Efficient Debiased Evidence Estimation by Multilevel Monte Carlo Sampling

Kei Ishikawa[1] **Takashi Goda**[2]

[1]ETH Zurich, Switzerland
[2]The University of Tokyo, Japan

Preliminary version

Abstract

In this paper, we propose a new stochastic optimization algorithm for Bayesian inference based on multilevel Monte Carlo (MLMC) methods. In Bayesian statistics, biased estimators of the model evidence have been often used as stochastic objectives because the existing debiasing techniques are computationally costly to apply. To overcome this issue, we apply an MLMC sampling technique to construct low-variance unbiased estimators both for the model evidence and its gradient. In the theoretical analysis, we show that the computational cost required for our proposed MLMC estimator to estimate the model evidence or its gradient with a given accuracy is an order of magnitude smaller than those of the previously known estimators. Our numerical experiments confirm considerable computational savings compared to the conventional estimators. Combining our MLMC estimator with gradient-based stochastic optimization results in a new scalable, efficient, debiased inference algorithm for Bayesian statistical models.

1 INTRODUCTION

In empirical Bayes estimation, the model evidence (or, the log marginal likelihood) is maximized to estimate parameters. As such, the evidence maximization is considered a fundamental problem in Bayesian statistics and has been studied extensively for a long time. Perhaps the most common approach for the evidence maximization would be the expectation-maximization (EM) algorithm (Dempster et al., 1977). In the EM algorithm, the analytical form of the posterior distribution given data and parameters is required. In the case where the exact posterior distribution is not available, the algorithm can be extended by various approximation techniques such as variational EM algorithm (Jordan et al., 1999) and Monte Carlo EM algorithm (Wei and Tanner, 1990), which maximized the lower bound of the evidence. However, such approximation methods usually maximize a lower bound of the model evidence and thus the resulting estimates are biased.

To reduce the bias from evidence estimation, application of importance sampling (Robert and Casella, 2013) is often considered. The estimate of the model evidence obtained by importance sampling is known to have a negative bias as discussed in Section 2.2, thus it serves as a stochastic lower bound of the true model evidence.

This stochastic lower bound can be maximized for the empirical Bayes method, by using its gradient with respect to the model parameters in stochastic optimization (Robbins and Monro, 1951). Especially, if the size of the data is too large to compute the gradient of the objective for all data points in each iteration, we can randomly pick a subset of the data to carry out doubly stochastic optimization. However, to reduce the bias of the objective, the number of Monte Carlo samples required to compute the gradient for each data point needs to get larger, leading to a computational inefficiency in the optimization of the objective based on importance sampling.

We tackle this problem by using a sophisticated Monte Carlo simulation technique called the multilevel Monte Carlo (MLMC) method. Although the MLMC was originally studied in the context of parametric integration (Heinrich, 1998) and stochastic differential equations (Giles, 2008), it can be applied to many other contexts as well, in situations where the computational cost per Monte Carlo sample increases as we reduce the bias of an objective. By considering a hierarchy of different bias levels from a true objective and constructing a tightly coupled Monte Carlo estimator for the difference between two successive biased objectives, the true objective can be estimated quite efficiently compared to the standard Monte Carlo method which only estimates a single biased objective at a fixed bias level. In a favorable setting, the MLMC estimator can be even made unbiased for

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

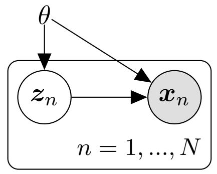


Figure 1: A graphical model of Bayesian models with local latent variables.

the true objective by some randomization (Rhee and Glynn, 2015).

After the seminal work by Giles (2008), the MLMC has been applied to various areas such as partial differential equations with random coefficients (Cliffe et al., 2011), continuous-time Markov chains (Anderson and Higham, 2012) and Markov Chain Monte Carlo sampling (Dodwell et al., 2015). We refer the reader to a review on recent developments of the MLMC (Giles, 2015). Recently, the MLMC has been studied intensively for an efficient estimation of nested expectations, motivated by various applications such as computational finance (Bujok et al., 2015), Bayesian computation (Giles et al., 2016), decision-making under uncertainty (Giles and Goda, 2019; Hironaka et al., 2020) and experimental design (Goda et al., 2020; Goda et al., 2021). In the appendix, we give a brief introduction to the MLMC method and its basic theory.

In this paper, the above-cited works on nested expectations are leveraged to obtain a computationally efficient, debiased estimator for the model evidence. We also provide an efficient, debiased estimator for the gradient of the model evidence with respect to the model parameters, which can be combined well with stochastic optimization to search for a good estimate of the parameters learned from a large data set.

2 BACKGROUND

2.1 PROBLEM SETTINGS

In this paper, we consider Bayesian statistical models with local latent variables that are formulated by the following i.i.d. data generating process:

$$\begin{aligned} \boldsymbol{z}_n &\sim p_\theta(z), \\ \boldsymbol{x}_n | \boldsymbol{z}_n = z_n &\sim p_\theta(x|z_n), \end{aligned} \tag{1}$$

for $n = 1, ..., N$. Here, the bold letters $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$ denote random variables, whereas the normal letter x_n denotes the corresponding realization. This latent variable model can be expressed as a graphical model shown in Figure 1. The problem we are interested in is to estimate the parameter θ from the data $x_1, ..., x_N$.

Throughout the paper, we will omit the dependence of the model p_θ on the parameter θ, and simply write p instead of p_θ where it is obvious from the context. We also abbreviate a vector such as $(x_1, ..., x_n)$ as $x_{1:n}$ for the simplicity of notation. To estimate the parameter θ, we maximize the model evidence of the data $x_1, ..., x_N$, which is defined by

$$\begin{aligned} \mathcal{L}(x_{1:N}; \theta) &= \log p_\theta(x_{1:N}) \\ &= \sum_{n=1}^{N} \log \int p_\theta(x_n|z_n) p_\theta(z_n) \, \mathrm{d}z_n. \end{aligned} \tag{2}$$

Additionally, we assume that the size N of the data is so large that stochastic or mini-batch optimization of the above objective is more desirable than batch optimization. Our aim of this paper is to introduce efficient, debiased Monte Carlo estimators of the model evidence (2) and its gradient with respect to θ and then to propose a new scalable, stochastic optimization algorithm of this objective.

Remark 1. Actually, it is possible to treat the parameter θ in a Bayesian manner and obtain a similar debiased, efficient estimator using MLMC. Bayesian treatment of parameters enables us to quantify the uncertainty of the estimated parameters. To calculate the debiased posterior, we can combine a gradient-based variant of the stochastic variational inference (Hoffman et al., 2013) with our proposed algorithm. Such a combination can be implemented very simply with almost no additional effort. The details on this point are discussed in the appendix.

2.2 NESTED MONTE CARLO ESTIMATION OF MODEL EVIDENCE AND ITS GRADIENT

As discussed before, the model evidence can be estimated by importance sampling. The estimator of the model evidence by importance sampling, denoted here by $\hat{\mathcal{L}}_K$, is given by

$$\hat{\mathcal{L}}_K(x_{1:N}) = \sum_{n=1}^{N} \log \left[\frac{1}{K} \sum_{k=1}^{K} \frac{p(x_n|Z_{n,k}) p(Z_{n,k})}{q_n(Z_{n,k})} \right],$$

where, for each n, $Z_{n,1}, ..., Z_{n,K}$ are i.i.d. random samples from a proposal distribution $q_n(z_n)$. In general, q_n is taken so that it approximates the true posterior distribution of $\boldsymbol{z}_n$ given $\boldsymbol{x}_n = x_n$. This is because the Monte Carlo average inside the logarithm becomes exactly equal to the marginal distribution $p(x_n)$ with variance 0, if we can set $q_n(z_n)$ to the conditional distribution $p(z_n|\boldsymbol{x}_n = x_n)$. In practice, we choose the proposal distribution q_n as an approximate posterior of $\boldsymbol{z}_n$ computed from x_n. So hereafter we will write $q(z_n; x_n)$ instead of $q_n(z_n)$ to express the proposal distribution of Z_n's.

There are a few useful properties of this estimator for the model evidence based on importance sampling, as proven in Theorem 1 of Burda et al. (2016). First, when we increase K, the number of the samples in the Monte Carlo average inside the logarithm, to infinity, we recover the model evidence thanks to the law of large numbers, i.e., we have

$$\lim_{K \to \infty} \hat{\mathcal{L}}_K(x_{1:N}) = \mathcal{L}(x_{1:N}).$$

Preliminary version

Second, when we denote the expectation of the estimator $\hat{\mathcal{L}}_K(x_{1:N})$ by $\mathcal{L}_K = \mathbb{E}[\hat{\mathcal{L}}_K(x_{1:N})]$, $\mathcal{L}_K$ is always smaller than or equal to $\mathcal{L}_{K+1}$. That is, we have

$$\mathcal{L}_1 \leq \cdots \leq \mathcal{L}_K \leq \mathcal{L}_{K+1} \leq \cdots \leq \mathcal{L}_\infty = \mathcal{L}(x_{1:N}), \quad (3)$$

which implies that the larger K we use, the better lower bound on the model evidence we obtain. Thus, the maximization of this lower bound with respect to θ is a good approximation of the evidence maximization if K is chosen large enough.

In order to process a large data set efficiently, it is sensible to apply a gradient-based doubly stochastic optimization. For this, instead of looking at all the data points at each iteration, we randomly pick a subset (mini-batch) of the data points with size M, and estimate the corresponding log-marginal likelihood (or its gradient). This is equivalent to rewriting the model evidence (2) into a *nested expectation*

$$\mathcal{L}(x_{1:N}) = N\mathbb{E}_X\left[\log \int p_\theta(X|Z)p_\theta(Z)\,\mathrm{d}Z\right],$$

where X is a random variable taking $x_1, \ldots, x_N$ uniformly and $\mathbb{E}_X$ denotes the average with respect to X, and to estimate it by the nested Monte Carlo method:

$$\hat{\mathcal{L}}_{M,K} = \frac{N}{M}\sum_{m=1}^{M} \log\left[\frac{1}{K}\sum_{k=1}^{K} \frac{p(X_m|Z_{m,k})p(Z_{m,k})}{q(Z_{m,k};X_m)}\right]. \quad (4)$$

Here, it is important to note that $X_1, \ldots, X_M$ denote i.i.d. random samples of X. For each sample X_m, $Z_{m,1}, \ldots, Z_{m,K}$ are conditionally i.i.d. random samples from a proposal distribution $q(z_m; X_m)$. Because of the linearity of expectation, we have $\mathbb{E}[\hat{\mathcal{L}}_{M,K}] = \mathbb{E}[\hat{\mathcal{L}}_{1,K}] = \mathcal{L}_K \leq \mathcal{L}(x_{1:N})$.

Later in the paper, when we have only one sample of X, we write $\hat{\mathcal{L}}_{1,K}$ as $\hat{\mathcal{L}}_K := \hat{\mathcal{L}}_K(X_1, Z_{1,1:K})$ for notational simplicity.

The gradient of the model evidence with respect to θ can be estimated by the gradient of the nested Monte Carlo estimator (4), which is explicitly given by

$$\begin{aligned}\nabla_\theta \hat{\mathcal{L}}_{M,K} &= \frac{N}{M}\sum_{m=1}^{M} \nabla_\theta \log\left[\frac{1}{K}\sum_{k=1}^{K} \frac{p(X_m|Z_{m,k})p(Z_{m,k})}{q(Z_{m,k};X_m)}\right] \\ &= \frac{N}{M}\sum_{m=1}^{M} \frac{\frac{1}{K}\sum_{k=1}^{K} \frac{\nabla_\theta\left(p(X_m|Z_{m,k})p(Z_{m,k})\right)}{q(Z_{m,k};X_m)}}{\frac{1}{K}\sum_{k=1}^{K} \frac{p(X_m|Z_{m,k})p(Z_{m,k})}{q(Z_{m,k};X_m)}}.\end{aligned}$$

Note that, by using automatic differentiation software, we can avoid coding the gradient in the presented form. We would emphasize here that $\nabla_\theta \hat{\mathcal{L}}_{M,K}$ is a biased estimator of the gradient of the true model evidence, and so, applying a gradient-based doubly stochastic optimization based on $\nabla_\theta \hat{\mathcal{L}}_{M,K}$ does not converge to the optimal parameter θ in general.

Let us focus on the nested Monte Carlo estimation of the model evidence at this moment. A similar argument applies to the gradient of the model evidence. The mean squared error of $\hat{\mathcal{L}}_{M,K}$ is decomposed into the sum of the variance and the squared bias:

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\mathcal{L}}_{M,K} - \mathcal{L}(x_{1:N})\right)^2\right] \\ &= \mathbb{V}\left[\hat{\mathcal{L}}_{M,K}\right] + \left(\mathbb{E}\left[\hat{\mathcal{L}}_{M,K}\right] - \mathcal{L}(x_{1:N})\right)^2 \\ &= \frac{\mathbb{V}[\hat{\mathcal{L}}_{1,K}]}{M} + \left(\mathcal{L}_K - \mathcal{L}(x_{1:N})\right)^2.\end{aligned}$$

Therefore, in order to make the mean squared error small, we need to increase both the mini-batch size M and the number of inner Monte Carlo samples K, so that the variance and the bias become small, respectively. Here, it follows from (3) that the bias converges to 0 as K approaches infinity. More precisely, in order to estimate the model evidence with a mean squared accuracy ε^2, it suffices to have

$$\frac{\mathbb{V}[\hat{\mathcal{L}}_{1,K}]}{M} \leq \frac{\varepsilon^2}{2} \quad \text{and} \quad \left(\mathcal{L}_K - \mathcal{L}(x_{1:N})\right)^2 \leq \frac{\varepsilon^2}{2}.$$

Assuming that $\mathbb{V}[\hat{\mathcal{L}}_{1,K}] \approx \mathbb{V}[\hat{\mathcal{L}}_{1,\infty}]$ for large enough K and that the bias $|\mathcal{L}_K - \mathcal{L}(x_{1:N})|$ decays with the order $K^{-\alpha}$ for some $\alpha > 0$, we need to set $M = O(\varepsilon^{-2})$ and $K = O(\varepsilon^{-1/\alpha})$, respectively. Since the computational cost of $\hat{\mathcal{L}}_{M,K}$ is given by the product $M \times K$, it is of $O(\varepsilon^{-2-1/\alpha})$. Although another balancing between the variance and the squared bias is possible, the cost of $O(\varepsilon^{-2-1/\alpha})$ cannot be improved by nested Monte Carlo methods.

2.3 RELATED METHODS

As discussed above, the nested Monte Carlo estimation is computationally inefficient. Thus, there have been some attempts to improve the efficiency of the debiased estimation for the model evidence and its gradient. In the context of variational autoencoder (VAE) (Kingma and Welling, 2014), the use of the nested Monte Carlo estimator has been actively studied (Burda et al., 2016) as it naturally extends the ELBO, the original objective of the VAE. As a more efficient variant of the nested Monte Carlo objective, applications of the Jackknife method and the Russian roulette estimator were proposed.

The Jackknife method is a bias removal method in statistics. It uses resampling techniques to remove low order bias, e.g., the bias of the first order Jackknife estimator becomes $O(n^{-2})$ as the $O(n^{-1})$ bias can be removed. Nowozin (2018) applied this idea to the estimation of the model evidence and its gradient of VAE.

Preliminary version

More related to our work is a Russian roulette estimator introduced in Luo et al. (2020). Their estimator, called SUMO, randomly picks a positive integer $\mathcal{K}$ from distribution

$$P(k \leq \mathcal{K}) = \begin{cases} 1/k & \text{if } k < a \\ 1/a \cdot (1-0.1)^{k-a} & \text{if } k \geq a, \end{cases}$$

for which $a = 80$ is recommended in the paper. Here, the exponential decay of $\mathbb{P}(k \leq \mathcal{K})$ for k larger than a serves as a soft truncation of the $\mathcal{K}$ as the $\mathcal{K}$ sufficiently larger than the a cannot be sampled with high probability. This random sampling is applied to each data point X and the SUMO outputs the following weighted sum:

$$\hat{\mathcal{L}}_a^{\text{SUMO}}(X) = \sum_{k=1}^{\infty} \frac{\mathbb{1}_{\{k \leq \mathcal{K}\}}}{\mathbb{P}(k \leq \tilde{\mathcal{K}})} [\mathcal{L}_k \widehat{- \mathcal{L}_{k}}_{-1}](X),$$

where $\tilde{\mathcal{K}}$ is distributed identically to $\mathcal{K}$. The differences in the summation are defined by $[\mathcal{L}_K \widehat{- \mathcal{L}_{K}}_{-1}](X) = \hat{\mathcal{L}}_K(X; Z_{1:K}) - \hat{\mathcal{L}}_{K-1}(X; Z_{1:(K-1)})$, where we explicitly wrote the dependence of $\hat{\mathcal{L}}_K(x; Z_{1:K}) = \log\left[\frac{1}{K}\sum_{k=1}^{K} \frac{p(x|Z_k)p(Z_k)}{q(Z_k;x)}\right]$ on $Z_{1:K}$ and $\hat{\mathcal{L}}_0$ is defined as 0. The function $\mathbb{1}_{\{k \leq \tilde{\mathcal{K}}\}}$ is the indicator function. Having defined the point-wise definition of the SUMO, the SUMO for mini-batch can simply defined as the following sum: $\hat{\mathcal{L}}_a^{\text{SUMO}}(X_{1:M}) = \sum_{m=1}^{M} \hat{\mathcal{L}}_a^{\text{SUMO}}(X_m)$.

It should be also noted that this estimator is similar to the randomized MLMC estimator discussed in the next section, in that they both use estimators of differences with shared inner Monte Carlo samples. However, even though the SUMO attempted to construct an unbiased estimator of the model evidence, a truly unbiased estimation is infeasible in the sense that both the expected computational cost and the variance of the SUMO approach infinity as the bias tends to zero (Luo et al., 2020). Our estimator using the MLMC method, on the other hand, can completely remove the bias while requiring finite expected computational cost and having a bounded variance. We refer the reader to the appendix for a detailed theoretical comparison between different estimators.

Aside from Jackknife method and SUMO, there are several works on unbiased estimation of the model evidence that are based on Markov chain Monte Carlo method (MCMC) instead of nested Monte Carlo method (Ruiz et al., 2020; Rischard et al., 2018; Wei and Murray, 2017). Though their approaches are different from ours, they share some of the key ideas such as the random truncation of telescoping sum decomposition and the coupling of Monte Carlo samples.

3 PROPOSED ALGORITHM

To reduce the necessary computational cost from $O(\varepsilon^{-2-1/\alpha})$ to $O(\varepsilon^{-2})$ for estimating the model evidence, we apply the MLMC methods. Later in this section, we also discuss the case for the gradient of the model evidence.

The main difference from the nested Monte Carlo estimation is to consider a geometric hierarchy of the biased objectives $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_{2^\ell}, \ldots$ and to represent the model evidence by the following telescoping sum

$$\mathcal{L}(x_{1:N}) = \sum_{\ell=0}^{\infty} (\mathcal{L}_{2^\ell} - \mathcal{L}_{2^{\ell-1}}),$$

where we set $\mathcal{L}_{2^{-1}} \equiv 0$. We call the term with $\ell = 0$ *main term* and the remaining terms *correction terms*. Note that truncating the infinite sum over ℓ up to the first L terms yields the objective $\mathcal{L}_{2^L}$. The nested Monte Carlo method estimates the single term $\mathcal{L}_{2^L}$ only, whereas the MLMC method estimates the main term and the correction terms (up to level L) independently and sums them up.

The key ingredient is in how to estimate the correction terms. In order to estimate $\mathcal{L}_{2^\ell}$ by $\hat{\mathcal{L}}_{M,2^\ell}$ for some mini-batch size M, we generate 2^ℓ i.i.d. samples of Z_m from a proposal distribution $q(z_m; X_m)$ for $m = 1, \ldots, M$. Here, for the same mini-batch, two halves of the i.i.d. samples of Z_m can be used to compute $\hat{\mathcal{L}}_{M,2^{\ell-1}}$ twice, which we denote by $\hat{\mathcal{L}}^{(a)}_{M,2^{\ell-1}}$ and $\hat{\mathcal{L}}^{(b)}_{M,2^{\ell-1}}$, respectively. Defining

$$\Delta\hat{\mathcal{L}}_{M,2^\ell} = \begin{cases} \hat{\mathcal{L}}_{M,1} & \text{if } \ell = 0, \\ \hat{\mathcal{L}}_{M,2^\ell} - \dfrac{\hat{\mathcal{L}}^{(a)}_{M,2^{\ell-1}} + \hat{\mathcal{L}}^{(b)}_{M,2^{\ell-1}}}{2} & \text{otherwise,} \end{cases}$$

the linearity of expectation ensures that

$$\mathbb{E}\left[\Delta\hat{\mathcal{L}}_{M,2^\ell}\right] = \mathcal{L}_{2^\ell} - \mathcal{L}_{2^{\ell-1}}.$$

This means that $\Delta\hat{\mathcal{L}}_{M,2^\ell}$ is an unbiased estimator for the correction term. Now the truncated telescoping sum is estimated by

$$\hat{\mathcal{L}}^{\text{MLMC}}_{2^L} := \sum_{\ell=0}^{L} \Delta\hat{\mathcal{L}}_{M_\ell,2^\ell}, \tag{5}$$

for mini-batch sizes $M_0, \ldots, M_L > 0$. This is our MLMC estimator for the model evidence, and we refer the readers to Algorithm 1 for its summary.

We give a heuristic explanation on why our MLMC estimator is more efficient than the nested Monte Carlo estimator. It is easy to see that the computational cost and the variance of (5) are given by

$$\sum_{\ell=0}^{L} M_\ell 2^\ell \quad \text{and} \quad \sum_{\ell=0}^{L} \frac{\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]}{M_\ell},$$

respectively. Because of the shared use of i.i.d. samples of Z_m in computing $\hat{\mathcal{L}}^{(a)}_{M,2^{\ell-1}}$, $\hat{\mathcal{L}}^{(b)}_{M,2^{\ell-1}}$ and $\hat{\mathcal{L}}_{M,2^\ell}$, the difference $\Delta\hat{\mathcal{L}}_{M,2^\ell}$ is expected quite small in magnitude, particularly for large levels ℓ. In fact, by Theorem 2 and 3,

Preliminary version

Algorithm 1 MLMC estimation of $\mathcal{L}_{2^L} = \mathbb{E}[\hat{\mathcal{L}}_{1,2^L}]$

```
for $m = 1, ..., M_0$ do
    sample $X_m$ randomly from $x_1, ..., x_N$
    sample $Z_m \sim q(z_m; X_m)$
end for
$\Delta\hat{\mathcal{L}}_{M_0,1} \leftarrow \frac{N}{M_0}\sum_{m=1}^{M_0} \log\left(\frac{p(X_m|Z_m)p(X_m)}{q(Z_m;X_m)}\right)$
for $\ell = 1, ..., L$ do
    for $m = 1, ..., M_\ell$ do
        (re-)sample $X_m$ randomly from $x_1, ..., x_N$
        for $k = 1, ..., 2^\ell$ do
            (re-)sample $Z_{m,k} \sim q(z_m; X_m)$
        end for
        $\hat{\mathcal{L}}^{(a)}_{2^{\ell-1}} \leftarrow N \cdot \log\left[\frac{1}{2^{\ell-1}}\sum_{k=1}^{2^{\ell-1}} \frac{p(X_m|Z_{m,k})p(Z_{m,k})}{q(Z_{m,k};X_m)}\right]$
        $\hat{\mathcal{L}}^{(b)}_{2^{\ell-1}} \leftarrow N \cdot \log\left[\frac{1}{2^{\ell-1}}\sum_{k=2^{\ell-1}+1}^{2^\ell} \frac{p(X_m|Z_{m,k})p(Z_{m,k})}{q(Z_{m,k};X_m)}\right]$
        $\hat{\mathcal{L}}_{2^\ell} \leftarrow N \cdot \log\left[\frac{1}{2^\ell}\sum_{k=1}^{2^\ell} \frac{p(X_m|Z_{m,k})p(Z_{m,k})}{q(Z_{m,k};X_m)}\right]$
        $\Delta_\ell\hat{\mathcal{L}}_m \leftarrow \hat{\mathcal{L}}_{2^\ell} - \frac{1}{2}\left(\hat{\mathcal{L}}^{(a)}_{2^{\ell-1}} + \hat{\mathcal{L}}^{(b)}_{2^{\ell-1}}\right)$
    end for
    $\Delta\hat{\mathcal{L}}_{M_\ell,2^\ell} \leftarrow \frac{1}{M_\ell}\sum_{m=1}^{M_\ell} \Delta_\ell\hat{\mathcal{L}}_m$
end for
$\hat{\mathcal{L}}^{\text{MLMC}}_{2^L} \leftarrow \sum_{\ell=0}^{L} \Delta\hat{\mathcal{L}}_{M_\ell,2^\ell}$
```

we can assume that the variance per one randomly chosen data point, i.e., $\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]$, decays exponentially fast with respect to ℓ:

$$\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,1}] \gg \mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2}] \gg \cdots \gg \mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}] \gg \cdots .$$

Thus, in order to estimate higher-level correction terms accurately so that

$$\sum_{\ell=0}^{L} \frac{\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]}{M_\ell} \leq \frac{\varepsilon^2}{2}$$

is satisfied, we can decrease mini-batch sizes M_ℓ exponentially fast:

$$M_0 \gg M_1 \gg \cdots \gg M_\ell \gg \cdots .$$

This leads to a substantial saving of the required total computational cost as compared to the nested Monte Carlo method.

Let us assume that $\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]$ decays with the order of $2^{-\beta\ell}$ for some $\beta > 0$. The method of Lagrange multipliers leads to an optimal allocation of mini-batch sizes $M_0, M_1, \ldots, M_L$ by minimizing the total cost with keeping the variance bounded by $\varepsilon^2/2$:

$$\sum_{\ell=0}^{L} M_\ell 2^\ell + \lambda\left(\sum_{\ell=0}^{L} \frac{\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]}{M_\ell} - \frac{\varepsilon^2}{2}\right).$$

It is an easy exercise to obtain

$$M_\ell \propto \sqrt{\frac{\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]}{2^\ell}} = O(2^{-(\beta+1)\ell/2}).$$

If $\beta > 1$ holds, the terms with small ℓ are dominant in the sum $\sum_{\ell=0}^{L} M_\ell 2^\ell$, whereas, if $\beta < 1$ holds, the terms with large ℓ are dominant. For the dividing case $\beta = 1$, all the terms are approximately equal.

Remark 2. For the case $\beta > 1$, our MLMC estimator can be even made unbiased by applying a randomization technique from Rhee and Glynn (2015). For any sequence $\boldsymbol{\omega} = (\omega_0, \omega_1, \ldots)$ such that $\omega_\ell > 0$ and $\|\boldsymbol{\omega}\|_1 = 1$, it is possible to represent the model evidence by the weighted telescoping sum

$$\mathcal{L}(x_{1:N}) = \sum_{\ell=0}^{\infty} \omega_\ell \frac{\mathcal{L}_{2^\ell} - \mathcal{L}_{2^{\ell-1}}}{\omega_\ell} = \sum_{\ell=0}^{\infty} \omega_\ell \frac{\mathbb{E}\left[\Delta\hat{\mathcal{L}}_{1,2^\ell}\right]}{\omega_\ell},$$

For a mini-batch size $M > 0$, let $\ell^{(1)}, \ldots, \ell^{(M)} \geq 0$ be i.i.d. random samples from a discrete distribution with probabilities $\omega_0, \omega_1, \ldots$. Then the *randomized* MLMC estimator

$$\frac{1}{M}\sum_{m=1}^{M} \frac{\Delta\hat{\mathcal{L}}_{1,2^{\ell^{(m)}}}}{\omega_{\ell^{(m)}}}$$

becomes an unbiased estimator of $\mathcal{L}(x_{1:N})$. The expected computational cost and the variance per one data point from the mini-batch are given by

$$\sum_{\ell=0}^{\infty} \omega_\ell 2^\ell \quad \text{and} \quad \sum_{\ell=0}^{\infty} \frac{\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]}{\omega_\ell},$$

respectively. In order for these quantities to be both finite, it suffices to set $\omega_\ell \propto 2^{-(\beta+1)\ell/2}$. Such a discrete probability distribution does not exist if $\beta \leq 1$.

We now come to estimation of the gradient of the model evidence. Similarly to the model evidence itself, we represent the gradient by the telescoping sum

$$\nabla_\theta \mathcal{L}(x_{1:N}) = \sum_{\ell=0}^{\infty} \left(\nabla_\theta \mathcal{L}_{2^\ell} - \nabla_\theta \mathcal{L}_{2^{\ell-1}}\right).$$

The correction terms can be estimated by

$$\nabla_\theta \Delta\hat{\mathcal{L}}_{M,2^\ell} = \nabla_\theta \hat{\mathcal{L}}_{M,2^\ell} - \frac{\nabla_\theta \hat{\mathcal{L}}^{(a)}_{M,2^{\ell-1}} + \nabla_\theta \hat{\mathcal{L}}^{(b)}_{M,2^{\ell-1}}}{2},$$

which is unbiased. In this way the truncated telescoping sum for the gradient is estimated by

$$\nabla_\theta \hat{\mathcal{L}}^{\text{MLMC}}_{2^L} := \sum_{\ell=0}^{L} \nabla_\theta \Delta\hat{\mathcal{L}}_{M_\ell,2^\ell},$$

Preliminary version

for mini-batch sizes $M_0, \dots, M_L > 0$. This is our MLMC estimator for the gradient of the model evidence. By a reasoning similar to before, we can expect a situation with

$$\mathbb{V}[\nabla_\theta \Delta \hat{\mathcal{L}}_{1,1}] \gg \cdots \gg \mathbb{V}[\nabla_\theta \Delta \hat{\mathcal{L}}_{1,2^\ell}] \gg \cdots,$$

which allows for a rapid decrease

$$M_0 \gg \cdots \gg M_\ell \gg \cdots,$$

resulting in a substantial computational saving as compared to the nested Monte Carlo estimator. Although it is not necessarily the case that $\mathbb{V}[\nabla_\theta \Delta \hat{\mathcal{L}}_{1,2^\ell}]$ decays with the order of $2^{-\beta\ell}$ for the same β appearing in the decay of $\mathbb{V}[\Delta \hat{\mathcal{L}}_{1,2^\ell}]$, the theoretical results given in the next section state under some assumptions that we have $\beta = 2$ for both the model evidence and its gradient. Therefore, by following Remark 2, the gradient of the model evidence can be estimated by the randomized MLMC method without any bias.

4 THEORETICAL RESULTS

In order to show that the necessary computational cost for our MLMC estimator of the model evidence to achieve a mean squared accuracy ε^2 is of $O(\varepsilon^{-2})$, we need to introduce the fundamental theorem on MLMC methods proven by Giles (2008) and Cliffe et al. (2011). Although a general statement is given in the appendix, the statement below is adapted for the current context.

Theorem 1. *Assume that there exist positive constants c_1, c_2, α, β such that*

1. $\alpha \geq \min(\beta, 1)/2$,
2. $|\mathcal{L}_{2^\ell} - \mathcal{L}(x_{1:N})| \leq c_1 2^{-\alpha\ell}$, *and*
3. $\mathbb{V}[\Delta \hat{\mathcal{L}}_{1,2^\ell}] \leq c_2 2^{-\beta\ell}$.

Then, for any given accuracy $\varepsilon < \exp(-1)$, there exists a positive constant c_3 such that there are the corresponding maximum level L and the mini-batch sizes $M_0, M_1, \dots, M_L$ for which the mean squared error of the MLMC estimator $\hat{\mathcal{L}}^{MLMC}_{2^L}$ is less than ε^2 with the total computational cost C bounded by

$$\mathbb{E}[C] \leq \begin{cases} c_3 \varepsilon^{-2}, & \beta > 1, \\ c_3 \varepsilon^{-2} |\log \varepsilon^{-1}|^2, & \beta = 1, \\ c_3 \varepsilon^{-2-(1-\beta)/\alpha}, & \beta < 1. \end{cases}$$

Remark 3. If $\beta > 1$, the MLMC estimator can achieve the optimal computational complexity $O(\varepsilon^{-2})$ to estimate the model evidence. Notably, even if $\beta < 1$, the cost of order $\varepsilon^{-2-(1-\beta)/\alpha}$ still compares favorably with the nested Monte Carlo estimator for which the cost is of $O(\varepsilon^{-2-1/\alpha})$.

Remark 4. A statement similar to Theorem 1 also holds for the gradient of the model evidence. The difference is that, since the gradient is represented as a vector, the second and third assumptions should be replaced, respectively, by

2. $\|\nabla_\theta \mathcal{L}_{2^\ell} - \nabla_\theta \mathcal{L}(x_{1:N})\|_2 \leq c_1 2^{-\alpha\ell}$, and
3. $\mathbb{E}\left[\|\nabla_\theta \Delta \hat{\mathcal{L}}_{1,2^\ell} - \mathbb{E}[\nabla_\theta \Delta \hat{\mathcal{L}}_{1,2^\ell}]\|_2^2\right] \leq c_2 2^{-\beta\ell}$,

and that the mean squared error is given by the sum of the mean squared errors over all the elements. We refer to Section 2.5 of Giles (2015) for an extension of the MLMC theory to multi-dimensional outputs.

Based on Theorem 1, it suffices to characterize the values of α and β for the MLMC estimator. The following result for the model evidence is an immediate consequence from Goda et al. (2020). We show a full proof in the appendix for the sake of completeness.

Theorem 2. *If there exist $s, t > 2$ with $(s-2)(t-2) \geq 4$ such that*

$$\mathbb{E}_X\left[\int \left|\frac{p_\theta(X|Z)p_\theta(Z)}{p_\theta(X)q(Z;X)}\right|^s \mathrm{d}Z\right] < \infty, \quad \textit{and}$$
$$\mathbb{E}_X\left[\int \left|\log \frac{p_\theta(X|Z)p_\theta(Z)}{p_\theta(X)q(Z;X)}\right|^t \mathrm{d}Z\right] < \infty,$$

the MLMC estimator for the model evidence satisfies

$$\alpha = \min\left\{\frac{s(t-1)}{2t}, 1\right\} \quad \textit{and} \quad \beta = \min\left\{\frac{s(t-2)}{2t}, 2\right\}.$$

It is important to see that $\beta \geq 1$ whenever $(s-2)(t-2) \geq 4$, which directly implies that the MLMC estimator achieves the optimal order of computational cost. Moreover, if $(s-4)(t-2) \geq 8$, we have $\beta = 2$.

Regarding the gradient of the model evidence, we need to extend the result from the work by Hironaka et al. (2020) which has been studied in a different context and only dealt with a scalar output instead of vector. The following result is an analogy of the one shown in Goda et al. (2021). A full proof is given in the appendix.

Theorem 3. *If there exists $s \geq 2$ such that*

$$\mathbb{E}_X\left[\int \left|\frac{p_\theta(X|Z)p_\theta(Z)}{p_\theta(X)q(Z;X)}\right|^s \mathrm{d}Z\right] < \infty, \quad \textit{and}$$
$$\sup_{x,z} \|\nabla_\theta \log p_\theta(x|z)p_\theta(z)\|_\infty < \infty,$$

the MLMC estimator for the gradient of the model evidence satisfies

$$\alpha = \min\{s/2, 1\} \quad \textit{and} \quad \beta = \min\{s/2, 2\}.$$

Again we see that $\beta \geq 1$, which directly implies that the MLMC gradient estimator achieves the optimal order of computational cost in Theorem 1. Moreover, if $s \geq 4$, we have $\beta = 2$. Therefore, the assumptions made in Theorems 2 and 3 are satisfied simultaneously for large s and t, the MLMC estimators for the model evidence and its gradient both attain $\beta = 2$, which will be supported by the numerical results given in the next section.

Preliminary version

5 EXPERIMENTS

To illustrate the effectiveness of our MLMC approach, we compared the computational efficiency of several evidence estimation methods using a random effect logistic regression model. In the appendix, we additionally provide experimental results of Bayesian version of random effect logistic regression and (sparse) Gaussian process classification. In all experiments, our algorithm was run on a single CPU, and Python codes used in our experiment are available at `https:\github.com/Goda-Research-Group/mlmc-model-evidence`.

5.1 EXPERIMENTAL SETTINGS

The random effect logistic regression is a model of the following i.i.d. data generating process for $n = 1, 2, \dots, N$ and $t = 1, \dots, T$:

$$\begin{aligned} \boldsymbol{z}_n &\sim N(0, \tau^2) \\ \boldsymbol{y}_{n,t} &\sim \text{Bernoulli}\left(p_n\right), \end{aligned}$$

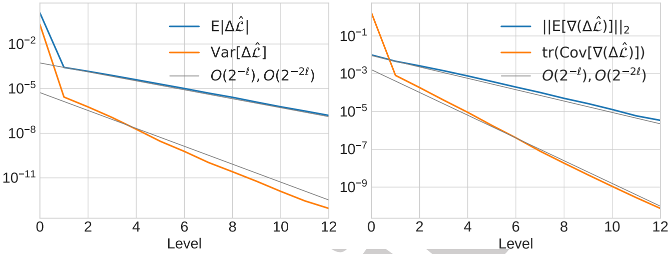
where we set the logit p_n to $p_n = \sigma(\boldsymbol{z}_n + w_0 + w^T x_{n,t})$ for the sigmoid link function $\sigma(x) = 1/(1 + \exp(-x))$. This model explains the binary response $\boldsymbol{y}_{n,t}$ of each individual n at each time point t given an explanatory variable $x_{n,t}$. By adding a random effect term $\boldsymbol{z}_n$ to the simple logistic regression model, we can estimate the effect (w_0, w) of $x_{n,t}$ on $\boldsymbol{y}_{n,t}$ more accurately by removing the individual variations in the data.

In our experiment, we used a synthetic data generated from a model whose parameters are given by $\eta = 1.0$, $w_0 = 0$, $w = (0.25, 0.50, 0.75)^T$. Here, we parametrized τ^2 with a non-constrained parameter η by softplus transformation as $\tau^2 = \log\left(1 + \exp(\eta)\right)$ to keep τ^2 positive. The explanatory variables $x_{n,t}$'s are all taken from a standard normal distribution and T was set to 2. For choosing a proposal distribution $q(z_n; x_{n,1:T})$, we used the Laplace approximation (Bishop, 2006). For the optimization, the Adam optimizer (Kingma and Ba, 2015) was used.

5.2 CONVERGENCE OF MLMC COUPLING

To examine whether the assumptions required for the MLMC estimation in Theorem 1 are satisfied, we evaluated the convergence behavior of the corrections $\Delta\hat{\mathcal{L}}_{1,2^\ell}$ and their gradient counterparts $\nabla_\theta \Delta\hat{\mathcal{L}}_{1,2^\ell}$.

Figure 2a shows the convergence behaviors of $\mathbb{E}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]$ and $\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]$. We see that $\mathbb{E}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]$ and $\mathbb{V}[\Delta\hat{\mathcal{L}}_{1,2^\ell}]$ approximately decay with the orders of $2^{-\ell}$ and $2^{-2\ell}$, respectively, implying that we have $\alpha = 1$ and $\beta = 2$ in the assumptions of Theorem 1.



(a) Decay of $\Delta\hat{\mathcal{L}}_{1,2^\ell}$ (b) Decay of $\nabla_\theta \Delta\hat{\mathcal{L}}_{1,2^\ell}$

Figure 2: Convergence of the mean and variance of the coupled correction estimators.

Figure 2b shows the convergence behaviors of $\mathbb{E}[\nabla_\theta \Delta\hat{\mathcal{L}}_{1,2^\ell}]$ and $\mathrm{Cov}[\nabla_\theta \Delta\hat{\mathcal{L}}_{1,2^\ell}]$ in L_2 norm and in trace, respectively. The trace of the covariance is equivalent to $\mathbb{E}\left[\|\nabla_\theta \Delta\hat{\mathcal{L}}_{1,2^\ell} - \mathbb{E}[\nabla_\theta \Delta\hat{\mathcal{L}}_{1,2^\ell}]\|_2^2\right]$ appearing in Remark 4. Again, the requirements for the MLMC method, i.e., the exponential decays of the corrections terms, are satisfied for the mean and the variance of the gradient counterparts. These numerical results support the theoretical findings given in Section 4.

5.3 ACCURACY OF ESTIMATION BY MLMC

Next, we compared the estimation accuracy of several evidence estimation methods by changing $K(= 2^L)$ for biased objective $\mathcal{L}_K$. In Table 1, the means and standard deviations of the estimated parameters obtained from different objectives are listed. For each method and objective, the parameters were estimated 100 times to obtain these quantities. The soft truncation of the SUMO was replaced by hard truncation, to match the bias of all estimators for given K. We can see that the biases become smaller as we increase K and the smallest bias is attained for $K = 512$ (or $L = 9$). When we look at the standard deviations of the estimates, both randomized (RandMLMC) and non-randomized MLMC methods yield smaller standard deviations and mean squared errors (MSEs) than other methods with the same bias. This is because the gradient estimation by the MLMC method has a smaller variance than other methods.

Additionally, we compared the progression of stochastic optimization in Figure 3. Each objective was optimized 100 times, and the means and the standard deviation are represented by the lines and error bands, respectively. Again, we can see that the MLMC method and the randomized MLMC method converge to the smallest MSEs than others. Though the SUMO did not converge as fast as other estimators in this experiment, it converged to good solutions after sufficient time was elapsed, as shown in Table 1.

Preliminary version

Table 1: Accuracy of Parameter Estimation by Different Objectives and Estimation Methods.

	η	w_0	w_1	w_2	w_3	MSE
Ground Truth	1.0	0.0	0.25	0.5	0.75	0.0
NMC (K=1)	-0.272 ± 0.121	-0.003 ± 0.023	0.231 ± 0.021	0.456 ± 0.021	0.684 ± 0.022	1.6412
NMC (K=8)	0.546 ± 0.086	-0.005 ± 0.023	0.244 ± 0.021	0.485 ± 0.020	0.712 ± 0.018	0.2167
NMC (K=64)	0.894 ± 0.059	**0.002** ± 0.024	0.252 ± 0.019	0.480 ± 0.021	0.744 ± 0.022	0.0169
NMC (K=512)	1.038 ± 0.049	0.012 ± 0.021	0.244 ± 0.021	**0.496** ± 0.020	**0.747** ± 0.022	0.0059
MLMC (L=9)	1.052 ± 0.033	0.010 ± 0.006	**0.250** ± 0.005	0.511 ± 0.003	0.741 ± 0.003	0.0041
RandMLMC (L=9)	**0.966** ± 0.033	-0.003 ± 0.006	0.241 ± 0.004	0.507 ± 0.003	0.744 ± 0.003	**0.0026**
SUMO (K=512)	0.951 ± 0.083	-0.011 ± 0.013	0.242 ± 0.008	0.506 ± 0.009	0.739 ± 0.009	0.0101
Jackknife (K=512)	0.959 ± 0.053	-0.016 ± 0.020	0.248 ± 0.021	0.494 ± 0.022	0.745 ± 0.016	0.0065

Preliminary version

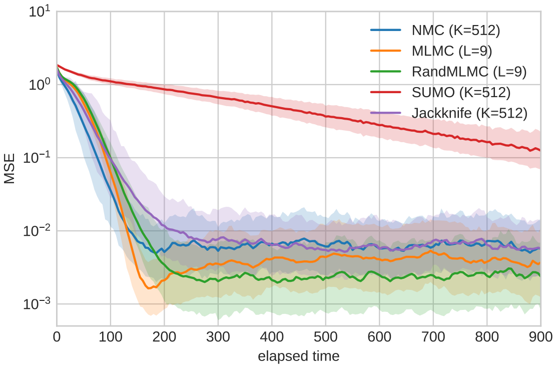


Figure 3: Learning curves of the different objective functions. Instead of the loss function, the mean squared errors from the true parameters are plotted.

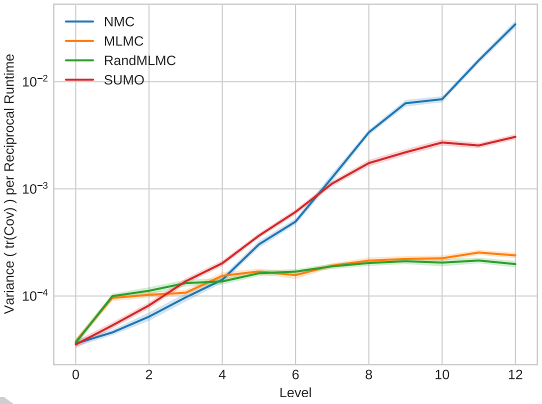


Figure 4: Computational efficiency of the gradient at each level. Two to the power of the level corresponds to the number of inner Monte Carlo samples, i.e. $2^L = K$.

5.4 COMPUTATIONAL EFFICIENCY OF MLMC

To quantitatively compare the computational efficiency of each estimator, we plotted the variance of the gradient estimator of $\nabla_\theta \mathcal{L}_{2^L}$ against each level L, for a given computational cost (runtime) in Figure 4. As the variance of Monte Carlo estimators decreases reciprocally to the number of samples (or, the computational cost), we multiplied the variance of each estimator by the runtime the estimator spent to obtain a measure of computational efficiency. The Jackknife estimator is not compared, because its bias for given L is not equal to those of the other estimators. Since the level corresponds to the magnitude of bias, the plot can also be interpreted as the comparison of the computational efficiency for different bias size. In this experiment, we used the largest batch size that fits in the memory to ignore the implementational inefficiencies of our Python code. Unlike the nested Monte Carlo estimator, the iteration over multiple levels in MLMC and SUMO cannot be written with basic array operations, and this causes runtime overhead when the batch size is small.

Theoretically, the variance per computational cost becomes $O(1)$ for the MLMC-based estimators, while $O(2^L)$ and $O(L^2)$ costs are required for the nested Monte Carlo and the SUMO, respectively. For large levels, the MLMC-based estimators are 10 to 100 times more efficient than other estimators. However, in the low-level regions ($L \leq 3$), although the corresponding objective is quite biased, the nested Monte Carlo estimator is the most efficient.

6 CONCLUSIONS

This paper introduced a new estimator for the model evidence and its gradient based on the MLMC sampling technique. In the theoretical analysis, we showed that the computational complexity of our MLMC estimator is an order of magnitude smaller than the standard nested Monte Carlo estimator and the estimator can be made unbiased while having finite variance and expected computational cost. This property is unprecedented by any other existing debiasing methods for the model evidence estimation. In the experiments, we confirmed that our MLMC estimator performs as expected from the theory and observed its superiority over the existing estimators.

Preliminary version

Acknowledgements

This work was supported by Physical Sciences Research Council (EPSRC) grant. K.I. is supported by the Heiwa Nakajima Foundation. T.G. is supported by JSPS KAKENHI Grant Number 20K03744.

References

David F Anderson and Desmond J Higham. Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. *Multiscale Modeling & Simulation*, 10(1):146–179, 2012.

Christopher M Bishop. *Pattern Recognition and Machine Learning*. springer, 2006.

Karolina Bujok, BM Hambly, and Christoph Reisinger. Multilevel simulation of functionals of Bernoulli random variables with application to basket credit derivatives. *Methodology and Computing in Applied Probability*, 17 (3):579–604, 2015.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *Fourth International Conference on Learning Representations (ICLR 2016)*. arXiv:1509.00519, 2016.

K Andrew Cliffe, Mike B Giles, Robert Scheichl, and Aretha L Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3, 2011.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

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.

Michael B Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

Michael B Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.

Michael B Giles and Takashi Goda. Decision-making under uncertainty: using MLMC for efficient estimation of EVPPI. *Statistics and Computing*, 29(4):739–751, 2019.

Mike Giles, Tigran Nagapetyan, Lukasz Szpruch, Sebastian Vollmer, and Konstantinos Zygalakis. Multilevel Monte Carlo for scalable Bayesian computations. *arXiv preprint arXiv:1609.06144*, 2016.

Takashi Goda, Tomohiko Hironaka, and Takeru Iwamoto. Multilevel Monte Carlo estimation of expected information gains. *Stochastic Analysis and Applications*, 38(4): 581–600, 2020.

Takashi Goda, Tomohiko Hironaka, Wataru Kitade, and Adam Foster. Unbiased MLMC stochastic gradient-based optimization of Bayesian experimental designs. *arXiv preprint arXiv:2005.08414v2*, 2021.

Stefan Heinrich. Monte Carlo complexity of global solution of integral equations. *Journal of Complexity*, 14(2):151–175, 1998.

Tomohiko Hironaka, Michael B Giles, Takashi Goda, and Howard Thom. Multilevel Monte Carlo estimation of the expected value of sample information. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):1236–1259, 2020.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR 2015)*. arXiv:1412.6980, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations (ICLR 2014)*. arXiv:1312.6114, 2014.

Yucen Luo, Alex Beatson, Mohammad Norouzi, Jun Zhu, David Duvenaud, Ryan P Adams, and Ricky TQ Chen. SUMO: Unbiased estimation of log marginal probability for latent variable models. In *Eighth International Conference on Learning Representations (ICLR 2020)*. arXiv:2004.00353, 2020.

Sebastian Nowozin. Debiasing evidence approximations: On importance-weighted autoencoders and Jackknife variational inference. In *Sixth International Conference on Learning Representations (ICLR 2018)*, 2018.

Chang-Han Rhee and Peter W Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

Maxime Rischard, Pierre E Jacob, and Natesh Pillai. Unbiased estimation of log normalizing constants with applications to Bayesian cross-validation. *arXiv preprint arXiv:1810.01382*, 2018.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Science & Business Media, 2013.

Francisco JR Ruiz, Michalis K Titsias, Taylan Cemgil, and Arnaud Doucet. Unbiased gradient estimation for variational auto-encoders using coupled Markov chains. *arXiv preprint arXiv:2010.01845*, 2020.

Colin Wei and Iain Murray. Markov chain truncation for doubly-intractable inference. In *20th International Conference on Artificial Intelligence and Statistics*, pages 776–784. PMLR, 2017.

Greg CG Wei and Martin A Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990.

Preliminary version