Statistics and Computing (2021) 31:21
https://doi.org/10.1007/s11222-021-09994-6

Unbiased estimation of the gradient of the log-likelihood in inverse problems

Ajay Jasra[1] · Kody J. H. Law[2] · Deng Lu[3]

Received: 10 March 2020 / Accepted: 7 January 2021 / Published online: 3 March 2021

Abstract

We consider the problem of estimating a parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ associated with a Bayesian inverse problem. Typically one must resort to a numerical approximation of gradient of the log-likelihood and also adopt a discretization of the problem in space and/or time. We develop a new methodology to unbiasedly estimate the gradient of the log-likelihood with respect to the unknown parameter, i.e. the expectation of the estimate has no discretization bias. Such a property is not only useful for estimation in terms of the original stochastic model of interest, but can be used in stochastic gradient algorithms which benefit from unbiased estimates. Under appropriate assumptions, we prove that our estimator is not only unbiased but of finite variance. In addition, when implemented on a single processor, we show that the cost to achieve a given level of error is comparable to multilevel Monte Carlo methods, both practically and theoretically. However, the new algorithm is highly amenable to parallel computation.

Keywords Parameter estimation · Inverse problems · Unbiased estimation · Stochastic gradient

1 Introduction

The problem of inferring unknown parameters associated with the solution of (partial) differential equations (PDEs) is referred to as an inverse problem. In such a context, when the forward problem is well posed, the inverse problem is often ill-posed and challenging to solve, even numerically. The area has a long history and a large literature (see e.g. Engl et al. (1996); Tikhonov and Glasko (1964)) yet the intersection with statistics is still comparatively small, particularly considering the significant overlap, in terms of methods, algorithms as well as objectives. If one adopts a Bayesian approach to solution of the inverse problem, then the object of interest is a posterior distribution and, in particular, expectations with respect to this distribution (Franklin 1970; Stuart 2010). While this provides an elegant solution and quantified uncertainty via well-defined target distribution, it is more computationally challenging to solve than its deterministic counterpart. Here, we assume solution of the Bayesian inverse problem (BIP) requires computationally intensive Monte Carlo methods for accurate estimation. We furthermore assume that the statistical model can only be defined up to some unknown parameters.

Consider a BIP with unknown $u \in \mathsf{X}$ and data $y \in \mathsf{Y}$, related through a PDE, and assume that there is an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ in the sense that one could consider the posterior density

$$p(u, \theta|y) \propto p(y|u, \theta)p(u|\theta)p(\theta)$$

where $p(y|u, \theta)$ is the conditional likelihood function of $Y|(u, \theta)$, $p(u|\theta)$ is the prior on the unknown u, for a given θ, and p is some prior density for θ. Due to the strong correlation of the unknown u with respect to the parameter θ, such posterior distributions can be highly complex and very challenging to sample from, even using quite advanced Markov

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

Ajay Jasra
ajay.jasra@kaust.edu.sa

Deng Lu
denglu@u.nus.edu

[1] Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Kingdom of Saudi Arabia

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

[3] Department of Statistics & Applied Probability,, National University of Singapore, Singapore 117546, Singapore

chain Monte Carlo (MCMC) algorithms. In this article, the unknown u is treated as a nuisance parameter and the goal is to maximize the marginal likelihood of the parameters

$$p_\theta(y) = \int_{\mathsf{X}} p_\theta(y|u)p_\theta(u)du$$

where we have dropped the conditioning on θ and used subscripts instead, as is classical in the literature (e.g. Cappé et al. (2005)) and du is a sigma-finite measure on some measurable space $(\mathsf{X}, \mathcal{X})$. Maximum likelihood estimation based upon the marginal likelihood appears in a wide variety of modelling contexts (e.g. Gland and Mevel (1997)), and the resulting estimators often have desirable mathematical properties, e.g. Cappé et al. (2005) in the context of hidden Markov models. In our given scenario, one is left with a finite-dimensional optimization problem, albeit with an objective function that is not available analytically. This intractability arises from two sources:

- first, for a given (u, θ) only a discretization of the likelihood $p_\theta(y|u)$ can be evaluated;
- second, the discretized marginal likelihood is a high-dimensional integral which itself must be approximated.

Moreover, the associated gradient of the log-likelihood is not available, which may be of interest in optimization algorithms. In the following, we will suppress the notation for fixed observation y and present the method generally. In particular, we use the notation $\gamma_\theta(u) = p_\theta(y|u)p_\theta(u)$. We will set $Z_\theta = p_\theta(y)$ and $\eta_\theta(du) = [\gamma_\theta(u)/Z_\theta]du$

In this article, we present a new scheme to provide finite variance estimates of the gradient of the log-likelihood that are unbiased. To be precise, let $E_\theta = \nabla_\theta \log(Z_\theta)$ denote the gradient of the log-likelihood with no discretization bias. The proposed method provides an estimator $\hat{E}_\theta$ such that $\mathbb{E}[\hat{E}_\theta] = E_\theta$, where $\mathbb{E}$ is the expectation with respect to the randomization induced by our numerical approach. Moreover, the estimator $\hat{E}_\theta$ is constructed so that one only needs access to finite resolution (discretized) approximations of the BIP. This scheme is of interest for several reasons:

1. Unbiased estimates of gradients help to facilitate stochastic gradient algorithms;
2. The method is easy to parallelize;
3. The method helps to provide a benchmark for other computations.

In terms of the first point, it is often simpler to verify the validity of stochastic gradient algorithms when the estimate of the noisy functional is unbiased e.g. Benveniste et al. (1990). While this is not always needed (see Tadic and Doucet (2018) for a special case, which does not apply in our context), it at least provides the user a peace of mind when implementing optimization schemes. The second point relates to efficiency of implementation relative to competing methods. Indeed the method delivers *independent and identically distributed (i.i.d.) unbiased estimators of expectations with respect to the posterior*, which are therefore "embarrassingly parallel" to simulate. Synchronization is only required to average these in order to construct the final estimator. This is a scalar operation, so is unlikely to impact the strong scaling with reasonably complex models and existing (super)computer architectures. The practical limitations in terms of memory will be investigated in a future publication. The third point simply states that one can check the precision of biased methodology. We now explain the approach in a little more detail.

The method that we use is based upon a technique developed in Jasra et al. (2020). In that article, the authors consider the filtering of a class of diffusion processes, which have to be discretized. The authors develop a method which allows one to approximate the filtering distribution, unbiasedly and without any discretization error. The methodology that is used in Jasra et al. (2020) is a double randomization scheme based upon the approaches in McLeish (2011); Rhee and Glynn (2015). The work in McLeish (2011); Rhee and Glynn (2015) provides a methodology to turn a sequence of convergent estimators into an unbiased estimator, using judicious randomization across the level of discretization. It is determined for the problem of interest in Jasra et al. (2020) that an additional randomization is required in order to derive efficient estimators, that is, estimators that are competitive with the existing state-of-the-art methods in the literature. In this article, we follow the basic approach that is used in Jasra et al. (2020), except that one cannot use the same estimation methodology for the current problem. An approach is introduced in Beskos et al. (2017b) which enables application of the related deterministic multilevel Monte Carlo identity (Vihola 2018) to a sequential Monte Carlo (SMC) sampler Del Moral (2013, 2004) for inference in the present context. In this article, we consider such a strategy to allow the application of the approach in Jasra et al. (2020) to unbiasedly estimate the gradient of the log-likelihood for BIPs. The method of Beskos et al. (2017b) is one of the most efficient techniques that could be used for estimation of the gradient of the log-likelihood for BIPs. However, this method is subject to discretization bias. In other words, suppose E_θ^l is the gradient of the log-likelihood with a choice of discretization bias level, e.g. 2^{-l}. The original method would produce an estimate $\hat{E}_\theta^l$ for which $\mathbb{E}[\hat{E}_\theta^l] \neq E_\theta^l$. On the other hand, under assumptions, it is proved that the new method introduced here can produce an estimate E_θ with finite variance and without bias, i.e. $\mathbb{E}[\hat{E}_\theta] = E_\theta^\infty$. We also show that the cost to achieve a given variance is very similar to the multilevel SMC (MLSMC) approach of Beskos et al. (2017b), with high probability. This is confirmed in numerical simulations. We furthermore numerically investigate the utility of our new estimator in the context of stochastic gradient algorithms, where it is shown that a huge improvement in efficiency is possible.

Our approach is one of the first which can in general provide unbiased and finite variance estimators of the gradient of the log-likelihood for BIPs. A possible alternative would be the approach of Agapiou et al. (2018); however, the methodology in that article is not as general as is presented here. In more detail, the work presented in Agapiou et al. (2018) considers unbiased estimation of expectations associated with particular classes of BIP, namely ones with either uniform or Gaussian priors. The article uses the randomization schemes of McLeish (2011); Rhee and Glynn (2015) along with particular coupled Markov chain simulation schemes. In the case of uniform priors, the methodology is based upon a global minorization condition, which is not reasonable in that case, but not if the priors (when truncated to finite dimension) lie on a non-compact state space. In the case of Gaussian priors, the approach is based upon a synchronous coupling of a type of Metropolis–Hastings algorithm. Such couplings require strong assumptions and may not always mix well. We do not constrain our BIP to uniform or Gaussian priors, and in principle, one does not need a particularly restrictive set of assumptions to apply our methodology. Certain hypotheses are made in order to ensure our proofs are reasonably concise, but these can be relaxed. See Del Moral et al. (2017) for an example.

This article is structured as follows. In Sect. 2, we explain the generic problem to which our approach is applicable. In particular, a concrete example in the context of Bayesian inverse problems is described. In Sect. 3, we present our methodology and the proposed estimator. In Sect. 4, we show that our proposed estimator is unbiased and of finite variance and we consider the cost to obtain the estimate. In Sect. 5, several numerical examples are presented to investigate performance of the estimator in practice, including the efficiency of the estimator when used in the relevant context of a stochastic gradient algorithm for parameter estimation. In "Appendix A", the proofs of some of our theoretical results can be found.

2 Problem setting

2.1 Generic problem

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space, and define a probability measure on it as

$$\eta_\theta(du) = \frac{\gamma_\theta(u)du}{\int_{\mathsf{X}} \gamma_\theta(u)du}$$

where $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, $\gamma : \Theta \times \mathsf{X} \to \mathbb{R}_+$ and recall du is a sigma-finite measure on $(\mathsf{X}, \mathcal{X})$. We are interested in computing

$$\begin{aligned}
&\nabla_\theta \log\Big(\int_{\mathsf{X}} \gamma_\theta(u)du\Big)\\
&\quad = \frac{1}{\int_{\mathsf{X}} \gamma_\theta(u)du}\nabla_\theta\Big(\int_{\mathsf{X}} \gamma_\theta(u)du\Big)\\
&\quad = \frac{1}{\int_{\mathsf{X}} \gamma_\theta(u)du}\int_{\mathsf{X}} \frac{\nabla_\theta\{\gamma_\theta(u)\}}{\gamma_\theta(u)}\gamma_\theta(u)du\\
&\quad = \int_{\mathsf{X}} \frac{\nabla_\theta\{\gamma_\theta(u)\}}{\gamma_\theta(u)}\eta_\theta(du)\\
&\quad = \int_{\mathsf{X}} \nabla_\theta \log\Big(\gamma_\theta(u)\Big)\eta_\theta(du)\\
&\quad = \int_{\mathsf{X}} \varphi_\theta(u)\eta_\theta(du)\,,
\end{aligned}$$

where we have defined $\varphi_\theta(u) = \nabla_\theta \log\Big(\gamma_\theta(u)\Big)$ and we are assuming it is legitimate to interchange differentiation and integration in the second line. From here on, we will use the following shorthand notation for a measure μ on $(\mathsf{X}, \mathcal{X})$ and a measurable $\mu-$integrable $\varphi : \mathsf{X} \to \mathbb{R}^d$

$$\mu(\varphi) := \int_{\mathsf{X}} \varphi(x)\mu(dx)\,,$$

which should be understood as a column vector of integrals.

In practice, we assume that we must work with an approximation of $\varphi_\theta(u)$ and $\eta_\theta(du)$. Let $l \in \mathbb{N}_0$, and set

$$\eta_\theta^l(du) = \frac{\gamma_\theta^l(u)du}{\int_{\mathsf{X}} \gamma_\theta^l(u)du}$$

where $\gamma^l : \Theta \times \mathsf{X} \to \mathbb{R}_+$. We are now interested in computing

$$\begin{aligned}
&\nabla_\theta \log\Big(\int_{\mathsf{X}} \gamma_\theta^l(u)du\Big)\\
&\quad = \int_{\mathsf{X}} \nabla_\theta \log\Big(\gamma_\theta^l(u)\Big)\eta_\theta^l(du)\\
&\quad = \int_{\mathsf{X}} \varphi_\theta^l(u)\eta_\theta^l(du).
\end{aligned}$$

It is assumed explicitly that $\forall \theta \in \Theta$

$$\lim_{l \to +\infty} \eta_\theta^l(\varphi_\theta^l) = \eta_\theta(\varphi_\theta).$$

2.2 Example of problem

We will focus on the following particular problem as an example. Let $D \subset \mathbb{R}^d$ with $\partial D \in C^1$ convex and $f \in L^2(D)$. Consider the following PDE on D:

$$- \nabla \cdot (\hat{u}^s(u) \nabla v) = f, \quad \text{on } D, \tag{1}$$

$$v = 0, \quad \text{on } \partial D\,. \tag{2}$$

Two examples of diffusion coefficients will be considered

$$\hat{u}^s(x;u) = \begin{cases} \bar{u} + \sum_{k=1}^{K} u_k \sigma_k \phi_k(x)\,, & s = 1\,, \\ \sum_{k=1}^{K} \exp(u_k) \phi_k(x)\,, & s = 2\,. \end{cases} \tag{3}$$

For any $s \in \{1, 2\}$, we define $u = \{u_k\}_{k=1}^{K}$, and the state space will be $\mathsf{X} = \prod_{k=1}^{K} [-1, 1]$. Let $v(\cdot; u)$ denote the weak solution of (1) for parameter value u. If $s = 1$, the prior is given by $u_k \sim U[-1, 1]$ (the uniform distribution on $[-1, 1]$) i.i.d. for $k = 1, \ldots, K$. If $s = 2$, the prior is given by the truncated normal $u \sim \mathbf{1}_{u \in \mathsf{X}} N(0, \Sigma_\theta)$, where $N(0, \Sigma_\theta)$ denotes the normal random variable with mean 0 and covariance matrix Σ_θ, and $\mathbf{1}_{\mathsf{X}}$ is the indicator function on the set X. In this case, the prior depends on some parameters $\theta \in \Theta$ through Σ_θ. The covariance matrix Σ_θ is defined elementwise by $\Sigma_{\theta,ij} = \theta_1 \exp(-\theta_2 \|x_i - x_j\|^2)$. The notation $\| \cdot \|$ is always used for the L^2-norm, either the Euclidean norm of the appropriate dimension or the $L^2(D)$ norm. The corresponding inner products will be denoted $\langle \cdot, \cdot \rangle$. The following will be assumed.

(H1) If $s = 1$, then $\phi_k \in C(D)$, $\|\phi_k\|_\infty \leq 1$, and there is a $u_* > 0$ such that $\bar{u} > \sum_{k=1}^{K} \sigma_k + u_*$. If $s = 2$ then ϕ_k are piecewise multilinear nodal basis functions as described in Sect. 2.2.1, such that $\hat{u}^s(u)$ interpolates between the coefficients $\exp(u_k)$.

Note that under Assumption (H1), $\hat{u}^s(u) > \min\{u_*, \exp(-1)\}$ uniformly in u for any $s \in \{1, 2\}$. Hence, there is a well-defined (weak) solution $v(\cdot; u)$ which is bounded in uniformly in u in $L^\infty(D)$ and $L^2(D)$ and its gradient is also bounded uniformly in u in $L^2(D)$ Ciarlet (2002); Dashti and Stuart (2011). Define the following vector-valued function

$$\mathcal{G}(u) = [\langle g_1, v(\cdot; u) \rangle, \ldots, \langle g_m, v(\cdot; u) \rangle]^\intercal,$$

where $g_i \in L^2(D)$ for $i = 1, \ldots, m$. We note that pointwise evaluation is also permissible since $u \in L^\infty(D)$, i.e. g_i can be Dirac delta functions; however, for simplicity we restrict the presentation to $L^2(D)$. It is assumed that the data take the form

$$y = \mathcal{G}(u) + \xi, \quad \xi \sim N(0, \theta_3^{-1} \cdot \boldsymbol{I}_m), \quad \xi \perp u\,,$$

where $\perp$ denotes independence. The un-normalized density of u for fixed θ is then given by the following (if $s = 1$ then $d_\theta = 1$ and we use the notation $\theta = \theta_3$)

$$\gamma_{\theta,s}(u) = \begin{cases} \theta^{m/2} \exp\Big(- \frac{\theta}{2} \|\mathcal{G}(u) - y\|^2 \Big) \mathbf{1}_{\mathbf{u} \in \mathsf{X}}\,, & s = 1\,, \\ \theta_3^{m/2} \exp\Big(- \frac{\theta_3}{2} \|\mathcal{G}(u) - y\|^2 \Big) \exp\Big(-\frac{1}{2} \| \Sigma_\theta^{-1/2} u \|^2 \Big) \mathbf{1}_{\mathbf{u} \in \mathsf{X}}\,, & s = 2\,. \end{cases} \tag{4}$$

The normalized density is given by

$$\eta_{\theta,s}(u) = \frac{\gamma_{\theta,s}(u)}{Z_{\theta,s}}\,,$$

where $Z_{\theta,s} = \int_{\mathsf{X}} \gamma_{\theta,s}(u) du$, and the quantity of interest is defined for $u \in \mathsf{X}$ as

$$\varphi_{\theta,s}(u) := \nabla_\theta \log\Big(\gamma_{\theta,s}(u)\Big) = \begin{cases} \frac{m}{2\theta} - \frac{1}{2} \|\mathcal{G}(u) - y\|^2\,, & s = 1\,, \\ \begin{pmatrix} \frac{1}{2} u^\intercal \Sigma_\theta^{-1} \frac{\partial \Sigma_\theta}{\partial \theta_1} \Sigma_\theta^{-1} u \\ \frac{1}{2} u^\intercal \Sigma_\theta^{-1} \frac{\partial \Sigma_\theta}{\partial \theta_2} \Sigma_\theta^{-1} u \\ \frac{m}{2\theta_3} - \frac{1}{2} \|\mathcal{G}(u) - y\|^2 \end{pmatrix}\,, & s = 2\,. \end{cases} \tag{5}$$

Henceforth, the notation s will be suppressed except when it is necessary to refer to the above explicit example. The method presented is very general and applicable to essentially any problem of the form presented in Sect. 2.1, as long as it satisfies appropriate assumptions which will be provided below.

2.2.1 Numerical approximation

The finite element method (FEM) is utilized for solution of (2) with piecewise multilinear nodal basis functions. For $s = 1$, we will let $d = 1$ and $D = [0, 1]$, while for $s = 2$ we will let $d = 2$ and $D = [0, 1]^2$. For brevity, we will completely present only the case $s = 1$. For the case $s = d = 2$, the product of such piecewise linear functions described below is used to generate a bilinear approximation on $[0, 1]^2$. The reader is referred to Section 5.3 of Beskos et al. (2017a) for details. We note the approach is easily generalized following standard FEM literature Brenner and Scott

(2007). As mentioned above, the notation for s is suppressed here.

The PDE problem at resolution level l is solved using FEM with piecewise linear shape functions on a uniform mesh of width $h_l = 2^{-l}$, for $l \geq 0$. Thus, on the lth level the finite element basis functions are $\{\psi_i^l\}_{i=1}^{2^l-1}$ defined as (for $x_i = i \cdot 2^{-l}$):

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

To solve the PDE, $v^l(x) = \sum_{i=1}^{2^l-1} v_i^l \psi_i^l(x)$ is plugged into (1) and projected onto each basis element:

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

resulting in the following linear system:

$$\boldsymbol{A}^l(u)\boldsymbol{v}^l = \boldsymbol{f}^l,$$

where we introduce the matrix $\boldsymbol{A}^l(u)$ with entries $A_{ij}^l(u) = \langle \hat{u} \nabla \psi_i^l, \nabla \psi_j^l \rangle$, and vectors $\boldsymbol{v}^l$, $\boldsymbol{f}^l$ with entries $v_i^l = \langle v, \psi_i^l \rangle$ and $f_i^l = \langle f, \psi_i^l \rangle$, respectively.

Define $\mathcal{G}^l(u) = [\langle g_1, v^l(\cdot; u) \rangle, \ldots, \langle g_m, v^l(\cdot; u) \rangle]^\intercal$. Denote the corresponding approximated un-normalized density by

$$\gamma_\theta^l(u) = \theta^{m/2} \exp\left\{ -\frac{\theta}{2} \|\mathcal{G}^l(u) - y\|^2 \right\} \mathbf{1}_{\mathbf{u} \in \mathsf{X}}, \tag{6}$$

and the approximated normalized density by

$$\eta_\theta^l(u) = \frac{\gamma_\theta^l(u)}{Z_\theta^l},$$

where $Z_\theta^l = \int_{\mathsf{X}} \gamma_\theta^l(u) du$. We further define

$$\begin{aligned} \varphi_\theta^l(u) &:= \nabla_\theta \log\left(\gamma_\theta^l(u)\right) \\ &= \frac{m}{2\theta} - \frac{1}{2} \|\mathcal{G}^l(u) - y\|^2. \end{aligned} \tag{7}$$

Recall from (5) that in the case $s = 2$ only $\partial \log\left(\gamma_\theta^l(u)\right)/\partial\theta_3$ depends upon l, and is given as above.

It is well known that under assumption (H1) $v^l(u)$ converges to $v(u)$ as $l \to \infty$ in $L^2(D)$ (as does its gradient), uniformly in u (Brenner and Scott 2007; Ciarlet 2002). Furthermore, continuity ensures $\gamma_\theta^l(u)$ converges to $\gamma_\theta(u)$ and $\varphi_\theta^l(u)$ converges to $\varphi_\theta(u)$ uniformly in u as well. See also Beskos et al. (2017a, b) and Sect. 4 for further details.

3 Methodology for unbiased estimation

We now describe our methodology for computing an unbiased estimate of $\eta_\theta(\varphi_\theta)$. For simplicity of exposition, we will suppose that for $i \in \{1, \ldots, d_\theta\}$, $(\varphi_\theta(u))_i \in \mathcal{B}_b(\mathsf{X})$, where $(x)_i$ denotes the i^{th} element of a vector and $\mathcal{B}_b(\mathsf{X})$ are the collection of bounded, measurable and real-valued functions on X. This constraint is not needed for the numerical implementation of the method, but shall reduce most of the technical exposition to follow. As remarked in Introduction, the basic approach follows that in Jasra et al. (2020) with some notable differences. We now detail how the approach will work.

3.1 Methodology in Jasra et al. (2020)

The underlying approach of Jasra et al. (2020) is a type of double randomization scheme. The first step is to use the single-term estimator as developed in Rhee and Glynn (2015). Suppose one wants to estimate $\eta_\theta(\varphi_\theta)$, but, only has access to a methodology that can approximate $\eta_\theta^l(\varphi_\theta^l)$ for each fixed $l \in \mathbb{N}_0$. Let $\mathbb{P}_L(l)$ be a positive probability mass function on $\mathbb{N}_0$ and suppose that one can construct a sequence of random variables $(\Xi_\theta^l)_{l \geq 0}$ such that

$$\mathbb{E}[\Xi_\theta^0] = \eta_\theta^0(\varphi_\theta^0) \tag{8}$$

$$\mathbb{E}[\Xi_\theta^l] = \eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1}) \qquad l \in \mathbb{N} \tag{9}$$

and that

$$\sum_{l \in \mathbb{N}_0} \frac{1}{\mathbb{P}_L(l)} \mathbb{E}[\|\Xi_\theta^l\|^2] < +\infty. \tag{10}$$

Now if one draws $L \sim \mathbb{P}_L(\cdot)$, then $\Xi_\theta^L/\mathbb{P}_L(L)$ is an unbiased and finite variance estimator of $\eta_\theta(\varphi_\theta)$. It should be noted that (8)-(9) are not necessary conditions, but are sufficient to ensure the unbiasedness of the estimator.

In the context of interest, it can be challenging to obtain a sequence of random variables which can possess the properties (8)-(10). We will detail one possible approach at a high level and then explain in detail how one can actually construct a simulation method to achieve this high-level description.

3.2 High-level approach

The objective of this section is to highlight the generic procedure that is used in Jasra et al. (2020) for producing estimates that satisfy (8)–(9). The basic idea is to use another application of randomization to construct such unbiased estimators from a consistent sequence of estimators. In particular, consider a given increasing sequence $(N_p)_{p \in \mathbb{N}_0}$ with $N_p \in \mathbb{N}$ for each $p \in \mathbb{N}_0$, $1 \leq N_0 < N_1 < \cdots$ and $\lim_{p \to \infty} N_p = \infty$. Then, we suppose that one can construct N_p-sample Monte

Carlo (type) estimators $\xi_\theta^{l,p}$ for $l \in \mathbb{N}_0$, such that almost surely the following consistency results hold

$$\lim_{p\to\infty} \xi_\theta^{0,p} = \eta_\theta^0(\varphi_\theta^0)\,, \tag{11}$$

$$\lim_{p\to\infty} \xi_\theta^{l,p} = \eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1})\,, \qquad l \in \mathbb{N}\,. \tag{12}$$

For a given $(l, p, p') \in \mathbb{N}_0^3$, $p \neq p'$ we do *not* require $\xi_\theta^{l,p}$ and $\xi_\theta^{l,p'}$ to be independent, nor do we require unbiasedness of the individual estimators as in

$$\mathbb{E}[\xi_\theta^{0,p}] = \eta_\theta^0(\varphi_\theta^0)\,,$$
$$\mathbb{E}[\xi_\theta^{l,p}] = \eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1})\,, \qquad l \in \mathbb{N}\,.$$

Now set

$$\Xi_\theta^{0,0} := \xi_\theta^{0,0}\,,$$
$$\Xi_\theta^{0,p} := \xi_\theta^{0,p} - \xi_\theta^{0,p-1}\,, \qquad p \in \mathbb{N}\,.$$

For $l \in \mathbb{N}$ given, set

$$\Xi_\theta^{l,0} := \xi_\theta^{l,0}\,,$$
$$\Xi_\theta^{l,p} := \xi_\theta^{l,p} - \xi_\theta^{l,p-1}\,, \qquad p \in \mathbb{N}\,.$$

Let $\mathbb{P}_P(p)$, $p \in \mathbb{N}_0$, be a positive probability mass function with $\overline{\mathbb{P}}_P(p) = \sum_{q=p}^{\infty} \mathbb{P}_P(q)$. Now if

$$\sum_{p\in\mathbb{N}_0} \frac{1}{\overline{\mathbb{P}}_P(p)} \mathbb{E}[\|\xi_\theta^{l,p} - \eta_\theta^0(\varphi_\theta^0)\|^2] < +\infty\,, \tag{13}$$

$$\sum_{p\in\mathbb{N}_0} \frac{1}{\overline{\mathbb{P}}_P(p)} \mathbb{E}[\|\xi_\theta^{l,p} - \{\eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1})\}\|^2] < +\infty\,, \quad l \in \mathbb{N} \tag{14}$$

and $P \sim \mathbb{P}_P(\cdot)$, then

$$\Xi_\theta^l = \sum_{p=0}^{P} \frac{1}{\overline{\mathbb{P}}_P(p)} \Xi_\theta^{l,p} \tag{15}$$

will allow $(\Xi_\theta^l)_{l\in\mathbb{N}_0}$ to satisfy (8)-(9), where expectations are understood to be with respect to $\mathbb{P}_P$ yet P is suppressed in the notation. Moreover, $(\Xi_\theta^l)_{l\in\mathbb{N}_0}$ will have finite variances. This result follows as we are simply using the coupled sum estimator as in Rhee and Glynn (2015) and using (Vihola 2018, Theorem 5), for instance, to verify the conditions required.

3.3 Details of the approach

We will now describe how to obtain the sequence $(\Xi_\theta^{l,p})_{p\in\mathbb{N}_0}$ for $l \in \mathbb{N}_0$ fixed.

3.3.1 MLSMC method of Beskos et al. (2017b)

To introduce our approach, we first consider the MLSMC method in Beskos et al. (2017b) which will form the basis for our estimation procedure. The approach in Beskos et al. (2017b) allows one to consistently estimate quantities such as $\eta_\theta^l(\varphi)$, $\varphi \in \mathcal{B}_b(\mathsf{X})$ and to estimate differences $\eta_\theta^l(\varphi) - \eta_\theta^{l-1}(\varphi)$, $l \in \mathbb{N}$, in a manner which will be suited to our objectives. The algorithm is simply a type of SMC sampler Del Moral et al. (2006) which generates a collection of samples, sequentially, to approximate η_θ^0 then η_θ^1 and so on.

The following notations will be used to accurately describe the method in Beskos et al. (2017b). Define for $l \in \mathbb{N}_0$

$$G_\theta^l(u) = \frac{\gamma_\theta^{l+1}(u)}{\gamma_\theta^l(u)}.$$

Note that $\eta_\theta^{l-1}(G_\theta^{l-1}) = \frac{1}{Z_\theta^{l-1}} \int_{\mathsf{X}} \gamma_\theta^l(u)du = Z_\theta^l / Z_\theta^{l-1}$, where we recall that $Z_\theta^l = \int_{\mathsf{X}} \gamma_\theta^l(u)du$. For $l \in \mathbb{N}$, M_θ^l is a η_θ^l-invariant Markov kernel; that is, for any $\varphi \in \mathcal{B}_b(\mathsf{X})$

$$\eta_\theta^l(\varphi) = \int_{\mathsf{X}} \Big(\int_{\mathsf{X}} \varphi(u') M_\theta^l(u, du') \Big) \eta_\theta^l(du). \tag{16}$$

Define for $\mu \in \mathcal{P}(\mathsf{X})$ (the collection of probability measures on $(\mathsf{X}, \mathcal{X})$), $l \in \mathbb{N}$

$$\Phi_\theta^l(\mu)(du') := \frac{1}{\mu(G_\theta^{l-1})} \int_{\mathsf{X}} G_\theta^{l-1}(u) M_\theta^l(u, du') \mu(du) \tag{17}$$

Noting that

$$\begin{aligned}\eta_\theta^l(\varphi) &= \frac{\eta_\theta^{l-1}(G_\theta^{l-1}\varphi)}{\eta_\theta^{l-1}(G_\theta^{l-1})} \\ &= \frac{Z_\theta^{l-1}}{Z_\theta^l} \eta_\theta^{l-1}(G_\theta^{l-1}\varphi) \quad = \frac{1}{Z_\theta^l} \int_{\mathsf{X}} (\gamma_\theta^l(u)\varphi(u))du\,,\end{aligned} \tag{18}$$

equations (16) and (17) lead to the recursion

$$\begin{aligned}\eta_\theta^l(\varphi) &= \frac{\eta_\theta^{l-1}(G_\theta^{l-1}\varphi)}{\eta_\theta^{l-1}(G_\theta^{l-1})} \\ &= \frac{1}{\eta_\theta^{l-1}(G_\theta^{l-1})} \int_{\mathsf{X}} G_\theta^{l-1}(u) \\ &\Big(\int_{\mathsf{X}} \varphi(u') M_\theta^l(u, du') \Big) \eta_\theta^{l-1}(du) \\ &= \Phi_\theta^l(\eta_\theta^{l-1})(\varphi)\,.\end{aligned} \tag{19}$$

The significance of this recursion is that it allows one to turn a simulation from η_θ^{l-1} into one from η_θ^l. This is important in a sequential context, where we will replace η_θ^{l-1} with approximations, but still use this given recursion.

Algorithm 1 A Multilevel Sequential Monte Carlo Sampler with a fixed number of samples $N \in \mathbb{N}$ and a given level $l \in \mathbb{N}_0$.

```
1. Initialization: For $i \in \{1, \dots, N\}$ sample $U_0^i$ from $\eta_\theta^0$. If $l = 0$ stop; otherwise set $s = 1$ and go to step 2.
2. Resampling and Sampling: For $i \in \{1, \dots, N\}$ sample $U_s^i$ from $\Phi_\theta^s(\eta_\theta^{s-1,N})$. This consists of sampling $a_s^i \in \{1, \dots, N\}$ with probability mass function
    $\mathsf{P}_\theta^N(a_s^i = j) = \frac{G_\theta^{s-1}(u_{s-1}^j)}{\sum_{k=1}^N G_\theta^{s-1}(u_{s-1}^k)}$ ,
   and then sampling $U_s^i$ from $M_\theta^s(u_{s-1}^{a_s^i}, \cdot)$. If $s = l$ stop; otherwise set $s = s + 1$ and return to the start of 2.
```

Consider $N \in \mathbb{N}$, and slightly modify the MLSMC algorithm used in Beskos et al. (2017b) to keep the number of samples across levels fixed, up to some given level $l \in \mathbb{N}$. Details are given in Algorithm 1. This algorithm yields samples distributed according to the following joint law

$$\mathsf{P}_\theta^N\left(d(u_0^{1:N}, \dots, u_l^{1:N})\right) = \left(\prod_{i=1}^N \eta_\theta^0(du_0^i)\right) \left(\prod_{s=1}^l \prod_{i=1}^N \Phi_\theta^s(\eta_\theta^{s-1,N})(du_s^i)\right), \tag{20}$$

where $\eta_\theta^{s-1,N}(du) = \frac{1}{N}\sum_{i=1}^N \delta_{u_{s-1}^i}(du)$ for $s \in \mathbb{N}$. The intuition of the algorithm can be understood by considering (19). For instance, if one observes $\Phi_\theta^1(\eta_\theta^{0,N})(du_1^i)$, we know that $\eta_\theta^{0,N}$ will converge to η_θ^0 as N grows, so we might expect that u_1^i might be approximately sampled according to η_θ^1. If this logic applies recursively, then one can approximate η_θ^2 and so on. One can compute an estimate of $\eta_\theta^0(\varphi_\theta^0)$ as

$$\eta_\theta^{0,N}(\varphi_\theta^0) := \frac{1}{N}\sum_{i=1}^N \varphi_\theta^0(u_0^i) .$$

Following from (18), for $l \in \mathbb{N}$, one can estimate $\eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1})$ with

$$\frac{\eta_\theta^{l-1,N}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N}(G_\theta^{l-1})} - \eta_\theta^{l-1,N}(\varphi_\theta^{l-1}) = \frac{\frac{1}{N}\sum_{i=1}^N G_\theta^{l-1}(u_{l-1}^i)\varphi_\theta^l(u_{l-1}^i)}{\frac{1}{N}\sum_{i=1}^N G_\theta^{l-1}(u_{l-1}^i)} - \frac{1}{N}\sum_{i=1}^N \varphi_\theta^{l-1}(u_{l-1}^i) .$$

The reason for using the samples generated at level $l-1$ to estimate $\eta_\theta^l(\varphi_\theta^l)$ as well as $\eta_\theta^{l-1}(\varphi_\theta^{l-1})$ is to construct estimators which satisfy conditions such as (10). Standard results (for instance in Del Moral (2013)) allow one to prove that almost surely

$$\lim_{N\to\infty} \eta_\theta^{0,N}(\varphi_\theta^0) = \eta_\theta^0(\varphi_\theta^0)$$

$$\lim_{N\to\infty} \left(\frac{\eta_\theta^{l-1,N}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N}(G_\theta^{l-1})} - \eta_\theta^{l-1,N}(\varphi_\theta^{l-1})\right) = \eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1}) , \qquad l \in \mathbb{N} .$$

Note that in general one has

$$\mathsf{E}_\theta^N\left[\left(\frac{\eta_\theta^{l-1,N}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N}(G_\theta^{l-1})} - \eta_\theta^{l-1,N}(\varphi_\theta^{l-1})\right)\right] \neq \eta_\theta^l(\varphi_\theta^l) - \eta_\theta^{l-1}(\varphi_\theta^{l-1}) , \qquad l \in \mathbb{N} ,$$

where E_θ^N is an expectation associated with the probability in (20).

3.3.2 Approach for constructing $(\Xi_\theta^{l,p})_{p\in\mathbb{N}_0}$

In order to calculate our approximation, we will consider the following approach, which was also used in Jasra et al. (2020). Given any $(l, P) \in \mathbb{N}_0^2$, we will run Algorithm 2 in order to obtain $(\Xi_\theta^{l,p})_{p\in\{0,1,\dots,P\}}$.

Algorithm 2 Approach to construct $(\Xi_\theta^{l,p})_{p\in\{0,1,\dots,P\}}$ for $(l, P) \in \mathbb{N}_0^2$ given.

```
1. Sample: Run Algorithm 1 *independently* with $N_p - N_{p-1}$ samples for $p \in \{0, 1, \dots, P\}$, up to level $(l-1) \vee 0$, where we define for convenience $N_{-1} := 0$.
2. Estimate: construct $\Xi_\theta^{l,p}$ as in equation (23), for $p \in \{0, 1, \dots, P\}$.
```

The joint probability law of the samples simulated according to Algorithm 2 is

$$\mathbb{P}_\theta\left(d(u_0^{1:N_p}, \dots, u_{(l-1)\vee 0}^{1:N_p})\right) = \prod_{p=0}^P \mathsf{P}_\theta^{N_p - N_{p-1}} \left((u_0^{N_{p-1}+1:N_p}, \dots, u_{(l-1)\vee 0}^{N_{p-1}+1:N_p})\right), \tag{21}$$

where $N_{-1} = 0$ and $\mathsf{P}_\theta^{N_p - N_{p-1}}$ is as defined in (20). For $(l, P) \in \mathbb{N}_0^2$ given, consider running Algorithm 2. Then for any $s \in \{0, 1, \dots, (l-1) \vee 0\}$ and any $p \in \{0, \dots, P\}$, we can construct the following empirical probability measure on $(\mathsf{X}, \mathcal{X})$

$$\eta_\theta^{s,N_{0:p}}(du_s) := \sum_{q=0}^{p}\Big(\frac{N_q - N_{q-1}}{N_p}\Big)\eta_\theta^{s,N_q-N_{q-1}}(du_s)\,. \tag{22}$$

Note the recursion

$$\eta_\theta^{s,N_{0:p}}(du_s) = \Big(\frac{N_p - N_{p-1}}{N_p}\Big)\eta_\theta^{s,N_p-N_{p-1}}(du_s)$$
$$+\frac{N_{p-1}}{N_p}\eta_\theta^{s,N_{0:p-1}}(du_s)\,.$$

Now define

$$\Xi_\theta^{l,p} := \begin{cases} \eta_\theta^{0,N_{0:p}}(\varphi_\theta^0) - \eta_\theta^{0,N_{0:p-1}}(\varphi_\theta^0) & \text{if } l=0 \\ \frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \eta_\theta^{l-1,N_{0:p}}(\varphi_\theta^{l-1}) - \Big(\frac{\eta_\theta^{l-1,N_{0:p-1}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p-1}}(G_\theta^{l-1})} - \eta_\theta^{l-1,N_{0:p-1}}(\varphi_\theta^{l-1})\Big) & \text{otherwise}\,, \end{cases} \tag{23}$$

where $\eta_\theta^{0,N_{0:-1}}(\varphi_\theta^0) := 0$, and

$$\frac{\eta_\theta^{l-1,N_{0:-1}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:-1}}(G_\theta^{l-1})}$$
$$-\eta_\theta^{l-1,N_{0:-1}}(\varphi_\theta^{l-1}) := 0\,.$$

For convenience in the next section, the conditions (13)-(14) translated to the notations used in this section are

$$\sum_{p\in\mathbb{N}_0}\frac{1}{\overline{\mathbb{P}}_P(p)}\mathbb{E}_\theta$$
$$[\|[\eta_\theta^{0,N_{0:p}} - \eta_\theta^0](\varphi_\theta^0)\|^2] < +\infty \tag{24}$$
$$\sum_{p\in\mathbb{N}_0}\frac{1}{\overline{\mathbb{P}}_P(p)}\mathbb{E}_\theta\Big[\Big\|\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \eta_\theta^{l-1,N_{0:p}}(\varphi_\theta^{l-1})$$
$$-\Big(\frac{\eta_\theta^{l-1}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1}(G_\theta^{l-1})} - \eta_\theta^{l-1}(\varphi_\theta^{l-1})\Big)\Big\|^2\Big]$$
$$< +\infty,\ \ l \in \mathbb{N}\,, \tag{25}$$

where $\mathbb{E}_\theta$ is used to denote expectation associated with the probability $\mathbb{P}_\theta$ in (21).

3.4 Method

The new method is now presented in Algorithm 3. The estimate of $\eta_\theta(\varphi_\theta)$ is given by (26). In Sect. 4, we will prove that it is both unbiased and of finite variance, as well as investigating the cost of computing the estimate. Indeed one has a finite variance unbiased estimator even if $M = 1$, and in the numerical examples, it will be shown that such an estimator can be useful in some contexts. However, the variance of such estimator may be rather large. Therefore, in the interest of accuracy one will typically want to adjust M to achieve suitable variance.

Algorithm 3 Method for Unbiasedly Estimating $\eta_\theta(\varphi_\theta)$.

```
For $i = 1, \dots, M$:
1. Generate $L_i \sim \mathbb{P}_L$ and $P_i \sim \mathbb{P}_P$.
2. Run Algorithm 2 with $l = L_i$ and $P = P_i$.
3.Compute:
$\Xi_\theta^{L_i} = \sum_{p=0}^{P_i} \frac{1}{\overline{\mathbb{P}}_P(p)} \Xi_\theta^{L_i,p}$
    where $\Xi_\theta^{L_i,p}$ is given in (23).
Return the estimate:
$\widehat{\eta_\theta(\varphi_\theta)} := \frac{1}{M}\sum_{i=1}^{M}\frac{1}{\mathbb{P}_L(L_i)}\Xi_\theta^{L^i}$ .        (26)
```

There are several points of practical interest to be made at this stage (the first two were noted already in Jasra et al. (2020)). First, the loop over the number of independent samples i in Algorithm 3 can be easily parallelized. Second, one does not need to make L and P independent; this is only assumed for simplicity of presentation, but is not required. Third, the current method uses only the level $l-1$ marginal of (21). All the samples for $s =\in \{0, \dots, l-2\}$ and associated empirical measures (22) are discarded and only the level $l-1$ empirical measure is utilized. This differs from Beskos et al. (2017b) where all the lower-level empirical measures are used. It is possible these samples could be utilized to improve the accuracy of the method, but it is not necessary and so is not investigated further here. The potential efficiency of the double randomization scheme, as well as a discussion of the overall efficiency of the approach, is given in (Jasra et al. 2020, Section 2.5).

4 Theoretical results

Our main objective is to show that $(\Xi_\theta^l)_{l\in\mathbb{N}_0}$ as defined in (15) with $(\Xi_\theta^{l,p})_{p\in\mathbb{N}_0}$ as in (23) will satisfy (8)-(10). To that end, one must first show that $(\Xi_\theta^{l,p})_{p\in\mathbb{N}_0}$ satisfy (24)-(25) which certainly verifies (8)-(9) and then one must establish that (10) holds. We make the following assumptions.

(A1) For each $\theta \in \Theta$, there exist $0 < \underline{C} < \overline{C} < +\infty$ such that

$$\sup_{l \ge 0} \sup_{u \in \mathsf{X}} G_\theta^l(u) \le \overline{C}$$
$$\inf_{l \ge 0} \inf_{u \in \mathsf{X}} G_\theta^l(u) \ge \underline{C}.$$

(A2) For each $\theta \in \Theta$, , there exist a $\rho \in (0,1)$ such that for any $l \ge 1$, $(u,v) \in \mathsf{X}^2$, $A \in \mathcal{X}$

$$\int_A M_\theta^l(u, du') \ge \rho \int_A M_\theta^l(v, dv').$$

(A3) For each $\theta \in \Theta$, there exists a $\widetilde{C} < +\infty$ such that for each $i \in \{1, \dots, d_\theta\}$

$$\sup_{l \ge 0} \sup_{u \in \mathsf{X}} |(\varphi_\theta^l(u))_i| \le \widetilde{C}.$$

For $\varphi \in \mathcal{B}_b(\mathsf{X})$, we set $\|\varphi\|_\infty = \sup_{u \in \mathsf{X}} |\varphi(u)|$. To simplify our notations, we will set for $l \in \mathbb{N}$

$$\|\varphi_\theta^l - \varphi_\theta^{l-1}\|_\infty^2 = \max_{i \in \{1,\dots,d_\theta\}} \left\{ \|(\varphi_\theta^l)_i - (\varphi_\theta^{l-1})_i\|_\infty^2 \right\}.$$

We begin with the following result, which is associated with verifying that (24)-(25) can hold.

Proposition 4.1 *Assume (A1-3). Then for any $\theta \in \Theta$ there exists a $C < +\infty$ such that for any $p \in \mathbb{N}_0$, $1 \le N_0 < N_1 < \cdots < N_p < +\infty$:*

$$\mathbb{E}_\theta[\|[\eta_\theta^{0,N_{0:p}} - \eta_\theta^0](\varphi_\theta^0)\|^2] \le \frac{C}{N_p}\Big(1 + \frac{p^2}{N_p}\Big).$$

In addition, for any $(l,p) \in \mathbb{N} \times \mathbb{N}_0$, $1 \le N_0 < N_1 < \cdots < N_p < +\infty$:

$$\mathbb{E}_\theta\Bigg[\Bigg\| \frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \eta_\theta^{l-1,N_{0:p}}(\varphi_\theta^{l-1})$$
$$-\Big(\frac{\eta_\theta^{l-1}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1}(G_\theta^{l-1})} - \eta_\theta^{l-1}(\varphi_\theta^{l-1})\Big)\Bigg\|^2\Bigg]$$
$$\le \frac{C}{N_p}\Big(1 + \frac{p^2}{N_p}\Big)\Big(\|\varphi_\theta^l - \varphi_\theta^{l-1}\|_\infty^2 + \Big\|G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big\|_\infty^2\Big).$$

Proof The first result follows by Lemma A.2 in appendix and the second from Lemma A.4 also in appendix. □

Remark 4.1 To show that (24)-(25) can hold, one can set, for instance $N_p = 2^p$ and $\mathbb{P}_P(p) \propto 2^{-p}(p+1)\log_2(p+2)^2$. See, for example, Jasra et al. (2020) and Rhee and Glynn (2015).

To continue our discussion, to complete our proof, we must know something about the quantity

$$\|\varphi_\theta^l - \varphi_\theta^{l-1}\|_\infty^2 + \Big\|G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big\|_\infty^2$$

in terms of a possible decay as a function of l. To that end, we shall assume that this term is $\mathcal{O}(h_l^\beta)$ for some $\beta > 0$. This assumption can be verified for the example in Sect. 2.2. Recall from Sect. 2.2 that $h_l = 2^{-l}$.

Proposition 4.2 *Assume (H1). Then there are $C > 0$ and $\beta > 0$, depending on f and u_*, such that for all $u \in \mathsf{X}$*

$$\|v^l(\cdot;u)\|, \|v(\cdot;u)\| < C\,, \qquad \|v^l(\cdot;u) - v(\cdot;u)\|^2 \le C h_l^\beta\,, \tag{27}$$

where the norm is $L^2(D)$. Given a function $F : \mathbb{N} \times \mathsf{X} \to \mathbb{R}^n$, suppose that there is a $C' > 0$ which does not depend on (l,u) such that for all $i = 1, \dots, n$

$$|F_i(l,u) - F_i(\infty,u)| \le C' \|v^l(\cdot;u) - v(\cdot;u)\|\,, \tag{28}$$

where the norm is $L^2(D)$, and $F(\infty,\cdot) := \lim_{l\to\infty} F(l,\cdot)$. Then there is another $C > 0$ which does not depend on (l,u) such that

$$\max_{i \in \{1,\dots,n\}} \left\{ \sup_{u \in \mathsf{X}} |F_i(l,u) - F_i(l-1,u)|^2 \right\} \le C h_l^\beta\,. \tag{29}$$

Proof This is a slight generalization of the results of Beskos et al. (2017b), Sect. 4. In particular, Eqs. 18, 21, 22 and 23 of Beskos et al. (2017b) provide (27). Plugging (27) into (28) immediately implies that there is a $C > 0$ independent of l such that $\sup_{u \in \mathsf{X}} |F_i(l,u) - F_i(\infty,u)|^2 \le C h_l^\beta$ for all $i = 1, \dots, n$. The triangle inequality implies

$$|F_i(l,u) - F_i(l-1,u)|^2 \le 2(|F_i(l,u) - F_i(\infty,u)|^2$$
$$+ |F_i(l-1,u) - F_i(\infty,u)|^2)\,.$$

Since $h_{l-1} = 2h_l$, combining the results at levels l and $l-1$ gives the result. □

This proposition simply states that if a vector-valued function has the continuity property (28) with respect to the solution of (2), then the quantity on the left-hand side of (29) inherits the same rate of convergence.

Note that $G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} = \frac{G_\theta^{l-1}}{\eta_\theta^{l-1}(G_\theta^{l-1})}$ and $G_\theta^\infty = 1$. So

$$
\begin{aligned}
&\|G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^{l}}-1\|_\infty\\
&\quad=\left\|\frac{1}{\eta_\theta^{l-1}(G_\theta^{l-1})}\left((G_\theta^{l-1}-1)+\eta_\theta^{l-1}(1-G_\theta^{l-1})\right)\right\|_\infty\\
&\quad\le\left\|\frac{2}{\eta_\theta^{l-1}(G_\theta^{l-1})}\right\|_\infty\|G_\theta^{l-1}-1\|_\infty\,.
\end{aligned}
$$

A slight generalization of Prop. 4.1 of Beskos et al. (2017b) to include β, together with Proposition 4.2 imply assumption (A1) and that there is a $C>0$ such that

$$
\left\|G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^{l}}-1\right\|_\infty^2\le Ch_l^\beta\,.
$$

Now defining $F(l,u):=\varphi_\theta^l(u)$ and $F(\infty,u)=\varphi_\theta(u)$, as in (7) and (5) for $s=1$, then it is easy to show that Proposition 4.2 ensures

$$
\|\varphi_\theta^l-\varphi_\theta^{l-1}\|_\infty^2\le Ch_l^\beta\,.
$$

In particular, notice that $n=d_\theta=1$ and

$$
\begin{aligned}
&\varphi_\theta^l-\varphi_\theta\\
&\quad=\frac{1}{2}\left(\|\mathcal{G}(u)-y\|^2-\|\mathcal{G}^l(u)-y\|^2\right)\\
&\quad=\frac{1}{2}\left(\|\mathcal{G}(u)\|^2-\|\mathcal{G}^l(u)\|^2+\langle y,\mathcal{G}^l(u)-\mathcal{G}(u)\rangle\right)\\
&\quad\le\frac{1}{2}(\|\mathcal{G}(u)\|+\|\mathcal{G}^l(u)\|+\|y\|)\|\mathcal{G}^l(u)-\mathcal{G}(u)\|\,, \qquad (30)
\end{aligned}
$$

where the scalar identity $(a+b)(a-b)=a^2-b^2$ is used along with Cauchy–Schwartz twice in going from the second to the third line (the norms and inner product are all Euclidean on $\mathbb{R}^m$). Now using Cauchy–Schwartz again m times

$$
\begin{aligned}
\|\mathcal{G}^l(u)-\mathcal{G}(u)\|&=\left(\sum_{i=1}^m\langle g_i,v^l(\cdot;u)-v(\cdot;u)\rangle^2\right)^{1/2}\\
&\le\left(\sum_{i=1}^m\|g_i\|^2\right)^{1/2}\|v^l(\cdot;u)-v(\cdot;u)\|\,.
\end{aligned}
$$

It should not cause confusion that the norms and inner products to the right of the first equal sign are on $L^2(D)$. Assumption (A3) then follows straightforwardly from (27). As mentioned in Sect. 2.2.1, (7), for the case $s=2$ only $(\varphi_\theta^l)_3$ depends explicitly on l, and it is given as above. For $i=1,2$, we have $(\varphi_\theta^l)_i=(\varphi_\theta^{l-1})_i$.

Theorem 1 *Assume (A1-3). Then there exist choices of* $\mathbb{P}_L,\mathbb{P}_P$ *and* $(N_p)_{p\in\mathbb{N}_0}$, $1\le N_0<N_1<\cdots$ *so that* $(\Xi_\theta^l)_{l\in\mathbb{N}_0}$ *as defined in* (15) *with* $(\Xi_\theta^{l,p})_{p\in\mathbb{N}_0}$ *as in* (23) *will satisfy* (8)-(10). *That is, (26) is an unbiased and finite variance estimator of* $\eta_\theta(\varphi_\theta)$.

Proof Throughout the proof, C is a finite constant that will not depend on l or p and whose value will change upon each appearance. Given the commentary above, we need only show that (10) can hold for some given choices of $\mathbb{P}_L,\mathbb{P}_P$ and $(N_p)_{p\in\mathbb{N}_0}$. Note that in the context of (10) Ξ_θ^l is as (15) and in the case of (15) $\Xi_\theta^{l,p}$ is as (23). We have that, by direct computation and the Cauchy–Schwarz inequality

$$
\begin{aligned}
&\sum_{l\in\mathbb{N}_0}\frac{1}{\mathbb{P}_L(l)}\mathbb{E}_\theta[\|\Xi_\theta^l\|^2]\\
&\quad\le\sum_{(l,p)\in\mathbb{N}_0^2}\frac{\mathbb{P}_P(p)}{\mathbb{P}_L(l)}\Bigg\{\sum_{s=0}^p\frac{\mathbb{E}_\theta[\|\Xi_\theta^{l,s}\|^2]}{\overline{\mathbb{P}}_P(s)^2}\\
&\qquad+2\sum_{0\le s<q\le p}\frac{\mathbb{E}_\theta[\|\Xi_\theta^{l,s}\|\|\Xi_\theta^{l,q}\|]}{\overline{\mathbb{P}}_P(s)\overline{\mathbb{P}}_P(q)}\Bigg\}. \qquad (31)
\end{aligned}
$$

The strategy is now to control the terms $\mathbb{E}_\theta[\|\Xi_\theta^{l,s}\|^2]$ and $\mathbb{E}_\theta[\|\Xi_\theta^{l,s}\|\|\Xi_\theta^{l,q}\|]$, which can then allow one to find choices of $\mathbb{P}_L,\mathbb{P}_P$ and $(N_p)_{p\in\mathbb{N}_0}$ to ensure that the R.H.S. of (31) is finite. Now recalling (23) and noting that for $p\in\mathbb{N}$

$$
\begin{aligned}
&\eta_\theta^{0,N_{0:p}}(\varphi_\theta^0)-\eta_\theta^{0,N_{0:p-1}}(\varphi_\theta^0)\\
&\quad=\eta_\theta^{0,N_{0:p}}(\varphi_\theta^0)-\eta_\theta^0(\varphi_\theta^0)\\
&\qquad-\{\eta_\theta^{0,N_{0:p-1}}(\varphi_\theta^0)-\eta_\theta^0(\varphi_\theta^0)\}
\end{aligned}
$$

and that for $p\in\mathbb{N}$

$$
\begin{aligned}
&\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})}\\
&\quad-\eta_\theta^{l-1,N_{0:p}}(\varphi_\theta^{l-1})\\
&\quad-\left(\frac{\eta_\theta^{l-1,N_{0:p-1}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p-1}}(G_\theta^{l-1})}-\eta_\theta^{l-1,N_{0:p-1}}(\varphi_\theta^{l-1})\right)\\
&=\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})}-\eta_\theta^{l-1,N_{0:p}}(\varphi_\theta^{l-1})\\
&\quad-\left(\frac{\eta_\theta^{l-1}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1}(G_\theta^{l-1})}-\eta_\theta^{l-1}(\varphi_\theta^{l-1})\right)\\
&\quad-\Bigg\{\left(\frac{\eta_\theta^{l-1,N_{0:p-1}}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1,N_{0:p-1}}(G_\theta^{l-1})}-\eta_\theta^{l-1,N_{0:p-1}}(\varphi_\theta^{l-1})\right)\\
&\quad-\left(\frac{\eta_\theta^{l-1}(G_\theta^{l-1}\varphi_\theta^l)}{\eta_\theta^{l-1}(G_\theta^{l-1})}-\eta_\theta^{l-1}(\varphi_\theta^{l-1})\right)\Bigg\}
\end{aligned}
$$

by Proposition 4.1 that for $(l, s) \in \mathbb{N}_0^2$

$$\mathbb{E}_\theta[\| \Xi_\theta^{l,s} \|^2] \leq \frac{Ch_l^\beta}{N_s}\Big(1 + \frac{s^2}{N_s}\Big). \tag{32}$$

Also, by using Cauchy–Schwarz, $(l, p, q) \in \mathbb{N}_0 \times \mathbb{N}^2$

$$\mathbb{E}_\theta[\| \Xi_\theta^{l,s} \| \| \Xi_\theta^{l,q} \|] \leq \frac{Ch_l^\beta}{N_s^{1/2}N_q^{1/2}}\Big(1+\frac{s^2}{N_s}\Big)^{1/2}\Big(1+\frac{q^2}{N_q}\Big)^{1/2}. \tag{33}$$

Then using the bounds (32)-(33) in (31) gives the upper bound (noting that the case $s = 0 = q$ the terms $\mathbb{E}_\theta[\| \Xi_\theta^{l,s} \|^2]$ and $\mathbb{E}_\theta[\| \Xi_\theta^{l,s} \| \| \Xi_\theta^{l,q} \|]$ are $\mathcal{O}(1)$ so one can find a C such that the following upper bound holds)

$$\sum_{l\in\mathbb{N}_0} \frac{1}{\mathbb{P}_L(l)}\mathbb{E}_\theta[\| \Xi_\theta^l \|^2]$$
$$\leq C \sum_{(l,p)\in\mathbb{N}_0^2} \frac{\mathbb{P}_P(p)h_l^\beta}{\mathbb{P}_L(l)}\Bigg\{\sum_{s=0}^{p} \frac{\Big(1+\frac{s^2}{N_s}\Big)}{N_s\overline{\mathbb{P}}_P(s)^2} +$$
$$\sum_{0\leq s<q\leq p} \frac{\Big(1+\frac{s^2}{N_s}\Big)^{1/2}\Big(1+\frac{q^2}{N_q}\Big)^{1/2}}{N_s^{1/2}N_q^{1/2}\overline{\mathbb{P}}_P(p)\overline{\mathbb{P}}_P(q)}\Bigg\}.$$

Now if one chooses, for instance $N_p = 2^p$ and $\mathbb{P}_P(p) \propto 2^{-p}(p+1)\log_2(p+2)^2$ and $\mathbb{P}_L(l) \propto h_l^{\alpha\beta}$ for any $\alpha \in (0, 1)$ then (10) is satisfied, and hence, the proof is completed. □

In most cases of practical interest, it is not possible to choose $\mathbb{P}_L$, $\mathbb{P}_P$ and $(N_p)_{p\in\mathbb{N}_0}$ so that (26) is an unbiased and finite variance estimator, as well as having finite expected cost. Suppose, in the case of Sect. 2.2, the cost to evaluate G_θ^l is $\mathcal{O}(h_l^{-1})$ and $\beta = 1$. Then, just as in Jasra et al. (2020), if we choose $\mathbb{P}_L(l) \propto 2^{-l}(l+1)\log_2(l+2)^2$, $N_p = 2^p$, and $\mathbb{P}_P(p) \propto 2^{-p}(p+1)\log_2(p+2)^2$, then to achieve a variance of $\mathcal{O}(\epsilon^2)$ (for $\epsilon > 0$ arbitrary) the cost is $\mathcal{O}(\epsilon^{-2}|\log(\epsilon)|^{2+\delta})$ for any $\delta > 0$, with high probability. For the MLSMC method in Beskos et al. (2017b), the cost to obtain a mean square error of $\mathcal{O}(\epsilon^2)$ is $\mathcal{O}(\epsilon^{-2}\log(\epsilon)^2)$, which is a mild reduction in cost. However, we note that this discussion is constrained to the case of a single processor. The unbiased method is straightforwardly parallelized.

5 Numerical results

First we will consider a toy example where we can analytically calculate the marginal likelihood and investigate the performance of the resulting estimator in comparison with the estimator obtained using the original MLSMC method of Beskos et al. (2017b) (not presented here). Subsequently we will consider the examples from Sect. 2.2. Finally, for all examples we will explore the potential applicability of our estimators within the context of parameter optimization using the stochastic gradient descent (SGD) method.

The forward model is the same for all problems, and the right-hand side $f \in L^2(D)$ is always smooth. Hence, the anticipated rate of convergence is the same, and is estimated as $\beta = 4$, just as in Beskos et al. (2017b). The cost is $\mathcal{O}(h_l^{-\gamma})$ in general for the problem in Sect. 2.2, where γ will depend upon the dimension d of the PDE (2) and the linear solver used for the forward problem. We redefine $h_l = 2^{-3-l}$ for all problems, so that the coarsest resolution is $h_0 = 2^{-3}$. The optimal probability distribution for levels is given by $\mathbb{P}_L(l) \propto 2^{-l(\beta+\gamma)/2}$ Rhee and Glynn 2015. For the toy problem and the $d = 1$ example, we have $\gamma = 1$ and $\mathbb{P}_L(l) \propto 2^{-2.5l}$. For the more complex $d = 2$ example, we have $\gamma = 2 + \epsilon$, for $\epsilon > 0$. We choose $\epsilon = 0.5$ in simulations so we can explicitly calculate $\mathbb{P}_L(l) \propto 2^{-3.25l}$. Both examples are in the so-called canonical regime ($\beta > \gamma$, see (Rhee and Glynn 2015, Section 4) for details), and therefore, we allow unbounded L, i.e. $L_{\max} = \infty$ in the terminology of Jasra et al. (2020). The reason for this is basically that the sum (10) and the corresponding cost series both converge, if the cost is deterministic and $\mathcal{O}(h_l^{-\gamma})$ as a function of h_l. However, in this case the cost depends upon the randomized estimator of the series in p. Since the rate of convergence is borderline in the p direction, $\beta_p = \gamma_p = 1$, as in Jasra et al. (2020) we impose a maximum $P_{\max}$ on P. This is necessary to prevent the possibility of the algorithm getting stuck with an extremely expensive sample. It is discussed further in that work. In particular, we choose $N_p = 2^{p+3}$ and

$$\mathbb{P}_P(p) \propto \mathbb{I}(0 \leq p \leq P_{\max}) \begin{cases} 2^{-p+4} & \text{if } p < 4, \\ 2^{-p} \cdot p \cdot \log_2(p)^2 & \text{otherwise}. \end{cases}$$

The piecewise definition of $\mathbb{P}_P$ ensures that it has the correct asymptotic behaviour, but is also monotonically decreasing. Note that in this regime, i.e. strongly canonical convergence in L, or large $\beta > \gamma$, the MLSMC method easily achieves the optimal complexity $\mathcal{O}(\epsilon^{-2})$. However, since the convergence rate in P is necessarily subcanonical, our method therefore suffers from a logarithmic penalty, i.e. $\mathcal{O}(\epsilon^{-2}\log(\epsilon)^{2+\delta})$, for any $\delta > 0$. This cannot be observed in the simulations though. Empirically we observe that we can set $P_{\max}$ rather small, which is perhaps afforded by the very fast convergence in the L direction. This may be why we cannot see the theoretically predicted log penalty in the simulations.

5.1 Toy example

We first consider an example where the marginal likelihood is analytically calculable. Let $d = 1$ and $D = [0, 1]$ and consider the following DE on D:

$$\begin{aligned} \nabla^2 v &= u, \quad \text{on } D, \\ v &= 0, \quad \text{on } \partial D\,. \end{aligned}$$

The solution of this DE is $v(x;u) = \frac{u}{2}(x^2 - x)$. Define the observation operator as

$$\mathcal{G}(u) = [v(x_1;u), v(x_2;u), \ldots, v(x_m;u)]^\top \triangleq Gu\,.$$

We note that pointwise observations are indeed defined thanks to the regularity of the right-hand side $u \in L^2(D)$. Suppose the observation takes the form $y = \mathcal{G}(u) + \xi$, $\xi \sim N(0, \theta^{-1} \cdot \boldsymbol{I}_m)$, $\xi \perp u$, and $\log(\theta)$ follows a normal prior distribution with mean 0 and variance 1. The explicit solution of the log-likelihood and its gradient are given in "Appendix B".

First, the performance of the unbiased algorithm for a single gradient estimation is verified. The data are generated with $\theta^* = 2$, $m = 50$, and observation operator $\mathcal{G}(u) = [p(x_1;u), p(x_2;u), \ldots, v(x_m;u)]^\top$ with $x_i = i/(m+1)$. The true value of the derivative of the log-likelihood at θ^* is calculated using the analytical solution given in "Appendix B". For each L, the MLSMC estimator is realized 50 times and the MSE is reported. Similarly, the MSE of unbiased algorithm is calculated based on 50 realizations as M increases. The results are presented in the left panel of Fig. 1. The cost reported in the plot is proportional to the sum of the cost per forward solve at level l (tridiagonal linear system), h_l^{-1}, multiplied by the total number of samples at level l. The effect of the truncation at $P_{\max}$ is not observed here, even for very small MSE. We observe that the rate of convergence is the same as MLSMC (the canonical 1/cost rate), and our estimator suffers a slightly higher constant. We note that this is for a single processor and does not account for the (embarrassing) parrallelizability of our algorithm over the M individual estimators – indeed, as mentioned several times already, each estimator is i.i.d. and so each can be in principle computed on a separate processor with 0 communication between them until the individual estimators are combined in (26) to produce the final estimator.

5.2 Examples of Sect. 2.2

We now consider the example from Sect. 2.2, for both cases of the potential defined in (3).

5.2.1 Case $s = 1$

For the case $s = 1$ in (3), we let $d = 1$, $D = [0, 1]$, $f(x) = 100x$. For the prior specification of u, we set $K = 2$, $\bar{u} = 0.15$, and for $k = 1, 2$, we let $\sigma_k = (2/5)4^{-k}$, $\phi_k(x) = \sin(k\pi x)$ if k is odd and $\phi_k(x) = \cos(k\pi x)$ if k is even. The observation operator is $\mathcal{G}(u) = [p(0.25;u), p(0.75;u)]^\top$, and the parameter in observation noise covariance is taken to be $\theta = 0.3$. We note again that pointwise observations are defined thanks to the regularity of the right-hand side $f \in L^2(D)$. Again, θ will follow the same log-normal distribution, and this will be included in the definition of $\gamma_\theta(u)$, as above, in other words, with respect to (5) the target for $s = 1$ here will be replaced with $\gamma_\theta(u) \leftarrow \gamma_\theta(u)\theta^{-1}\exp(-(\log(\theta))^2/2)$.

Here, we do not have an analytical solution, so the true value of the target was first estimated with the MLSMC algorithm with $L = 12$. This sampler was realized 50 times and the average of the estimator is taken as the ground truth. Now for each L, the MLSMC estimator is realized 50 times and the MSE is reported. Similarly, the MSE of unbiased algorithm is calculated based on 50 realizations as M increases. The results are presented in the middle panel of Fig. 1. The cost in the plot is proportional to the sum of the cost per forward solve at level l (tridiagonal linear system), h_l^{-1}, multiplied by the total number of samples at level l. In this case, the bias from truncation at $P_{\max}$ is observed, but only for very small MSE. Again, the comparison shows that our method achieves approximately the same canonical rate, and again, we iterate that this comparison does not leverage the parrallelizability of our method.

5.2.2 Case $s = 2$

For the case $s = 2$ in (3), we let $d = 2$, $D = [0, 1]^2$, and $f(x) = \sum_{i=1}^4 a_i \exp(-\frac{1}{2}\|x - m_i\|^2)$, where $m_i \in \{(0.3, 0.3), (0.3, 0.7), (0.7, 0.3), (0.7, 0.7)\}$ and $a_i \in \{2, -3, -2, 3\}$, respectively. In this case, ϕ_i are piecewise bilinear finite elements of the type described in Sect. 2.2.1 between nodes x_i on a mesh in $[0, 1]^2$ with a fixed $l^* = 2$, for $i = 1, \ldots, K = (2^{l^*} - 1)^2$, such that $\hat{u}(x)$ is itself piecewise bilinear. Observations of the potential function v are collected at 25 measurement points, evenly spaced within $[0.2, 0.6]^2$ (boundaries included). Again pointwise observations are defined even for $d = 2$ due to the regularity of the right-hand side $f \in L^2(D)$. The observation precision $\theta_3^* = 2.64$ is chosen such that a prescribed signal-to-noise ratio (SNR) $\max_{x \in D} |v(x)|\theta_3^* = 10$, where v^* is the truth solution at which the observations are collected. The values of the other parameters are given by $\theta_1^* = 3$ and $\theta_2^* = 5$ Now the priors on θ_1 and θ_2 are taken as Gamma distributions with mean and variance 1, and the prior on θ_3 is again log-normal as above. This will be included in the definition of $\gamma_\theta(u)$, as above, in other words, with respect to (5) the target for $s = 2$ here will be replaced with $\gamma_\theta(u) \leftarrow \gamma_\theta(u)\theta_3^{-1}\exp(-(\log(\theta_3))^2/2 - \theta_1 - \theta_2)$.

Again we do not have an analytical solution here, so the true value of the target was first estimated with the MLSMC algorithm with $L = 8$. This sampler was again realized 50

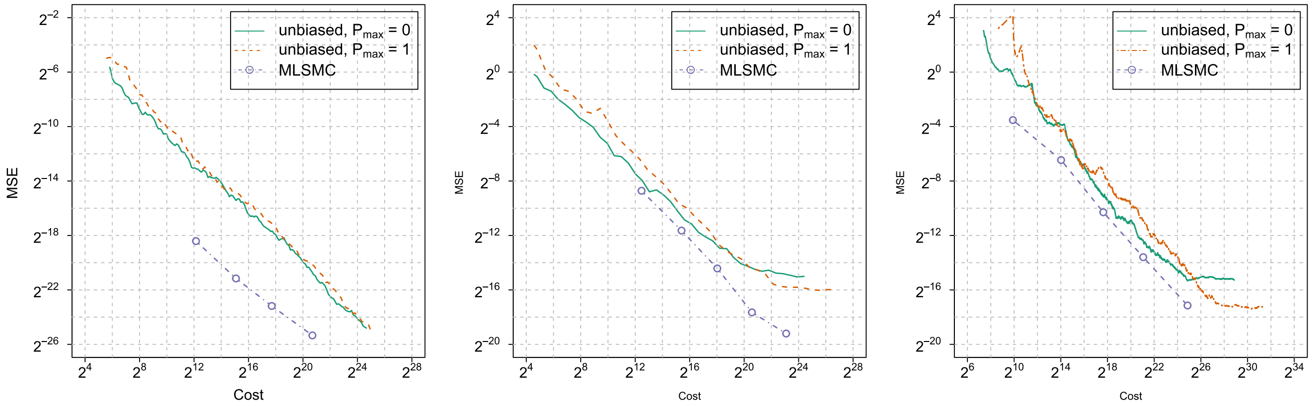

Fig. 1 Single estimator asymptotic complexity in terms of MSE vs cost for (i) the unbiased algorithm with different choices of $P_{\max}$ and (ii) MLSMC. Left: Toy example, single estimator. Middle: Example of Sect. 2.2 with $s = d = 1$ Right: Example of Sect. 2.2 with $s = d = 2$

times and the average of the estimator is taken as the ground truth. Now for each L, the MLSMC estimator is realized 50 times and the MSE is reported. Similarly, the MSE of unbiased algorithm is calculated based on 50 realizations. The results are presented in the right panel of Fig. 1. The forward model inversion is done by the conjugate gradient method with the number of iteration steps proportional to level l, and we find the cost is given by $\gamma = 2$ with a log penalty. The cost in the plot is proportional to the sum of the cost per forward solve at level l multiplied by the total number of samples at level l. The MSE in the plot is the summation of three individual MSE. (We multiply the MSEs of the estimators of θ_1 and θ_2 by a constant so the scale is consistent with the MSE of the estimator of θ_3.) Here, we observe that the curves of our unbiased algorithm are more wiggly than the cures when $s = d = 1$, presumably due to complexity of the model.

5.3 Application to stochastic gradient descent

In this section, we investigate the potential to use our unbiased estimators within the SGD method. The objective here is to find the maximum a posteriori (MAP) estimator of θ by minimizing $-\log p_\theta(y) = -\log(\int \gamma_\theta(u)du)$. Our estimator given in equation (26) provides an unbiased estimator $\widehat{\eta_\theta(\varphi_\theta)}$ of $\nabla_\theta \log p_\theta(y)$, for any choice of $M \geq 1$. In other words $\mathbb{E}\widehat{\eta_\theta(\varphi_\theta)} = \nabla_\theta \log p(y, \theta)$. We will see that it is most efficient to choose $M = 1$. To ensure the output of the SGD algorithm satisfies $\theta > 0$, we let $\theta = \exp(\xi)$ and optimize ξ. The details are given in Algorithm 4.

As above, it makes sense to first explore the toy model with analytical solution, as described in Sect. 5.1. The MAP estimate is first computed using gradient descent. The MSE is then calculated based on 50 realizations, and the cost in the plot is again proportional to the sum of the cost per forward solve at level l (tridiagonal linear system), h_l^{-1}, multiplied by the total number of samples at level l.

Algorithm 4 SGD using new unbiased estimator.

```
1. Initialize $\xi_1$ and choose a sequence $\{\alpha_k\}_{k=1}^{\infty}$ and a value $M \in \mathbb{N}$.
2. For $k = 1, \ldots, K$ (or until convergence criterion is met)
   - Compute $\widehat{\eta_\theta(\varphi_{e^{\xi_k}})}$ using (26)
     - Update $\xi_{k+1} = \xi_k - \alpha_k \widehat{\eta_\theta(\varphi_{e^{\xi_k}})} \exp(\xi_k)$.
3. Return $\theta_{K+1} = \exp(\xi_{K+1})$.
```

There is a well-developed literature on SGD methods. The choice of nuisance parameter α_k, or "learning rate" as it is referred to in the machine learning literature, is still an active topic of investigation (Kushner and Yin 2003; Gower et al. 2019). It is well known that convergence to a local optimum is achieved for unbiased gradients when $\alpha_k \propto k^{-a}$, for $a \in (1/2, 1]$ (Kushner and Yin 2003; Gower et al. 2019), but beyond that we are not aware of a general rule. Furthermore, in our case we are able to control the variance of the gradient, via the number M of individual estimators which are averaged in (26). Our experiments revealed that the method is quite sensitive to the choice of these nuisance parameters. So, despite that it is not the focus of the present work and a general or theoretical treatment is outside of scope, we present the results of several experiments for the analytical model, which have guided our choices and can be useful for practitioners and guide future investigation. In the left panel of Fig. 2, we explore the performance of the unbiased estimator with different choices of $\alpha_k = \alpha_1/k$, $\alpha_1 \in \{0.1, 0.025\}$, and different choices of the number of samples M used to construct $\widehat{\eta_\theta(\varphi_{e^{\xi_k}})}$ using (26) in step 2 of Algorithm 4. The two takeaways from this experiment are that (1) it is more efficient to take fewer samples M (in particular $M = 1$), and (2) it is more efficient to choose a larger constant $\alpha_1 = 0.1$. In particular, the dynamics of the algorithm experiences a phase transition as one varies the constant α_1. A large enough value provides

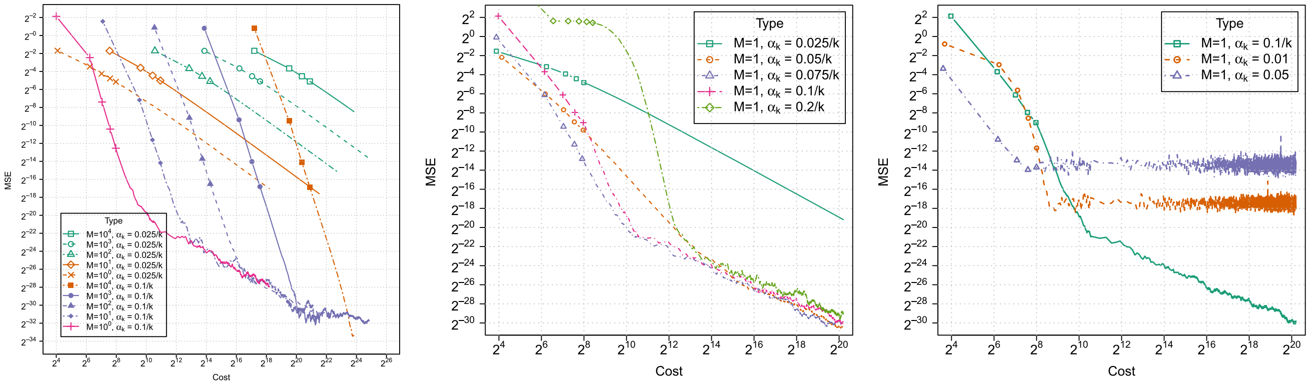

Fig. 2 Toy example, SGD. MSE vs cost for (left) $\alpha_k = 0.025/k$ and $\alpha_k = 0.1/k$ for a range of sample sizes M, (middle) $\alpha_k = \alpha_1/k$ and a range of α_1, and (right) some examples of constant α. $P_{\max} = 0$ is fixed

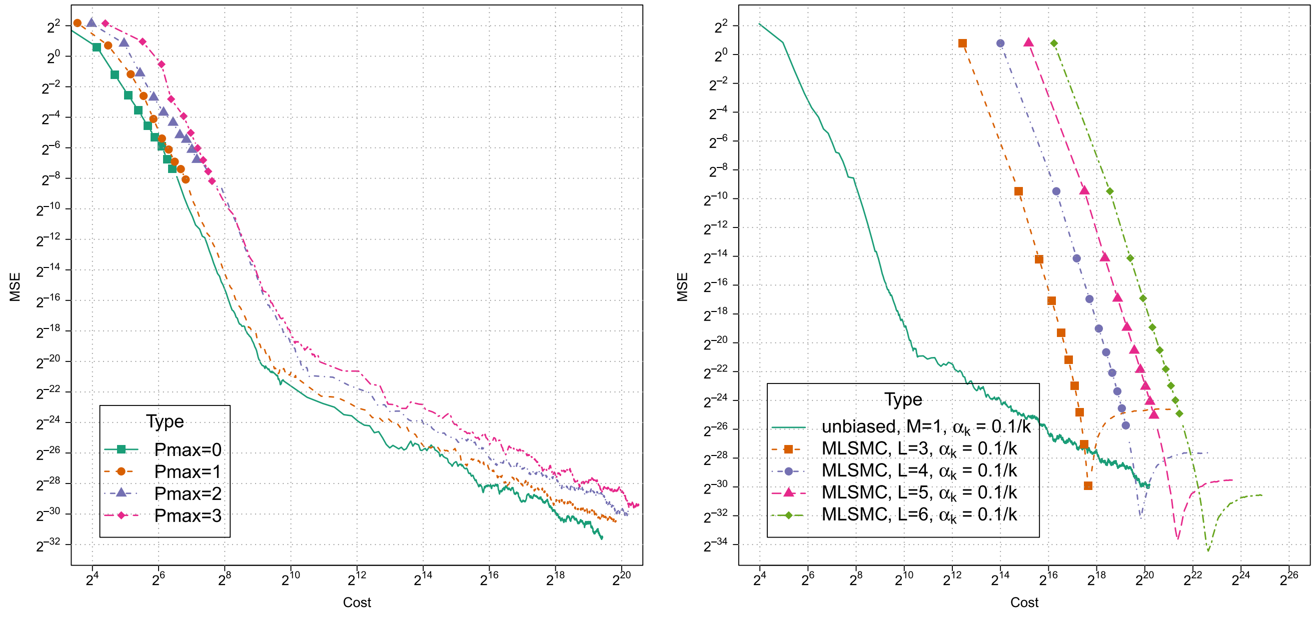

Fig. 3 Toy example, SGD. Left panel: MSE vs cost for $M = 1$ fixed and $\alpha_k = 0.1/k$, and various choices of $P_{\max}$. Right panel: MSE vs cost for $\alpha_k = 0.1/k$, and unbiased estimator with $P_{\max} = 0$ and $M = 1$ in comparison with MLSMC estimator with different choices of L

gradient descent type exponential convergence $\mathcal{O}(e^{-\text{cost}})$, while a value which is too small yields Monte Carlo (MC) type $\mathcal{O}(1/\text{cost})$. It is notable that the exponential convergence eventually gives way to MC type convergence, and that the point where this occurs increases proportionally to the additional constant in cost incurred with larger sample size M, so that the error curves for different values of M eventually intersect. A similar phenomenon has been documented in the recent literature on SGD Gower et al. (2019).

Natural questions are then whether there is a limit to how large one can choose α_1 and at which value precisely the phase transition occurs. These questions are partially answered by the experiments presented in the middle panel of Fig. 2, where we see that α_1 should not be chosen larger than 0.2 and the phase transition happens in between 0.025 and 0.05. The right panel of Fig. 2 illustrates the benefits and drawbacks of using a constant α. In particular, the algorithm may converge more quickly at first, but plateaus when it reaches the induced bias.

In the left panel of Fig. 3, we explore various choices of $P_{\max}$, for $M = 1$ fixed and $\alpha_1 = 0.1$. It is apparent that it is preferable to choose a smaller value of $P_{\max}$. We note, however, that there will be an induced bias, which will be larger for smaller $P_{\max}$. However, for this particular problem we do not even observe that bias over the range of MSE and cost considered.

As a last experiment with the toy example, we compare the convergence of SGD using our unbiased algorithm with $P_{\max} = 0$, $M = 1$, and $\alpha_1 = 0.1$ to the analogous algorithm where an MLSMC estimator with various L (single gradient estimator MSE $\propto 2^{-\beta L}$) replaces the unbiased estimator in step 2 of Algorithm 4. Similar behaviour was observed for

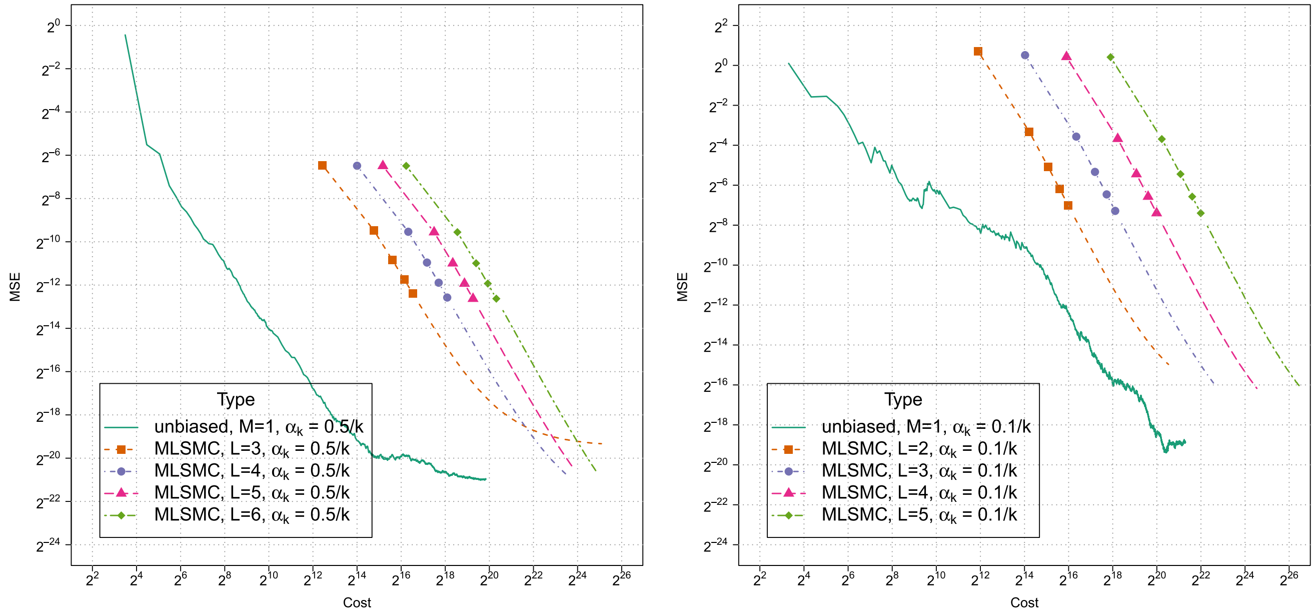

Fig. 4 Examples of Sect. 2.2, SGD. MSE vs cost for unbiased estimator with $P_{\max} = 0$ and $M = 1$ in comparison with MLSMC estimator with different choices of L. Left: $s = d = 1$ and $\alpha_k = 0.5/k$. Right: $s = d = 2$ and $\alpha_k = 0.1/k$

MLSMC relative to different choices of α_k as compared to the unbiased estimator. The results are shown in the right panel of Fig. 3. Here, it is clear that over a wide range of MSE the unbiased estimator provides a significantly more efficient alternative to the MLSMC estimator.

Next, we consider the same last experiment except with the examples of Sect. 2.2. The results are presented in Fig. 4. Again, over a wide range of MSE the unbiased estimator provides a significantly more efficient alternative to the MLSMC estimator, although it is less conspicuous for the $s = 2$ case (right) than when $s = 1$ (left). Here, one can already observe the induced $P_{\max} = 0$ bias for the unbiased estimator around 2^{-20} ($s = 1$) and 2^{-19} ($s = 2$). Adjusting the various tuning parameters resulted in similar behaviour as was observed in the earlier experiments with the toy example. These results are not presented.

Acknowledgements AJ was supported by KAUST baseline funding. Some of this research was supported by Singapore MOE tier 1 grant R-155-000-182-114. KJHL was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1. We thank two referees for their comments which have greatly improved the article.

Appendix A Proofs

We begin with some technical results that have been proved in previous works (e.g. Beskos et al. (2017b); Del Moral (2004)). Before continuing, we recall the C_p inequality which is used several times in our proofs. This states that for two real-valued random variables X and Y associated with an expectation operator $\mathbb{E}$ one has for any $p \in (0, \infty)$ that

$$\mathbb{E}[|X + Y|^p] \leq C_p(\mathbb{E}[|X|^p] + \mathbb{E}[|Y|^p])$$

where $C_p = 2^{p-1}$ if $p \geq 1$ and $C_p = 1$ if $p \in (0, 1)$, assuming all expectations are well defined. Recall that $\mathbb{E}^N_{\theta,m}$ is an expectation w.r.t. the probability with finite-dimensional law (20), associated with the simulation of Algorithm 1.

Lemma A.1 *Assume (A1-2). Then for any $\theta \in \Theta$ there exists a $C < +\infty$ such that for any $(l, N, \varphi) \in \mathbb{N}_0 \times \mathbb{N} \times \mathcal{B}_b(\mathsf{X})$:*

$$\mathbb{E}^N_{\theta,m}[[\eta^{l,N}_\theta - \eta^l_\theta](\varphi)^2] \leq \frac{C\|\varphi\|^2_\infty}{N}$$

$$|\mathbb{E}^N_{\theta,m}[[\eta^{l,N}_\theta - \eta^l_\theta](\varphi)]| \leq \frac{C\|\varphi\|_\infty}{N}.$$

Proof The first statement is (Del Moral 2004, Theorem 7.4.4.) and the second follows easily from (e.g.) (Beskos et al. 2017b, eq. (A.2.), Lemma A.1.(iii)). □

Recall that we use $\mathbb{E}_\theta$ to denote expectation associated with the probability $\mathbb{P}_\theta$ in (21) of which is associated with the generation of Algorithm 2.

Lemma A.2 *Assume (A1-2). Then for any $\theta \in \Theta$ there exists a $C < +\infty$ such that for any $(l, p, \varphi) \in \mathbb{N}_0 \times \mathbb{N}_0 \times \mathcal{B}_b(\mathsf{X})$, $1 \leq N_0 < N_1 < \cdots < N_p < +\infty$:*

$$\mathbb{E}_\theta[[\eta_\theta^{l,N_{0:p}} - \eta_\theta^l](\varphi)^2] \leq \frac{C\|\varphi\|_\infty^2}{N_p}\Big(1 + \frac{p^2}{N_p}\Big).$$

Proof Follows by a similar approach to the proof of (Jasra et al. 2020, Proposition A.1.), which needs the results in Lemma A.1. □

Lemma A.3 *Assume (A1-2). Then for any $\theta \in \Theta$ there exists a $C < +\infty$ such that for any $(l, p, i) \in \mathbb{N}_0 \times \mathbb{N}_0 \times \{1, \ldots, d_\theta\}$, $1 \leq N_0 < N_1 < \cdots < N_p < +\infty$:*

$$\mathbb{E}_\theta\Big[\Big(\frac{\eta_\theta^{l,N_{0:p}}(G_\theta^l\{(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\})}{\eta_\theta^{l,N_{0:p}}(G_\theta^l)} - \frac{\eta_\theta^l(G_\theta^l\{(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\})}{\eta_\theta^l(G_\theta^l)}\Big)^2\Big] \leq \frac{C\|(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\|_\infty^2}{N_p}\Big(1 + \frac{p^2}{N_p}\Big).$$

Proof As

$$\frac{\eta_\theta^{l,N_{0:p}}(G_\theta^l\{(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\})}{\eta_\theta^{l,N_{0:p}}(G_\theta^l)} - \frac{\eta_\theta^l(G_\theta^l\{(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\})}{\eta_\theta^l(G_\theta^l)} = \frac{1}{\eta_\theta^{l,N_{0:p}}(G_\theta^l)}\Big([\eta_\theta^{l,N_{0:p}} - \eta_\theta^l](G_\theta^l\{(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\})\Big) + \frac{\eta_\theta^l(G_\theta^l\{(\varphi_\theta^{l+1})_i - (\varphi_\theta^l)_i\})}{\eta_\theta^{l,N_{0:p}}(G_\theta^l)\eta_\theta^l(G_\theta^l)}[\eta_\theta^l - \eta_\theta^{l,N_{0:p}}](G_\theta^l)$$

one can simply use the C_2 inequality, (A2) and Lemma A.2 to complete the proof. □

Lemma A.4 *Assume (A1-3). Then for any $\theta \in \Theta$ there exists a $C < +\infty$ such that for any $(l, p, i) \in \mathbb{N} \times \mathbb{N}_0 \times \{1, \ldots, d_\theta\}$, $1 \leq N_0 < N_1 < \cdots < N_p < +\infty$:*

$$\mathbb{E}_\theta\Bigg[\Bigg(\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}(\varphi_\theta^l)_i)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \eta_\theta^{l-1,N_{0:p}}((\varphi_\theta^{l-1})_i) - \Bigg(\frac{\eta_\theta^{l-1}(G_\theta^{l-1}(\varphi_\theta^l)_i)}{\eta_\theta^{l-1}(G_\theta^{l-1})} - \eta_\theta^{l-1}((\varphi_\theta^{l-1})_i)\Bigg)\Bigg)^2\Bigg] \leq \frac{C}{N_p}\Big(1+\frac{p^2}{N_p}\Big)\Big(\|(\varphi_\theta^l)_i - (\varphi_\theta^{l-1})_i\|_\infty^2 + \Big\|G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big\|_\infty^2\Big).$$

Proof We have the decomposition

$$\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}(\varphi_\theta^l)_i)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \eta_\theta^{l-1,N_{0:p}}((\varphi_\theta^{l-1})_i) - \Big(\frac{\eta_\theta^{l-1}(G_\theta^{l-1}(\varphi_\theta^l)_i)}{\eta_\theta^{l-1}(G_\theta^{l-1})} - \eta_\theta^{l-1}((\varphi_\theta^{l-1})_i)\Big) = \sum_{j=1}^3 T_j$$

where

$$T_1 = \frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}\{(\varphi_\theta^l)_i - (\varphi_\theta^{l-1})_i\})}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \frac{\eta_\theta^{l-1}(G_\theta^{l-1}\{(\varphi_\theta^l)_i - (\varphi_\theta^{l-1})_i\})}{\eta_\theta^{l-1}(G_\theta^{l-1})}$$

$$T_2 = -\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})}\eta_\theta^{l-1,N_{0:p}}\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big) + \frac{\eta_\theta^{l-1}(G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1}(G_\theta^{l-1})}\eta_\theta^{l-1}\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big)$$

$$T_3 = \eta_\theta^{l-1,N_{0:p}}\Big((\varphi_\theta^{l-1})_i\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big)\Big) - \eta_\theta^{l-1}\Big((\varphi_\theta^{l-1})_i\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big)\Big).$$

Thus, one can apply the C_2 inequality twice and deal individually with the terms $\sum_{j=1}^3 \mathbb{E}_\theta[T_j^2]$. For $\mathbb{E}_\theta[T_1^2]$, one can use Lemma A.3. For $\mathbb{E}_\theta[T_3^2]$, one can use Lemma A.2. So to conclude, we consider $\mathbb{E}_\theta[T_2^2]$. We have

$$T_2 = T_4 + T_5$$

where

$$T_4 = -\Big(\frac{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})} - \frac{\eta_\theta^{l-1}(G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1}(G_\theta^{l-1})}\Big) \times \eta_\theta^{l-1,N_{0:p}}\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big)$$

$$T_5 = \frac{\eta_\theta^{l-1}(G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1}(G_\theta^{l-1})}\Big(\eta_\theta^{l-1}\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big) - \eta_\theta^{l-1,N_{0:p}}\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^l} - 1\Big)\Big).$$

Applying the C_2 inequality once again allows one to consider just $\mathbb{E}_\theta[T_4^2]$ and $\mathbb{E}_\theta[T_5^2]$ individually. For $\mathbb{E}_\theta[T_5^2]$, one can use (A3) and Lemma A.2. As

$$T_4 = -\Big(\frac{[\eta_\theta^{l-1,N_{0:p}} - \eta_\theta^{l-1}](G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})}$$

$$-\frac{\eta_\theta^{l-1}(G_\theta^{l-1}(\varphi_\theta^{l-1})_i)}{\eta_\theta^{l-1}(G_\theta^{l-1})\eta_\theta^{l-1,N_{0:p}}(G_\theta^{l-1})}[\eta_\theta^{l-1}-\eta_\theta^{l-1,N_{0:p}}](G_\theta^{l-1})\Big)$$
$$\times\eta_\theta^{l-1,N_{0:p}}\Big(G_\theta^{l-1}\frac{Z_\theta^{l-1}}{Z_\theta^{l}}-1\Big).$$

One can then conclude the result by applying the C_2 inequality and using (A2) and Lemma A.2. □

Appendix B Explicit solution of the toy model in Sect. 5.1

The un-normalized target is given by

$$\gamma_\theta(u)=\theta^{\frac{M}{2}}\cdot\exp\Big\{-\frac{\theta}{2}\|Gu-y\|^2\Big\}\cdot\frac{1}{\theta}\exp\Big\{-\frac{(\log(\theta))^2}{2}\Big\},$$

and the marginal is

$$\begin{aligned}Z_\theta &= \int_{-1}^{1}\gamma(u,\theta)\mathrm{d}u\\
&=\theta^{\frac{m-2}{2}}\exp\Big\{-\frac{(\log(\theta))^2}{2}\Big\}\int_{-1}^{1}\exp\Big\{-\frac{\theta}{2}\|Gu-y\|^2\Big\}\mathrm{d}u\\
&=\theta^{\frac{m-2}{2}}\exp\Big\{-\frac{\theta}{2}\Big(\|y\|^2-\frac{(G^\intercal y)^2}{\|G\|^2}\Big)-\frac{(\log(\theta))^2}{2}\Big\}\\
&\quad\times\int_{-1}^{1}\exp\Big\{-\frac{\theta\|G\|^2}{2}\Big(u-\frac{G^\intercal y}{\|G\|^2}\Big)^2\Big\}\mathrm{d}u\\
&=\theta^{\frac{m-3}{2}}\frac{\sqrt{\pi/2}}{\|G\|}\exp\Big\{-\frac{\theta}{2}\Big(\|y\|^2-\frac{(G^\intercal y)^2}{\|G\|^2}\Big)-\frac{(\log(\theta))^2}{2}\Big\}\\
&\quad\Big(\mathrm{erf}\Big(\sqrt{\frac{\theta}{2}}\|G\|(1-\frac{G^\intercal y}{\|G\|^2})\Big)-\\
&\quad\mathrm{erf}\Big(\sqrt{\frac{\theta}{2}}\|G\|(-1-\frac{G^\intercal y}{\|G\|^2})\Big)\Big).\end{aligned}$$

The logarithm is given by

$$\begin{aligned}\log\big(Z_\theta\big) &= \frac{m-3}{2}\log(\theta)-\frac{\theta}{2}\Big(\|y\|^2-\frac{(G^\intercal y)^2}{\|G\|^2}\Big)\\
&\quad-\frac{(\log(\theta))^2}{2}+\log\Big(\mathrm{erf}\Big(\sqrt{\frac{\theta}{2}}\|G\|(1-\frac{G^\intercal y}{\|G\|^2})\Big)-\\
&\quad\mathrm{erf}\Big(\sqrt{\frac{\theta}{2}}\|G\|(-1-\frac{G^\intercal y}{\|G\|^2})\Big)\Big)+C,\end{aligned}$$

and the derivative of the logarithm is

$$\begin{aligned}&\frac{\partial\log\big(Z_\theta\big)}{\partial\theta}\\
&=\frac{m-3}{2\theta}-\frac{\Big(\|y\|^2-\frac{(G^\intercal y)^2}{\|G\|^2}\Big)}{2}-\frac{\log(\theta)}{\theta}\end{aligned}$$

$$\begin{aligned}&+\frac{1}{\mathrm{erf}\Big(\sqrt{\frac{\theta}{2}}\|G\|(1-\frac{G^\intercal y}{\|G\|^2})\Big)-\mathrm{erf}\Big(\sqrt{\frac{\theta}{2}}\|G\|(-1-\frac{G^\intercal y}{\|G\|^2})\Big)}\cdot\\
&\frac{2}{\sqrt{\pi}}\Big(\exp\Big\{-\frac{\theta\|G\|^2}{2}(1-\frac{G^\intercal y}{\|G\|^2})^2\Big\}\frac{\|G\|(1-\frac{G^\intercal y}{\|G\|^2})}{2\sqrt{2\theta}}-\\
&\quad\exp\Big\{-\frac{\theta\|G\|^2}{2}(-1-\frac{G^\intercal y}{\|G\|^2})^2\Big\}\frac{\|G\|(-1-\frac{G^\intercal y}{\|G\|^2})}{2\sqrt{2\theta}}\Big).\end{aligned}$$

References

Agapiou, S., Roberts, G.O., Vollmer, S.: Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models. Bernoulli **24**, 1726–1786 (2018)

Benveniste, A., Métivier, M., Priouret, P.: Adaptive Algorithms and Stochastic Approximation. Springer-Verlag, New York (1990)

Beskos, A., Jasra, A., Law, K.J.H., Marzouk, Y., Zhou, Y.: Multilevel sequential Monte Carlo with dimension-independent likelihood-informed proposals. SIAM/ASA J. Uncertain. Quantif. **6**(2), 762–786 (2017)

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

Brenner, S., Scott, R.: The Mathematical Theory of Finite Element Methods. Springer, New York (2007)

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

Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. SIAM, Philadelphia (2002)

Dashti, M., Stuart, A.M.: Uncertainty quantification and weak approximation of an elliptic inverse problem. SIAM J. Numer. Anal. **49**(6), 2524–2542 (2011)

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

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

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

Del Moral, P., Jasra, A., Law, K.J.H.: Multilevel sequential Monte Carlo: Mean square error bounds under verifiable conditions. Stoch. Anal. Appl. **35**(3), 478–498 (2017)

Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Springer, New York (1996)

Franklin, J.N.: Well-posed stochastic extensions of ill-posed linear problems. J. Math. Anal. App. **31**(3), 682–716 (1970)

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., Richtarik, P.: SGD: General analysis and improved rates. In: *Proceedings of the 36th International Conference on Machine Learning, in PMLR* **97**, 5200-5209, (2019)

Jasra, A., Law, K.J.H., Yu, F.: Unbiased filtering of a class of partially observed diffusions. arXiv preprint, (2020)

Kushner, H., Yin, G.G.: Stochastic approximation and recursive algorithms and applications, vol. 35. Springer Science & Business Media, Berlin/Heidelberg (2003)

Le Gland, F., Mevel, M.: Recursive identification in hidden Markov models. In: *Proceedings of the 36th IEEE Conference on Decision and Control*, pp. 3468-3473, (1997)

McLeish, D.: A general method for debiasing a Monte Carlo estimator. Monte Carlo Meth. Appl. **17**, 301–315 (2011)

Rhee, C.H., Glynn, P.: Unbiased estimation with square root convergence for SDE models. Oper. Res. **63**, 1026–1043 (2015)

Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010)

Tadic, V., Doucet, A.: Asymptotic properties of recursive maximum likelihood estimators in non-linear state-space models. arXiv preprint, (2018)

Tikhonov, A.N., Glasko, V.B.: The approximate solution of Fredholm integral equations of the first kind. USSR Comput. Math. Math. Phys. **4**(3), 236–247 (1964)

Vihola, M.: Unbiased estimators and multilevel Monte Carlo. Oper. Res. **66**, 448–462 (2018)