exp(x(s))ds\right)^{-n} = \exp\left\{\sum_{j=1}^{n} x(s_j)\right\} \left(\mathbb{E}_u\left[e^{\sum_{i=1}^{n} x(u_i)}\right]\right)^{-1} \\
&\leq \exp\left\{\sum_{j=1}^{n} x(s_j)\right\} \mathbb{E}_u\left[e^{-\sum_{i=1}^{n} x(u_i)}\right] = \mathbb{E}_u\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \sum_{i=1}^{n} x(u_i)\right\}\right]
\end{aligned}
\tag{49}
$$

The inequality above applies Jensen's Inequality. Thus, we see we can bound Z by bounding the expectation of $\exp\{\sum_{j=1}^{n} x(s_j) - \sum_{i=1}^{n} x(u_i)\}$ with respect to π_0 and u. By Fubini's Theorem [62], we can hold u fixed and take the expectation with respect to π_0. Notice that under π_0, x is a Gaussian process with bounded mean and covariance. We let μ and σ bound the mean and covariance of x. So, conditional on u, it holds that

$$
\mathbb{E}_{\pi_0}\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \sum_{i=1}^{n} x(u_i)\right\}\right] \leq e^{n\mu + n\sigma^2/2}.
$$

Since the above upper bound is independent of u, we have that

$$
\mathbb{E}_{\pi_0}\left[\mathbb{E}_u\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \sum_{i=1}^{n} x(u_i)\right\}\right]\right] \leq e^{n\mu + n\sigma^2/2}.
\tag{50}
$$

Note that Fubini for positive random variables allows interchanging integrals without a priori finite guarantees [62]. Combining inequalities (49) and (50) we have

$$
Z = \mathbb{E}_{\pi_0}[\tilde{Z}] \leq \mathbb{E}_{\pi_0}\left[\mathbb{E}_u\left[\exp\left\{\sum_{j=1}^{n} x(s_j) - \sum_{i=1}^{n} x(u_i)\right\}\right]\right] \leq e^{n\mu + n\sigma^2/2} < \infty,
$$

which gives the required upper bound on Z. $\qquad\square$

We now restate and prove Proposition 2.5.

Proposition B.3 *For both LGP and LGC, there is a $C > 0$ such that for $x \sim \pi_0$ and $x_\alpha = P_{\mathcal{A}_\alpha} x$, where $P_{\mathcal{A}_\alpha}$ denotes the projection onto the index set $\mathcal{A}_\alpha$ defined in (14), the following rate estimate holds for all $q < (\beta - 1)/2$*

$$
\mathbb{E}|L_\alpha(x_\alpha) - L(x)|^2 \leq C 2^{-2\min\{q,1\}\min\{\alpha_1, \alpha_2\}}.
$$

Proof The result is proven for the more difficult case of LGC. The LGP case follows similarly. Define

$$\Phi(x) := \sum_{j=1}^{n} x(z_j) - \int_{\Omega} \exp(x(z)) dz \tag{51}$$

$$\Phi_\alpha(x_\alpha) := \sum_{j=1}^{n} \hat{x}_\alpha(z_j) - Q(\exp(x_\alpha)), \tag{52}$$

where we recall the definition of $\hat{x}_\alpha$ above (15). Now $L(x) = \exp(\Phi(x))$ and $L_\alpha(x_\alpha) = \exp(\Phi_\alpha(x_\alpha))$. Note that $\pi_0 = N(0, \mathcal{C})$ and that $\mathcal{C}$ has the kernel representation

$$\mathcal{C}(z, z') = \sum_{k_1} \rho_{k_1}^2 \phi_{k_1}(z_1) \phi_{k_1}(z_1') \sum_{k_2} \rho_{k_2}^2 \phi_{k_2}(z_2) \phi_{k_2}(z_2') =: \mathcal{C}_1(z_1, z_1') \mathcal{C}_2(z_2, z_2').$$

This means that the dependence between z_1 and z_2 is only statistical, via ξ_k in (10), and therefore a given realisation $x \sim \pi_0$ admits factorisation

$$x(z) = \sum_{k_1} \rho_{k_1} \phi_{k_1}(z_1) x_{2,k_1}(z_2),$$

where, for each k_1, the i.i.d. random variables

$$x_{2,k_1}(z_2) = \sum_{k_2} \xi_{k_1,k_2} \rho_{k_2} \phi_{k_2}(z_2)$$

have the properties of $x_2' \sim N(0, \mathcal{C}_2)$, i.e. $x_{2,k_1} \in H_{\beta/2}$ and the one-dimensional Sobolev embedding theorem (see, e.g. [59]) provides $\|x_{2,k_1}\|_{L^\infty(\Omega)} \leq C \|x_{2,k_1}\|_r$, for $r \in [1/2, \beta/2]$. Furthermore, letting $\hat{x}_1(z_1) := \sum_{k_1} \rho_{k_1} \|x_{2,k_1}\|_r \phi_{k_1}(z_1)$, it is clear that $\|\hat{x}_1\|_r = \|x\|_r < \infty$ for $r < \beta/2$. Hence, applying Sobolev embedding theorem again on $\hat{x}_1(z_1)$, we have $\|\hat{x}_1\|_{L^\infty(\Omega)} \leq C \|\hat{x}_1\|_r$, for $r \in [1/2, \beta/2]$. Now, since $\|\phi_{k_1}\|_{L^\infty(\Omega)} = 1$ for all k_1 and $\rho_{k_1} \|x_{2,k_1}\|_r \geq 0$,

$$\|\hat{x}_1\|_{L^\infty(\Omega)} = \sup_{z_1} \sum_{k_1} \rho_{k_1} \|x_{2,k_1}\|_r |\phi_{k_1}(z_1)| \tag{53}$$

$$\geq C \sup_{z_1} \sum_{k_1} \rho_{k_1} \|x_{2,k_1}\|_{L^\infty(\Omega)} |\phi_{k_1}(z_1)| \tag{54}$$

$$\geq C \sup_{z_1,z_2} \left| \sum_{k_1} \sum_{k_2} \rho_{k_1} \phi_{k_1}(z_1) \xi_{k_1,k_2} \rho_{k_2} \phi_{k_2}(z_2) \right| \tag{55}$$

$$= C \|x\|_{L^\infty(\Omega)}. \tag{56}$$

Sobolev embedding theorem is used on x_{2,k_1} in the second line, and definitions are used in the third and fourth lines. In conclusion, $\|x\|_{L^\infty(\Omega)} \leq C \|x\|_r$, for $r \geq 1/2$. $\qquad \square$

We have that

$$\sum_{j=1}^{n} x(z_j) \leq n \sup_{z \in \Omega} x(z) \leq C \|x\|_r . \tag{57}$$

Therefore,

$$\Phi(x), \Phi_\alpha(x_\alpha) \leq \|x\|_r . \tag{58}$$

Furthermore, observe that this implies that for all $\epsilon > 0$

$$\Phi(x), \Phi_\alpha(x_\alpha) \leq \epsilon \|x\|_r^2 + \epsilon^{-1} . \tag{59}$$

To see this consider the cases $\|x\|_r > \epsilon^{-1}$ and $\|x\|_r \leq \epsilon^{-1}$ separately.

By the mean value theorem,

$$\mathbb{E}[|L_\alpha(x_\alpha) - L(x)|^2] = \int_{H_{\beta/2}^m} (\exp(\Phi(x)) - \exp(\Phi_\alpha(x_\alpha))^2 \, d\pi_0 \tag{60}$$

$$\leq \int_{H_{\beta/2}^m} \exp\left(\max\{\Phi(x), \Phi_\alpha(x_\alpha)\}\right) |\Phi(x) - \Phi_\alpha(x_\alpha)|^2 \, d\pi_0 . \tag{61}$$

Note that for the exponential term in the integral

$$\exp(\max\{\Phi(x), \Phi_\alpha(x_\alpha)\}) \leq \exp(\Phi(x)) + \exp(\Phi_\alpha(x_\alpha)) \tag{62}$$

$$\leq C(\epsilon) \exp(\epsilon \|x\|_r^2) \tag{63}$$

and for the squared term

$$|\Phi(x) - \Phi_\alpha(x_\alpha)|^2 = \left| \sum_{j=1}^{n} x(z_j) - \int_\Omega \exp(x(z)) dz - \sum_{j=1}^{n} \hat{x}_\alpha(z_j) + Q(\exp(x_\alpha)) \right|^2 \tag{64}$$

$$\leq 2 \left| \sum_{j=1}^{n} x(z_j) - \sum_{j=1}^{n} \hat{x}_\alpha(z_j) \right|^2 + 2 \left| \int_\Omega \exp(x(z)) dz - Q(\exp(x_\alpha)) dz \right|^2 \tag{65}$$

The bound (63) is clear following (59). Now consider the bound of $|\Phi(x) - \Phi_\alpha(x_\alpha)|^2$. For the first term of (65), let $\hat{x}_\alpha$ correspond to piecewise linear interpolation for simplicity. As in (57) we have

$$\left| \sum_{j=1}^{n} (x(z_j) - \hat{x}_\alpha(z_j)) \right| \leq C \|x - \hat{x}_\alpha\|_r .$$

Given standard piecewise linear approximation estimates which lead to Proposition 2.1, and weaker versions such as [25]

$$\|x_\alpha - \hat{x}_\alpha\| \leq 2^{-\min\{\alpha_1,\alpha_2\}} \|\nabla x_\alpha\| \leq C 2^{-\min\{\alpha_1,\alpha_2\}} \|x_\alpha\|_1 \,,$$

it is natural to assume the following generalisation, for $p > 0$ and $r + p \leq 2$,

$$\|x_\alpha - \hat{x}_\alpha\|_r \leq 2^{-p\min\{\alpha_1,\alpha_2\}} \|x_\alpha\|_{r+p} \,. \tag{66}$$

For $p = (\beta - 1)/2$, and $r = 1/2$, the bound is

$$\|x - \hat{x}_\alpha\|_r \leq \|x - x_\alpha\|_r + \|x_\alpha - \hat{x}_\alpha\|_r \leq \|x - x_\alpha\|_r + 2^{-\frac{(\beta-1)}{2}\min\{\alpha_1,\alpha_2\}} \|x_\alpha\|_{\beta/2} \,. \tag{67}$$

Recall that, by 2.3, $x \sim \pi_0$ implies that $x \in H^m_{\beta/2}$ a.s. and hence $x_\alpha \in H^m_{\beta/2}$ a.s.

Now consider the second term of (65). For the sake of concreteness, we use the trapezoidal quadrature rule so that $Q(\exp(x_\alpha)) = 2^{-(\alpha_1+\alpha_2)} \sum_{h \in \prod_{i=1}^2 \{0,2^{-\alpha_i},...,1\}} w_h \exp(x_\alpha(h))$ (where $w_h = (1/2)^I$ and $I = \#\{i; h_i \in \{0,1\}\}$, i.e. the boundary terms are down-weighted, by $1/2$ on edges and $1/4$ at corners). Now

$$\int_\Omega \exp(x(z))dz - Q(\exp(x_\alpha)) = \int_\Omega (\exp(x(z)) - \exp(x_\alpha(z)))dz$$
$$+ \int_\Omega \exp(x_\alpha(z))dz - Q(\exp(x_\alpha)) \,. \tag{68}$$

For the first term, we have

$$\int_\Omega (\exp(x(z)) - \exp(x_\alpha(z)))dz \leq \|\exp(\max\{x(z), x_\alpha(z)\})\| \|x - x_\alpha\|$$
$$\leq C \exp(\|x\|_r) \|x - x_\alpha\| \,, \tag{69}$$

where the first line follows from the mean value theorem and Cauchy Schwartz inequality, while in the second line, the first factor follows from the inequality $\|x\|_{L^\infty(\Omega)} \leq \|x\|_r$, and the fact that $|\Omega| < \infty$. Note that we are restricted to the $\|\cdot\|_r$ estimate as a result of the pointwise observations, as in (58), but $\|x\| \leq \|x\|_r$ so (69) is suitable. For the second term of (68), since the trapezoidal rule for $D = 2$ follows from iterating the $D = 1$ rules, Theorem 1.8 of [18], along with similar manipulations as above, implies

$$\int_\Omega \exp(x_\alpha(z))dz - Q(\exp(x_\alpha)) \leq C 2^{-\min\{\alpha_1,\alpha_2\}} \exp(\|x\|_r) \|x_\alpha\|_1 \,. \tag{70}$$

Following from Proposition 2.3, we require $\beta \geq 2$ so that $x \in H^m_1$ a.s. and the constant in the second term (70) is controlled. Combining (69) and (70) in (68), we have

$$\int_\Omega \exp(x(z))dz - Q(\exp(x_\alpha)) \leq C \exp(\|x\|_r)\|x - x_\alpha\|$$

$$+ C2^{-\min\{\alpha_1,\alpha_2\}} \exp(\|x\|_r)\|x_\alpha\|_1 . \quad (71)$$

Now let $r = 1/2 + \delta$ for $\delta > 0$ arbitrarily small. Then for $q \in (0, (\beta - 1)/2)$, plugging (67) and (71) into (65) and using the same argument leading to (59) and then Proposition 2.3, we have

$$|\Phi(x) - \Phi_\alpha(x_\alpha)|^2 \leq C(\epsilon) \exp(2\epsilon\|x\|_{\beta/2}^2) 2^{-2\min\{q,1\}\min\{\alpha_1,\alpha_2\}} . \quad (72)$$

We note that our interest here is in rough priors with $q \leq 1$. In case $q > 1$, one would employ higher-order interpolation and quadrature such that these errors do not limit the rate of convergence.

Finally,

$$\mathbb{E}[|L_\alpha(x_\alpha) - L(x)|^2] \leq C(\epsilon) \int_{H_{\beta/2}^m} \exp(3\epsilon\|x\|_{\beta/2}^2) d\pi_0 2^{-2\min\{q,1\}\min\{\alpha_1,\alpha_2\}}$$

$$\leq C2^{-2\min\{q,1\}\min\{\alpha_1,\alpha_2\}}$$

The first inequality is by substituting (63) and (72) in (61) and applying $\|x\|_r^2 \leq \|x\|_{\beta/2}^2$. We note that the first inequality holds for all $\epsilon > 0$. Fernique theorem (e.g. Theorem 6.9 of [59]) guarantees that $\pi_0(\exp(3\epsilon\|x\|_{\beta/2}^2)) < \infty$ for some $\epsilon > 0$, and allows us to conclude with the second line. □

C Additional Numerical Results

C.1 1D Toy Example

We consider a 1D Toy Example first, whose likelihood is analytically tractable. This example is taken from [41]. Note that the multi-index methods are the same as multilevel methods in 1D. Considering the PDE (2)–(3) with $D = 1$, let $\Omega = [0, 1]$, $a = 1$, and the forcing term be $f = x$, where x is a random input with a uniform prior such that $x \sim U[-1, 1]$. This differential equation can be solved analytically as $u(x) = -0.5x(z^2 - z)$. Assume the observation operator as (7) and the observation taking the form as (8). The pointwise observations are well defined in 1D with $x \in L^2(\Omega)$. We take the observations at ten points in the interval $(0,1)$ with a step size $1/10$. Let $\Xi = 0.2$. Observations are generated by $y = -0.5x^*(z^2 - z) + \nu$, where $y = [y_1, \ldots, y_{10}]$, $z = [z_1, \ldots, z_{10}]$, $x^* = 0.2581$ drawn from $U[-1, 1]$ and $\nu \sim N(0, 0.2^2)$.

For this example, the quantity of interest used is x^2. By applying the FEM and discretising the differential equation with the step size $h_l = 2^{-l-1}$, we have $s = 2$, $\beta = 4$. This is shown in Fig 6a and b. The value of γ is 1 because we use a linear nodal basis function for FEM and tridiagonal solver. The algorithm is applied with Metropolis–Hastings method and a fixed tempering schedule for all α, where $J = 3$.

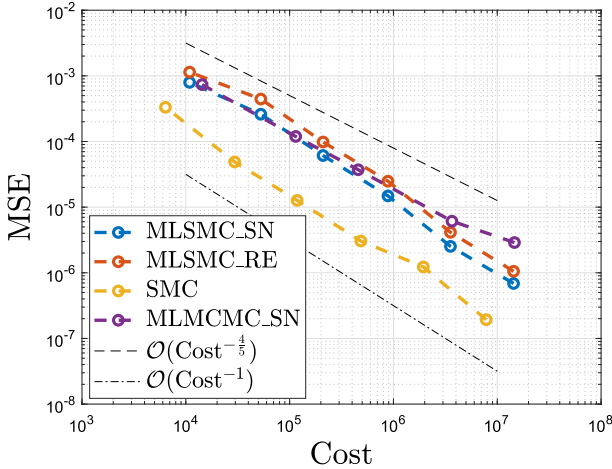**Fig. 5** 1D toy example, SMC, MLMCMC and MISMC rate of convergence by MSE vs Cost. Rates of regression: (1) MLSMC_SN: -1.011 (2) MLSMC_RE: -1.005 (3) SMC: -0.753 (4) MLMCMC_SN: -1.005
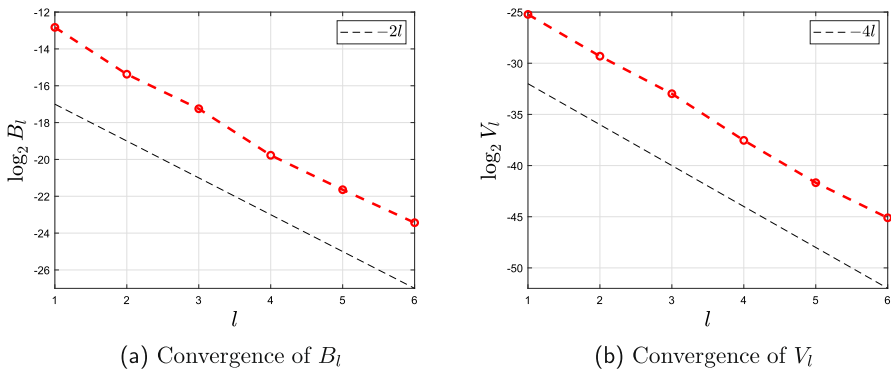

Fig. 6 1D toy example convergence rates. B_l is computed with 100 realisations and shown in panel 6a along with a line corresponding to $s = 2$. V_l is computed with 100 realisations and shown in panel 6a along with a line corresponding to $\beta = 4$

The MSE shown in Fig. 5 is calculated with 100 realisations, where the reference solution can be worked out as in [41]. The total computational cost is of $\mathcal{O}(\sum_{l=0}^{L} N_l C_l)$.

For comparison, single-level SMC, MLMCMC and MLSMC with the self-normalised increment estimator are applied in this example. It is difficult to observe the approximate rates from the plot directly, so we fit the rates and demonstrate those in the caption. The rate of convergence of single-level SMC is close to -4/5. The rate of convergence of MLMCMC with the self-normalised increment estimator, MLSMC with the self-normalised increment estimator and our MLSMC with ratio estimator are all approximately -1, which is the canonical complexity and better in terms of rate of convergence than the single-level methods as expected. The difference of performance between MLMCMC and MLSMC with either of the two estimators is only up

to a constant. MLMCMC has a smaller constant here, presumably as a consequence of the simplicity of the problem and the tuning of MLSMC. Our MLSMC with the ratio estimator appears to have a slightly larger constant, while the theoretical results remain its advantage.

C.2 2D Elliptic PDE with Random Diffusion Coefficient

See Figs. 7, 8, 9 and 10.
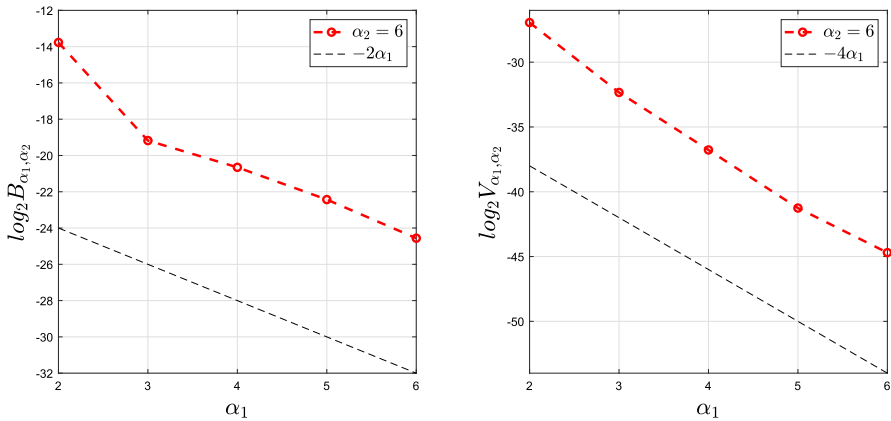
Fig. 7 2D Elliptic PDE with random diffusion coefficient. Verification of MISMC rates as in Assumption 4.2 for α_1 given $\alpha_2 = 7$, computed with 20 realisations and 1000 samples for each realisation. Left: s_1. Right: β_1. The same result holds for an $\alpha_1 = 7$ (not shown)
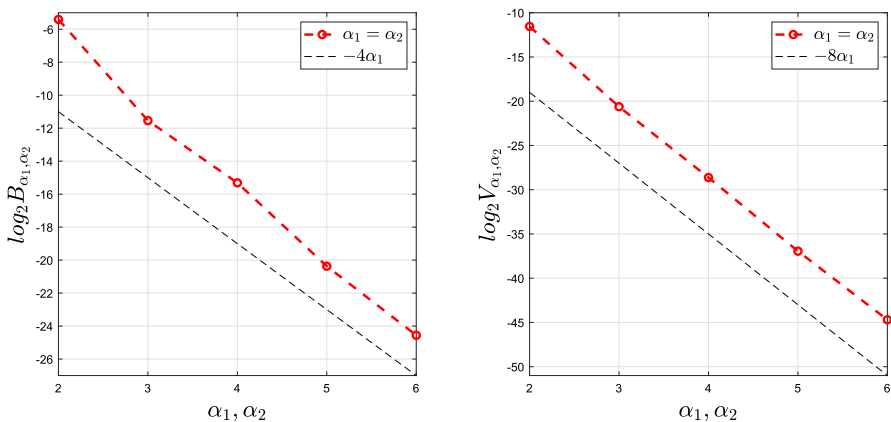
Fig. 8 2D Elliptic PDE with random diffusion coefficient. Verification of MISMC rates as in Assumption 4.2 for $\alpha_1 = \alpha_2$, computed with 20 realisations and 1000 samples for each realisation. Left: $s_1 + s_2$. Right: $\beta_1 + \beta_2$
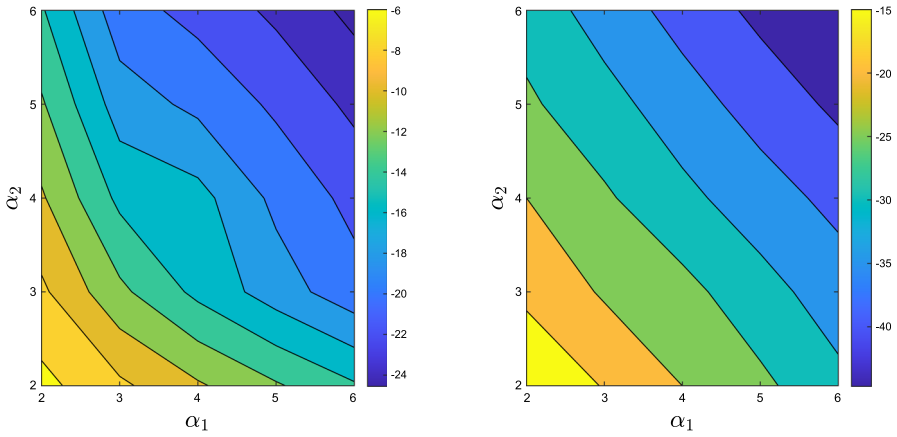
Fig. 9 2D Elliptic PDE with random diffusion coefficient. Verification of MISMC rates as in Assumption 4.2, computed with 20 realisations and 1000 samples for each realisation. Left: s. Right: β
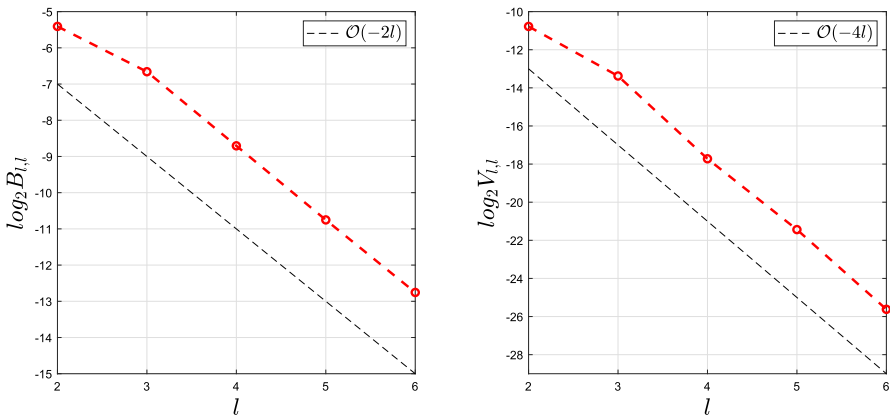
Fig. 10 2D Elliptic PDE with random diffusion coefficient. Verification of increment rates associated with MLSMC. Computed with 20 realisations and 2000 samples for each realisation. Left: s. Right: β
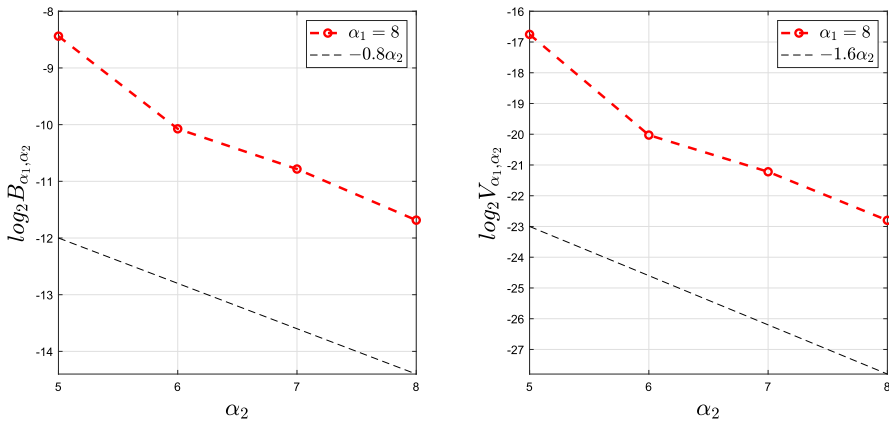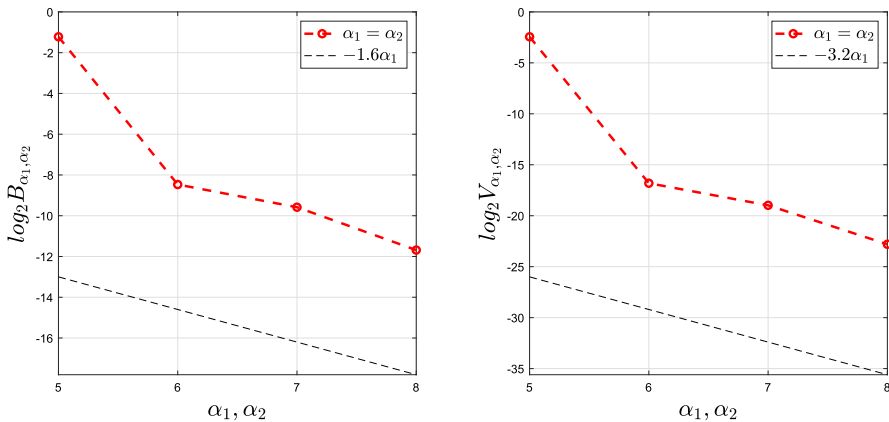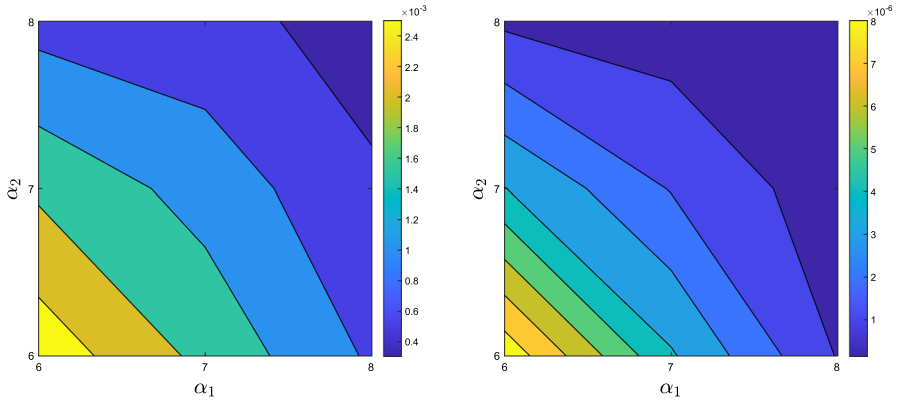
C.3 LGC

See Figs. 11, 12, 13 and 14.

Fig. 11 LGC model. Verification of mixed rates associated with Assumption 4.2 for MISMC, over α_2 given $\alpha_1 = 8$, computed with 20 realisations and 1000 samples for each realisation. Left: s_2. Right: β_2. The same result holds over α_1 for $\alpha_2 = 8$ (not shown)

Fig. 12 LGC model. Verification of mixed rates associated with Assumption 4.2 for MISMC, over $\alpha_2 = \alpha_1$, computed with 20 realisations and 1000 samples for each realisation. Left: $2s$. Right: 2β

Fig. 13 LGC model. Verification of mixed rates associated with Assumption 4.2 for MISMC, over α_2 and α_1, computed with 20 realisations and 1000 samples for each realisation. Left: $2s$. Right: 2β
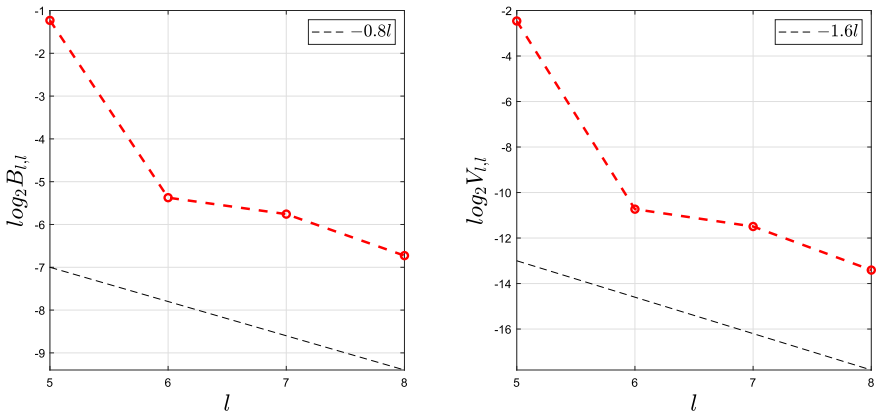

Fig. 14 LGC model. Verification of increment rates for MLSMC, computed with 20 realisations and 1000 samples for each realisation. Left: s. Right: β
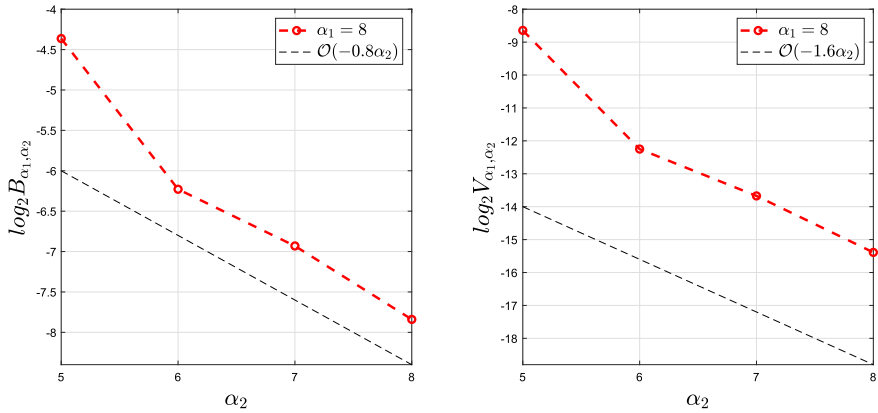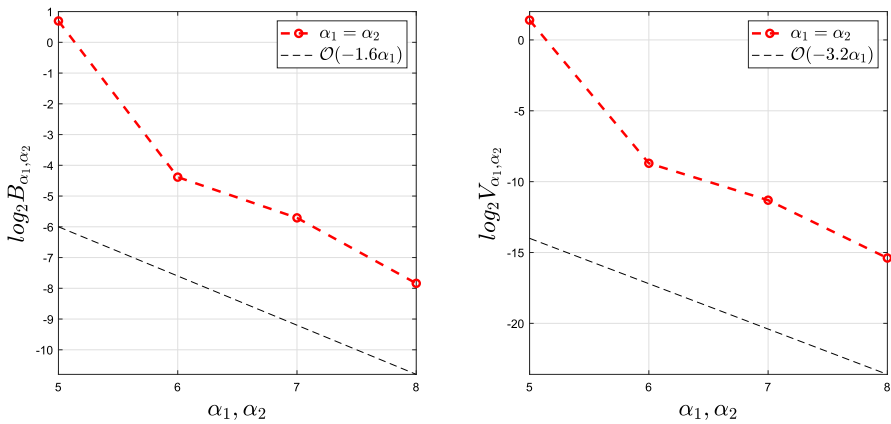
C.4 LGP

See Figs. 15, 16, 17 and 18.

Fig. 15 LGP model. Verification of mixed rates associated with Assumption 4.2 for MISMC, over α_2 given $\alpha_1 = 8$, computed with 20 realisations and 1000 samples for each realisation. Left: s_2. Right: β_2. The same result holds over α_1 for $\alpha_2 = 8$ (not shown)

Fig. 16 LGP model. Verification of mixed rates associated with Assumption 4.2 for MISMC, over $\alpha_2 = \alpha_1$, computed with 20 realisations and 1000 samples for each realisation. Left: $2s$. Right: 2β
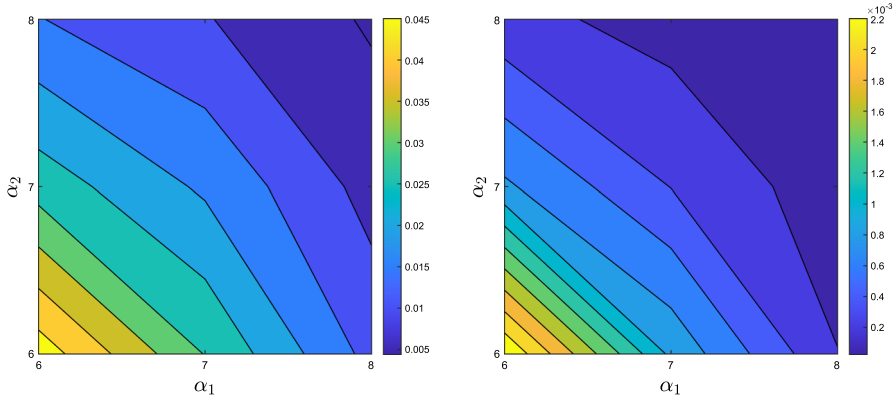
Fig. 17 LGP model. Verification of mixed rates associated with Assumption 4.2 for MISMC, over α_2 and α_1, computed with 20 realisations and 1000 samples for each realisation. Left: $2s$. Right: 2β
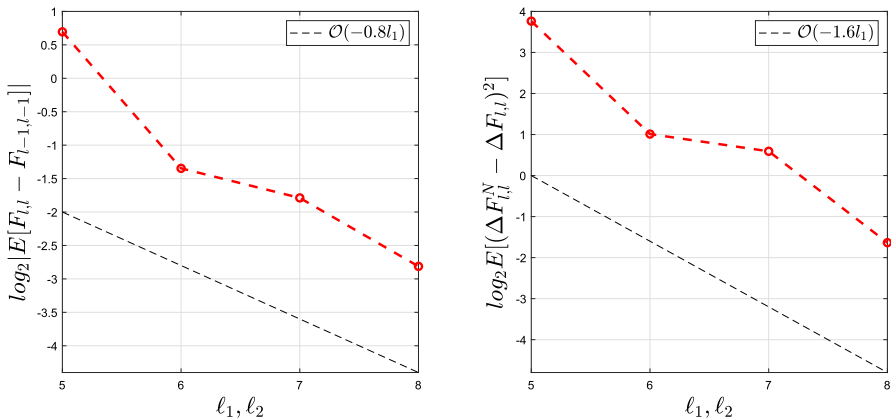

Fig. 18 LGP model. Verification of increment rates for MLSMC, computed with 20 realisations and 1000 samples for each realisation. Left: s. Right: β

References

1. S Agapiou, Omiros Papaspiliopoulos, D Sanz-Alonso, AM Stuart, et al. Importance sampling: Intrinsic dimension and computational cost. Statistical Science, 32(3):405–431, 2017.
2. Marco Ballesio, Ajay Jasra, Erik von Schwerin, and Raul Tempone. A Wasserstein coupled particle filter for multilevel estimation. arXiv:2004.03981, 2020.
3. Thomas Bengtsson, Peter Bickel, Bo Li, et al. Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In Probability and statistics: Essays in honor of David A. Freedman, pages 316–334. Institute of Mathematical Statistics, 2008.
4. Alexandros Beskos, Dan Crisan, Ajay Jasra, Nikolas Kantas, and Hamza Ruzayqat. Score-based parameter estimation for a class of continuous-time state space models. SIAM Journal on Scientific Computing, 43(4):A2555–A2580, 2021.

5. Alexandros Beskos, Ajay Jasra, Kody Law, Youssef Marzouk, and Yan Zhou. Multilevel sequential Monte Carlo with dimension-independent likelihood-informed proposals. SIAM/ASA Journal on Uncertainty Quantification, 6(2):762–786, 2018.

6. Alexandros Beskos, Ajay Jasra, Kody J.H. Law, Raul Tempone, and Yan Zhou. Multilevel sequential Monte Carlo samplers. Stochastic Processes and their Applications, 127(5):1417–1440, 2017.

7. William E Boyce, Richard C DiPrima, and Douglas B Meade. Elementary differential equations. John Wiley & Sons, 2017.

8. Dietrich Braess. Finite elements: Theory, fast solvers, and applications in solid mechanics. Cambridge University Press, 2007.

9. Susanne Brenner and Ridgway Scott. The mathematical theory of finite element methods, volume 15. Springer Science & Business Media, 2007.

10. Hans-Joachim Bungartz and Michael Griebel. Sparse grids. Acta numerica, 13:147–269, 2004.

11. Neil K Chada, Jordan Franks, Ajay Jasra, Kody J Law, and Matti Vihola. Unbiased inference for discretely observed hidden markov model diffusions. SIAM/ASA Journal on Uncertainty Quantification, 9(2):763–787, 2021.

12. Sourav Chatterjee, Persi Diaconis, et al. The sample size required in importance sampling. The Annals of Applied Probability, 28(2):1099–1135, 2018.

13. Alexey Chernov, Håkon Hoel, Kody JH Law, Fabio Nobile, and Raul Tempone. Multilevel ensemble Kalman filtering for spatio-temporal processes. Numerische Mathematik, 147(1):71–125, 2021.

14. Nicolas Chopin. A sequential particle filter method for static models. Biometrika, 89(3):539–552, 2002.

15. Nicolas Chopin, Omiros Papaspiliopoulos, et al. An introduction to sequential Monte Carlo, volume 4. Springer, 2020.

16. Philippe G Ciarlet. The finite element method for elliptic problems. SIAM, 2002.

17. Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. Statistical Science, pages 424–446, 2013.

18. David Cruz-Uribe and CJ Neugebauer. Sharp error bounds for the trapezoidal rule and simpson's rule. J. Inequal. Pure Appl. Math, 3(4):1–22, 2002.

19. T. Cui, Ajay Jasra, and Kody J. H. Law. Multi-index sequential Monte Carlo methods. Preprint.

20. Pierre Del Moral. Feynman-Kac formulae. In Feynman-Kac Formulae, pages 47–93. Springer, 2004.

21. Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(3):411–436, 2006.

22. Tim J Dodwell, Christian Ketelsen, Robert Scheichl, and Aretha L Teckentrup. A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow. SIAM/ASA Journal on Uncertainty Quantification, 3(1):1075–1108, 2015.

23. Howard Elman, Alison Ramage, and David Silvester. Algorithm 866: IFISS, a Matlab toolbox for modelling incompressible flow. ACM Trans. Math. Softw., 33:2–14, 2007.

24. Howard C Elman, David J Silvester, and Andrew J Wathen. Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics. Numerical Mathematics and Scie, 2014.

25. Alexandre Ern and Jean-Luc Guermond. Theory and practice of finite elements, volume 159. Springer, 2004.

26. Charles J Geyer. Practical Markov chain Monte Carlo. Statistical science, pages 473–483, 1992.

27. Michael B Giles. Multilevel Monte Carlo methods. Acta Numerica, 24:259, 2015.

28. Alastair Gregory, Colin J Cotter, and Sebastian Reich. Multilevel ensemble transform particle filtering. SIAM Journal on Scientific Computing, 38(3):A1317–A1338, 2016.

29. Abdul-Lateef Haji-Ali, Fabio Nobile, and Raúl Tempone. Multi-index Monte Carlo: when sparsity meets sampling. Numerische Mathematik, 132(4):767–806, 2016.

30. Raúl Tempone Håkon Hoel, Gaukhar Shaimerdenova. Multilevel ensemble Kalman Filtering based on a sample average of independent EnKF estimators. Foundations of Data Science, 2(4):351–390, 2020.

31. Jeremy Heng, Adrian N Bishop, George Deligiannidis, and Arnaud Doucet. Controlled sequential Monte Carlo. The Annals of Statistics, 48(5):2904–2929, 2020.

32. Jeremy Heng and Pierre E Jacob. Unbiased Hamiltonian Monte Carlo with couplings. Biometrika, 106(2):287–302, 2019.

33. Jeremy Heng, Ajay Jasra, Kody JH Law, and Alexander Tarakanov. On unbiased estimation for discretized models. arXiv preprint arXiv:2102.12230, 2021.

34. Viet Ha Hoang, Christoph Schwab, and Andrew M Stuart. Complexity analysis of accelerated MCMC methods for Bayesian inversion. Inverse Problems, 29(8):085010, 2013.

35. Håkon Hoel, Kody JH Law, and Raul Tempone. Multilevel ensemble Kalman filtering. SIAM Journal on Numerical Analysis, 54(3):1813–1839, 2016.
36. Christopher Jarzynski. Nonequilibrium equality for free energy differences. Physical Review Letters, 78(14):2690, 1997.
37. Ajay Jasra, Kengo Kamatani, Kody J. H. Law, and Yan Zhou. Bayesian static parameter estimation for partially observed diffusions via multilevel Monte Carlo. SIAM Journal on Scientific Computing, 40(2):A887–A902, 2018.
38. Ajay Jasra, Kengo Kamatani, Kody J. H. Law, and Yan Zhou. A multi-index Markov chain Monte Carlo method. International Journal for Uncertainty Quantification, 8(1), 2018.
39. Ajay Jasra, Kengo Kamatani, Kody JH Law, and Yan Zhou. Multilevel particle filters. SIAM Journal on Numerical Analysis, 55(6):3068–3096, 2017.
40. Ajay Jasra, Kody Law, and Fangyuan Yu. Unbiased filtering of a class of partially observed diffusions. To appear in Advances in Applied Probability, arXiv preprint arXiv:2002.03747, 2020.
41. Ajay Jasra, Kody JH Law, and Deng Lu. Unbiased estimation of the gradient of the log-likelihood in inverse problems. Statistics and Computing, 31(3):1–18, 2021.
42. Ajay Jasra, Kody JH Law, and Prince Peprah Osei. Multilevel particle filters for lévy-driven stochastic differential equations. Statistics and Computing, 29(4):775–789, 2019.
43. Ajay Jasra, Kody JH Law, and Yaxian Xu. Multi-index sequential Monte Carlo methods for partially observed stochastic partial differential equations. International Journal for Uncertainty Quantification, 11(3), 2021.
44. Ajay Jasra, Kody JH Law, and Fangyuan Yu. Randomized multilevel Monte Carlo for embarrassingly parallel inference. To appear in SMC 2022 Proceedings, arXiv preprint arXiv:2107.01913, 2021.
45. Ajay Jasra, Fangyuan Yu, and Jeremy Heng. Multilevel particle filters for the non-linear filtering problem in continuous time. Statistics and Computing, 30(5):1381–1402, 2020.
46. Gabriel J Lord, Catherine E Powell, and Tony Shardlow. An introduction to computational stochastic PDEs, volume 50. Cambridge University Press, 2014.
47. Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log Gaussian Cox processes. Scandinavian journal of statistics, 25(3):451–482, 1998.
48. Pierre Del Moral, Ajay Jasra, Kody JH Law, and Yan Zhou. Multilevel sequential Monte Carlo samplers for normalizing constants. ACM Transactions on Modeling and Computer Simulation (TOMACS), 27(3):1–22, 2017.
49. Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 541–548. JMLR Workshop and Conference Proceedings, 2010.
50. Radford Neal. Regression and classification using Gaussian process priors. Bayesian statistics, 6:475, 1998.
51. Radford M Neal. Annealed importance sampling. Statistics and computing, 11(2):125–139, 2001.
52. Jorge Nocedal and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.
53. Bernt Øksendal. Stochastic differential equations. In Stochastic differential equations. Springer, 2003.
54. Grigorios A Pavliotis. Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations, volume 60. Springer, 2014.
55. Christian Robert and George Casella. Monte Carlo statistical methods. Springer Science & Business Media, 2013.
56. Hamza Ruzayqat, Neil K Chada, and Ajay Jasra. Multilevel estimation of normalization constants using the ensemble Kalman-Bucy filter. arXiv preprint arXiv:2108.03935, 2021.
57. Robert Scheichl, Andrew M Stuart, and Aretha L Teckentrup. Quasi-Monte Carlo and multilevel Monte Carlo methods for computing posterior expectations in elliptic inverse problems. SIAM/ASA Journal on Uncertainty Quantification, 5(1):493–518, 2017.
58. Walter A Strauss. Partial differential equations: An introduction. John Wiley & Sons, 2007.
59. Andrew M Stuart. Inverse problems: a Bayesian perspective. Acta numerica, 19:451–559, 2010.
60. Luke Tierney. A note on metropolis-hastings kernels for general state spaces. Annals of applied probability, pages 1–9, 1998.
61. Surya T Tokdar and Jayanta K Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. Journal of statistical planning and inference, 137(1):34–42, 2007.
62. David Williams. Probability with martingales. Cambridge university press, 1991.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.