

UNBIASED MLMC-BASED VARIATIONAL BAYES FOR LIKELIHOOD-FREE INFERENCE*

ZHIJIAN HE[†], ZHENGHANG XU[‡], AND XIAOQUN WANG[§]

Abstract. Variational Bayes (VB) is a popular tool for Bayesian inference in statistical modeling. Recently, some VB algorithms are proposed to handle intractable likelihoods with applications such as approximate Bayesian computation. In this paper, we propose several unbiased estimators based on multilevel Monte Carlo (MLMC) for the gradient of Kullback-Leibler divergence between the posterior distribution and the variational distribution when the likelihood is intractable, but can be estimated unbiasedly. The new VB algorithm differs from the VB algorithms in the literature which usually render biased gradient estimators. Moreover, we incorporate randomized quasi-Monte Carlo (RQMC) sampling within the MLMC-based gradient estimators, which was known to provide a favorable rate of convergence in numerical integration. Theoretical guarantees for RQMC are provided in this new setting. Numerical experiments show that using RQMC in MLMC greatly speeds up the VB algorithm, and finds a better parameter value than some existing competitors do.

Key words. Multilevel Monte Carlo, quasi-Monte Carlo, variational Bayes, intractable likelihood, nested simulation

AMS subject classifications. 65C05, 62F15

1. Introduction. In this article, we are interested in variational Bayes (VB), which is widely used as a computationally effective method for approximating the posterior distribution of a Bayesian problem. Let y^* be the observed data and $\theta \in \mathbb{R}^p$ be the parameter of interest. The posterior distribution $p(\theta|y^*) \propto p(\theta)p(y^*|\theta)$, where $p(\theta)$ is the prior and $p(y^*|\theta)$ is the likelihood function. VB approximates the posterior by a tractable distribution $q(\theta)$ within certain distribution families, chosen to minimize the Kullback-Leibler (KL) divergence between the VB distribution $q(\theta)$ and the posterior $p(\theta|y^*)$. The optimization problem is usually solved by using the stochastic gradient descent (SGD) algorithm [8]. It calls for computing the gradient of the KL divergence. A difficulty with SGD is that plain Monte Carlo (MC) sampling to estimate the gradient can be error prone or inefficient. Some variance reduction methods have been adopted to improve SGD [24, 29]. On the other hand, randomized quasi-Monte Carlo (RQMC) methods have been used to improve SGD in the VB setting [4]. Recently, Liu and Owen [23] combined RQMC with a second order limited memory method known as L-BFGS for VB. RQMC methods such as scrambled digital nets proposed by [26] were known to provide a favorable rate of convergence in numerical integration [27]. Improved sampling accuracy translates directly to improved optimization as shown in [4, 23].

A second difficulty with SGD is due to the absence of the likelihood function $p(y^*|\theta)$. In many applications, the likelihood function is intractable making it difficult

*Submitted to the editors DATE.

Funding: This work of the first author was funded by the National Science Foundation of China (No. 12071154), Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515010275), Guangzhou Science and Technology Program (No. 202102020407). And the third author was funded by the National Science Foundation of China (No. 720711119).

[†]School of Mathematics, South China University of Technology, Guangzhou 510641, People's Republic of China (hezhijian@scut.edu.cn).

[‡]Corresponding author. Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China (xzh17@mails.tsinghua.edu.cn).

[§]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China (wangxiaoqun@mail.tsinghua.edu.cn).

to render an unbiased gradient estimator of the KL divergence. For example, the likelihood is an intractable high-dimensional integral over the state variables governed by a Markov process in state space-space models [9]. More examples can be found in the context of approximate Bayesian computation (ABC). ABC methods provide a way of approximating the posterior $p(\theta|y^*)$ when the likelihood function is difficult to compute but it is possible to simulate data from the model [30, 34].

Likelihood-free inference is an active area in Bayesian computation. There are some progresses on using VB in the likelihood-free context. Barthelmé and Chopin [2] used a variational approximation algorithm known as expectation propagation in approximating ABC posteriors. Tran et al. [35] developed a new VB with intractable likelihood (VBIL) method, which can be applied to commonly used statistical models without requiring an analytical solution to model-based expectations. Ong et al. [25] modified the VBIL method to work with unbiased log-likelihood estimates in the synthetic likelihood framework, resulting in the VB synthetic likelihood (VBSL) method.

We focus on the problems in which the likelihoods are formulated as an intractable expectation. The KL divergence turns out to be a nested expectation and so does its gradient. It is natural to use nested simulation for estimating these quantities. However, the plain nested estimator is biased. It is critical to develop unbiased gradient estimators for stochastic gradient-based optimization algorithms. To this end, we use the unbiased multilevel Monte Carlo (MLMC) proposed by [33] in the framework of nested simulation. MLMC is a sophisticated variance reduction technique introduced by [19] for parametric integration and by [12] for the estimation of the expectations arising from stochastic differential equations. Nowadays MLMC methods have been extended extensively. For a thorough review of MLMC methods, we refer to [13]. Nested simulation combined with the MLMC method has been widely studied in the literature due to its broad applicability [5, 14, 15, 17].

In this paper, we develop an unbiased nested MLMC-based VB method to deal with intractable likelihoods. Our work is related to [18], who developed an unbiased MLMC stochastic gradient-based optimization method for Bayesian experimental designs. Our proposed VB algorithm finds a better parameter value and a larger evidence lower bounded (ELBO) thanks to unbiased gradient and ELBO estimators. This leads to a better estimate of the marginal likelihood $p(y^*)$ compared to the VBIL method, which is an important factor in model selection. We also incorporate the RQMC sampling within the gradient and the ELBO estimators, which reduces the computational complexity effectively. Goda et al. [18] worked on the MC sampling rather than RQMC. We provide some numerical analysis for both MC and RQMC settings.

The rest of this paper is organized as follows. In Section 2, we review some VB methods with intractable likelihoods, such as VBIL and VBSL, and illuminate their limitations. In Section 3, we provide our unbiased MLMC methods for VB and discuss two different estimators of gradient, which are the score function gradient and re-parameterization gradient. In Section 4, we provide the details of our algorithms when using Gaussian variational family in VB. In Section 5, we improve the algorithms by incorporating RQMC and do some numerical analysis. Finally, in Section 6, some numerical experiments are conducted to support the advantages of our proposed methods. Section 7 concludes this paper.

2. Variational Bayes with an intractable likelihood. Recall that our target is to estimate the posterior distribution

$$(2.1) \quad p(\theta|y^*) = \frac{p(\theta)p(y^*|\theta)}{p(y^*)},$$

where $p(y^*) = \int p(\theta)p(y^*|\theta)d\theta$ is usually an unknown constant (called the marginal likelihood or evidence). In many applications such as state-space models and ABC, the likelihood is analytically intractable. For these cases, the likelihood $p(y^*|\theta)$ is usually formulated as an expectation

$$(2.2) \quad p(y^*|\theta) = \mathbb{E}[f(x; y^*)|\theta],$$

where $x \sim p(x|\theta)$ is the latent variable.

Suppose that there exists an unbiased estimator $\hat{p}_N(y^*|\theta)$ for the intractable likelihood $p(y^*|\theta)$ for given θ , where N is an algorithmic parameter relating to the precision in estimating the likelihood. For estimating (2.2), one can take the sample-mean estimator

$$(2.3) \quad \hat{p}_N(y^*|\theta) = \frac{1}{N} \sum_{i=1}^N f(x_i; y^*),$$

where x_i are iid copies of x for a given θ . In this paper, we restrict our attention to the sample-mean estimator (2.3). We should note that for the state-space models, the likelihood can be unbiasedly estimated by an importance sampling estimator [10], or by a particle filter estimator [31]. The later case does not fit into our framework.

VB approximates the posterior distribution $p(\theta|y^*)$ by a tractable density $q_\lambda(\theta)$ with a variational parameter λ , chosen to minimize the KL divergence from $q_\lambda(\theta)$ to $p(\theta|y^*)$, which is defined by

$$\text{KL}(\lambda) = \text{KL}(q_\lambda(\theta)||p(\theta|y^*)) = \mathbb{E}_{q_\lambda(\theta)}[\log q_\lambda(\theta) - \log p(\theta|y^*)].$$

Using (2.1), we have

$$\log p(y^*) = \text{KL}(\lambda) + L(\lambda),$$

where $L(\lambda)$ is defined by

$$L(\lambda) = \mathbb{E}_{q_\lambda(\theta)}[\log p(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta)].$$

Since $\text{KL}(\lambda) \geq 0$, $L(\lambda)$ is a lower bound of the log-evidence $\log p(y^*)$, which is called the ELBO. The minimization of KL is translated to the maximization of the ELBO since the marginal likelihood $p(y^*)$ is fixed. The problem turns out to solve

$$\lambda^* = \arg \max_{\lambda \in \Lambda} L(\lambda),$$

where Λ is the feasible region of λ . Stochastic gradient method and its variants are widely used to solve such a problem. They use a sequence of steps

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t \nabla_\lambda L(\lambda^{(t)}),$$

where $\nabla_\lambda L(\lambda)$ is gradient of the ELBO and $\rho_t > 0$ is the learning rate satisfying the Robbins-Monro conditions: $\sum_{t=0}^{\infty} \rho_t = \infty$ and $\sum_{t=0}^{\infty} \rho_t^2 < \infty$. A simple choice is

$\rho_t = a/(t + b)$ for some constants $a, b > 0$. Some adaptive methods for choosing the learning rate ρ_t were proposed in the literature, notably AdaGrad [7] and Adam [20].

The key in stochastic gradient methods is to estimate the gradient $\nabla_\lambda L(\lambda)$ unbiasedly. In the literature, the re-parameterization (RP) trick [21] and the score function (SF) are two popular methods to derive unbiased gradient estimators. Allowing the interchange of differentiation and expectation as required in the SF method, we have

$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathbb{E}_{q_\lambda(\theta)}[\log p(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta)] \\ &= \mathbb{E}_{q_\lambda(\theta)}[\nabla_\lambda \log q_\lambda(\theta)(\log p(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta))],\end{aligned}$$

where we used the fact that $\mathbb{E}_{q_\lambda(\theta)}[\nabla_\lambda \log q_\lambda(\theta)] = 0$. If the likelihood function $p(y^*|\theta)$ is known, it is straightforward to derive an unbiased estimator for $\nabla_\lambda L(\lambda)$ by sampling $\theta \sim q_\lambda(\theta)$ repeatedly. However, in our setting, $\log p(y^*|\theta)$ is intractable. The question is how to use the unbiased estimator $\hat{p}_N(y^*|\theta)$ of the likelihood to construct an unbiased SF estimator for $\nabla_\lambda L(\lambda)$.

On the other hand, for applying the RP trick, we assume that there exists a transformation $\theta = \Gamma(\mathbf{u}; \lambda) \sim q_\lambda(\theta)$, where the random variate $\mathbf{u} \sim p_1(\mathbf{u})$ independently of λ . Allowing the interchange of differentiation and expectation again, we have

$$\begin{aligned}\nabla_\lambda L(\lambda) &= \nabla_\lambda \mathbb{E}_{q_\lambda(\theta)}[\log p(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta)] \\ &= \nabla_\lambda \mathbb{E}_{\mathbf{u}}[\log p(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta)] \\ (2.4) \quad &= \mathbb{E}_{\mathbf{u}}[\nabla_\lambda \Gamma(\mathbf{u}; \lambda) \cdot (\nabla_\theta \log p(y^*|\theta) + \nabla_\theta \log p(\theta) - \nabla_\theta \log q_\lambda(\theta))],\end{aligned}$$

where $\nabla_\lambda \Gamma(\mathbf{u}; \lambda)$ is the Jacobian matrix with entries $[\nabla_\lambda \Gamma(\mathbf{u}; \lambda)]_{ij} = \partial \Gamma_j(\mathbf{u}; \lambda) / \partial \lambda_i$. The RP gradient is much complicated than the SF gradient. In (2.4), one needs to estimate the intractable gradient of log-likelihood $\nabla_\theta \log p(y^*|\theta)$ unbiasedly. Due to the absence of likelihood, the SF and RP methods for the traditional VB cannot be applied directly.

The VBIL method proposed by [35] works with the augmented space (θ, z) , where $z = \log \hat{p}_N(y^*|\theta) - \log p(\theta|y^*)$. Let $g_N(z|\theta)$ be the distribution of z given θ . Tran et al. [35] applied the variational inference for the target distribution

$$p_N(\theta, z) = p(\theta|y^*) \exp(z) g_N(z|\theta)$$

with a family of distributions of the form $q_\lambda(\theta, z) = q_\lambda(\theta) g_N(z|\theta)$. The KL divergence in the augmented space is

$$\widetilde{\text{KL}}(\lambda) = \text{KL}(q_\lambda(\theta, z) || p_N(\theta, z)) = \mathbb{E}_{q_\lambda(\theta, z)}[\log q_\lambda(\theta) - \log p(\theta|y^*) - z].$$

The ELBO in the augmented space is

$$\begin{aligned}(2.5) \quad \tilde{L}(\lambda) &= \mathbb{E}_{q_\lambda(\theta, z)}[\log \hat{p}_N(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta)] \\ &= L(\lambda) + \mathbb{E}_{q_\lambda(\theta, z)}[z].\end{aligned}$$

Note that

$$\mathbb{E}[z|\theta] = \mathbb{E}[\log \hat{p}_N(y^*|\theta)] - \log p(y^*|\theta) \leq \log \mathbb{E}[\hat{p}_N(y^*|\theta)] - \log p(y^*|\theta) = 0$$

by using Jensen's inequality. As a result, $L(\lambda) \geq \tilde{L}(\lambda)$. The equality holds if and only if $\hat{p}_N(y^*|\theta)$ is a constant with probability 1 (w.p.1). Generally, the maximization of $\tilde{L}(\lambda)$ is not the same as the maximization of $L(\lambda)$ unless $\mathbb{E}_{q_\lambda(\theta, z)}[z]$ is independent of λ .

Tran et al. [35] made an attempt to choose N as a function of θ such that $\mathbb{E}[z|\theta] \equiv \tau$ does not depend on θ . By doing so, $\mathbb{E}_{q_\lambda(\theta, z)}[z] = \tau$ does not depend on λ . Hence, in practice, one needs to adapt N so that the variance of the log-likelihood estimator is approximately constant with θ . Ong et al. [25] suggested to set some minimum value N' for the initially estimating the likelihood. Then, if some target value for the log-likelihood variance is exceeded based on an empirical estimate, an additional number of samples is repeatedly simulated until the target accuracy is achieved. Although the two ELBOs have the same maximizer, there is a gap (i.e., τ) between the maximums of the two ELBOs. The smaller the target accuracy is, the more work is required in estimating the likelihood. Actually, $L(\lambda)$ is a locally marginalized version of $\tilde{L}(\lambda)$, which is tighter. This can help to approximate the evidence better. Furthermore, this tighter lower bound can potentially help to compute the criterion for model selection such as perplexity used in topic modeling.

In fact, if we use an unbiased estimator of $\log p(y^*|\theta)$ to replace $\log \hat{p}_N(y^*|\theta)$ in (2.5), then the resulting ELBO corresponds to the original ELBO $L(\lambda)$. However, an unbiased estimator of $\log p(y^*|\theta)$ is not trivial. To overcome this, [25] proposed to use a synthetic likelihood. Suppose we have a summary statistic $\mathcal{S} = \mathcal{S}(y^*)$ of dimension $d \geq p$ and the inference is based on the observed value s of the summary statistic \mathcal{S} , which is thought to be informative about θ . Assume that the statistic \mathcal{S} is exactly Gaussian conditional on each value of θ , that is $p(s|\theta) = \phi(s; \mu(\theta), \Sigma(\theta))$, where ϕ is the density of multivariate normal with $\mu(\theta) = \mathbb{E}[\mathcal{S}|\theta]$ and $\Sigma(\theta) = \text{Cov}(\mathcal{S}|\theta)$. Now the posterior density is given by

$$p(\theta|s) \propto p(\theta)p(s|\theta) = p(\theta)\phi(s; \mu(\theta), \Sigma(\theta)).$$

For a given θ , we may simulate summary statistics $\mathcal{S}_1, \dots, \mathcal{S}_N$ under the model given θ . The mean vector $\mu(\theta)$ and the covariance matrix $\Sigma(\theta)$ are then estimated by

$$\hat{\mu}(\theta) = \frac{1}{N} \sum_{i=1}^N \mathcal{S}_i,$$

$$\hat{\Sigma}(\theta) = \frac{1}{N-1} \sum_{i=1}^N (\mathcal{S}_i - \hat{\mu}(\theta))(\mathcal{S}_i - \hat{\mu}(\theta))^\top,$$

respectively. Then an unbiased estimate of the log-synthetic likelihood $\log p(s|\theta)$ is given by

$$\hat{\ell}_N(s|\theta) = -\frac{d}{2} \log(2\pi) - \frac{1}{2} \left\{ \log |\hat{\Sigma}(\theta)| + d \log \left(\frac{N-1}{2} \right) - \sum_{i=1}^d \psi \left(\frac{N-i}{2} \right) \right\}$$

$$- \frac{1}{2} \left\{ \frac{N-d-2}{N-1} (s - \hat{\Sigma}(\theta))^\top \hat{\Sigma}(\theta)^{-1} (s - \hat{\Sigma}(\theta)) - \frac{d}{N} \right\},$$

where $\psi(t) = \Gamma'(t)/\Gamma(t)$ denotes the digamma function and $N > d + 2$. By replacing $\log \hat{p}_N(y^*|\theta)$ with $\hat{\ell}_N(s|\theta)$, then $\tilde{L}(\lambda) = L(\lambda)$. However, it should be noted that the unbiasedness of $\hat{\ell}_N(s|\theta)$ relies heavily on the assumption of the normality of $\mathcal{S}|\theta$, and the inference is based on the information of the summary statistic s rather than the full data y^* .

3. Unbiased MLMC for variational Bayes. To fix our idea, we work on the likelihood (2.2) with an unbiased estimate (2.3). Now the ELBO is a nested expectation

$$L(\lambda) = \mathbb{E}_{q_\lambda(\theta)} [\log \mathbb{E}[f(x; y^*)|\theta] + \log p(\theta) - \log q_\lambda(\theta)].$$

3.1. Score function gradient. Applying the SF method, we reformulate the gradient as

$$\nabla_\lambda L(\lambda) = \mathbb{E}_{q_\lambda(\theta)}[\nabla_\lambda \log q_\lambda(\theta)(\log \mathbb{E}[f(x; y^*)|\theta] + \log p(\theta) - \log q_\lambda(\theta))],$$

which is a nested expectation. Define

$$(3.1) \quad \text{SF}_N(\lambda) = \nabla_\lambda \log q_\lambda(\theta)[\log \hat{p}_N(y^*|\theta) + \log p(\theta) - \log q_\lambda(\theta)],$$

where $\hat{p}_N(y^*|\theta)$ is given by (2.3), and $(\theta, x) \sim q_\lambda(\theta)p(x|\theta)$. Although $\hat{p}_N(y^*|\theta)$ is an unbiased likelihood estimator, $\text{SF}_N(\lambda)$ is generally biased for estimating the gradient $\nabla_\lambda L(\lambda)$. We next show how to find an unbiased estimator for the log-likelihood by using unbiased MLMC. Let $\psi_{\theta, N} = \log \hat{p}_N(y^*|\theta)$. It is clear that

$$\lim_{N \rightarrow \infty} \mathbb{E}[\psi_{\theta, N}|\theta] = \log p(y^*|\theta).$$

Consider an increasing sequence $0 < M_0 < M_1 < \dots$ such that $M_\ell \rightarrow \infty$ as $\ell \rightarrow \infty$. Then the following telescoping sum holds,

$$\log p(y^*|\theta) = \mathbb{E}[\psi_{\theta, M_0}|\theta] + \sum_{\ell=1}^{\infty} \mathbb{E}[\psi_{\theta, M_\ell} - \psi_{\theta, M_{\ell-1}}|\theta].$$

More generally, if we have a sequence of correction random variables $\Delta\psi_{\theta, \ell}$, $\ell \geq 0$ such that $\mathbb{E}[\Delta\psi_{\theta, 0}|\theta] = \mathbb{E}[\psi_{\theta, M_0}|\theta]$ and for $\ell > 0$,

$$\mathbb{E}[\Delta\psi_{\theta, \ell}|\theta] = \mathbb{E}[\psi_{\theta, M_\ell} - \psi_{\theta, M_{\ell-1}}|\theta],$$

then it follows that

$$\log p(y^*|\theta) = \sum_{\ell=0}^{\infty} \mathbb{E}[\Delta\psi_{\theta, \ell}|\theta].$$

Let $w_\ell > 0$ satisfying $\sum_{\ell=0}^{\infty} w_\ell = 1$, and let I be an independent discrete random variable with $\mathbb{P}(I = \ell) = w_\ell$. We then have

$$\log p(y^*|\theta) = \mathbb{E} \left[\frac{\Delta\psi_{\theta, I}}{w_I} \middle| \theta \right].$$

Define

$$(3.2) \quad \text{SF}_{\text{MLMC}}(\lambda) = \nabla_\lambda \log q_\lambda(\theta) \left[\frac{\Delta\psi_{\theta, I}}{w_I} + \log p(\theta) - \log q_\lambda(\theta) \right],$$

which is unbiased for the gradient $\nabla_\lambda L(\lambda)$. For any number of outer samples $S \geq 1$, the following gradient estimator,

$$(3.3) \quad \widehat{\nabla_\lambda L}^{\text{SF}}(\lambda) = \frac{1}{S} \sum_{i=1}^S \text{SF}_{\text{MLMC}}^{(i)}(\lambda),$$

is unbiased, where $\text{SF}_{\text{MLMC}}^{(i)}(\lambda)$ are iid copy of $\text{SF}_{\text{MLMC}}(\lambda)$ for the MC sampling.

Now

$$\psi_{\theta, M_\ell} = \log \hat{p}_{M_\ell}(y^*|\theta) = \log \left(\frac{1}{M_\ell} \sum_{i=1}^{M_\ell} f(x_i; y^*) \right),$$

where $x_i \sim p(x|\theta)$ independently. We take $\Delta\psi_{\theta,0} = \psi_{\theta,M_0}$. For $\ell \geq 1$, we take an antithetic coupling estimator

$$\Delta\psi_{\theta,\ell} = \psi_{\theta,M_\ell} - \frac{1}{2} \left(\psi_{\theta,M_{\ell-1}}^{(a)} + \psi_{\theta,M_{\ell-1}}^{(b)} \right),$$

where

$$\psi_{\theta,M_{\ell-1}}^{(a)} = \log \left(\frac{1}{M_{\ell-1}} \sum_{i=1}^{M_{\ell-1}} f(x_i; y^*) \right), \quad \psi_{\theta,M_{\ell-1}}^{(b)} = \log \left(\frac{1}{M_{\ell-1}} \sum_{i=M_{\ell-1}+1}^{M_\ell} f(x_i; y^*) \right).$$

The strategy of antithetic coupling is widely used in the MLMC literature [16, 17], which yields a better rate of convergence for smooth functions. Denote C_ℓ as the expected cost of computing $\Delta\psi_{\theta,\ell}$, which is proportional to M_ℓ . To ensure a finite variance and finite expected computational cost of $\text{SF}_{\text{MLMC}}(\lambda)$, it is required that

$$(3.4) \quad \sum_{\ell=0}^{\infty} \frac{\mathbb{E}[\Delta\psi_{\theta,\ell}^2 \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2]}{w_\ell} < \infty \text{ and } \sum_{\ell=0}^{\infty} C_\ell w_\ell < \infty.$$

In this paper, we take $M_\ell = M_0 2^\ell$ for some $M_0 \geq 1$ and all $\ell \geq 0$, implying $C_\ell = O(2^\ell)$. Assume that $\mathbb{E}[\Delta\psi_{\theta,\ell}^2 \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2] = O(2^{-r\ell})$ for some $r > 1$. Let $w_\ell = w_0 2^{-\alpha\ell}$ for $w_0 = 1 - 2^{-\alpha}$ and $\alpha > 0$. Then (3.4) holds if we take $\alpha \in (1, r)$. The expected computational cost is then proportional to

$$(3.5) \quad C(\alpha, M_0) = \sum_{\ell=0}^{\infty} M_\ell w_\ell = \sum_{\ell=0}^{\infty} M_0 w_0 2^{(1-\alpha)\ell} = \left(1 + \frac{1}{2^\alpha - 2} \right) M_0.$$

LEMMA 3.1. *Let X be a random variable with zero mean, and let \bar{X}_N be an average of N iid samples of X . If $\mathbb{E}[|X|^p] < \infty$ for $p > 2$, then there exists a constant C_p depending only on p such that*

$$\mathbb{E}[|\bar{X}_N|^p] \leq C_p \frac{\mathbb{E}[|X|^p]}{N^{p/2}}.$$

Lemma 3.1 is stated as Lemma 1 in [15], with which we have the following theorem.

THEOREM 3.1. *Suppose that $f(x; y^*) > 0$ w.p.1, and there exist $p, q > 2$ with $(p-2)(q-2) > 4$ such that*

$$\mathbb{E} \left[\left| \frac{f(x; y^*)}{p(y^*|\theta)} \right|^p \right] < \infty \text{ and } \mathbb{E} \left[\left(1 + \left| \log \frac{f(x; y^*)}{p(y^*|\theta)} \right|^q \right) \|\nabla_\lambda \log q_\lambda(\theta)\|_2^q \right] < \infty,$$

where the expectations are taken with respect to $(\theta, x) \sim q_\lambda(\theta)p(x|\theta)$, then

$$\mathbb{E}[\Delta\psi_{\theta,\ell}^2 \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2] = O(2^{-r\ell}) \text{ with } r = \min \left(\frac{p(q-2)}{2q}, 2 \right) \in (1, 2].$$

Proof. This proof is in line with Theorem 2 of [17], which developed MLMC for a nested expectation of the form $\mathbb{E}_{X,Y}[\log[g(X,Y)|Y]]$. Let

$$R = \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} \frac{f(x_i; y^*)}{p(y^*|\theta)},$$

$$R^{(a)} = \frac{1}{M_{\ell-1}} \sum_{i=1}^{M_{\ell-1}} \frac{f(x_i; y^*)}{p(y^*|\theta)}, \quad R^{(b)} = \frac{1}{M_{\ell-1}} \sum_{i=M_{\ell-1}+1}^{M_{\ell}} \frac{f(x_i; y^*)}{p(y^*|\theta)}.$$

We then have

$$\Delta\psi_{\theta,\ell} = (\log R - R + 1) - \frac{1}{2} \left[(\log R^{(a)} - R^{(a)} + 1) + (\log R^{(b)} - R^{(b)} + 1) \right].$$

Applying Jensen's inequality gives

$$\Delta\psi_{\theta,\ell}^2 \leq 2(\log R - R + 1)^2 + (\log R^{(a)} - R^{(a)} + 1)^2 + (\log R^{(b)} - R^{(b)} + 1)^2.$$

Note that $|\log x - x + 1| \leq |x - 1|^r \max(-\log x, 1)$ for any $x > 0$ and any $1 < r \leq 2$. By Holder's inequality, we have

$$\begin{aligned} \mathbb{E}[(\log R - R + 1)^2 \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^2] &\leq \mathbb{E}[(R - 1)^{2r} \max(-\log R, 1)^2 \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^2] \\ &\leq \mathbb{E}[(R - 1)^{2rs}]^{1/s} \mathbb{E}[\max(-\log R, 1)^{2t} \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^{2t}]^{1/t} \end{aligned}$$

for any $s, t \geq 1$ satisfying $1/s + 1/t = 1$.

Note that $\mathbb{E}[R - 1] = 0$. Hence, if $2rs \leq p$, then it follows from [Lemma 3.1](#) that

$$\mathbb{E}[(R - 1)^{2rs}] \leq \frac{C_{2sr}}{M_{\ell}^{sr}} \mathbb{E}[|f(x; y^*)/p(y^*|\theta) - 1|^{2rs}],$$

where $\mathbb{E}[|f(x; y^*)/p(y^*|\theta) - 1|^{2rs}] < \infty$. Notice that the function $\max(-\log x, 1)^{2t}$ is convex for $x > 0$. Thus, applying Jensen's inequality and using $f(x_i; y^*) > 0$, we have

$$\begin{aligned} \max(-\log R, 1)^{2t} &= \max\left(-\log \frac{1}{M_{\ell}} \sum_{i=1}^{M_{\ell}} \frac{f(x_i; y^*)}{p(y^*|\theta)}, 1\right)^{2t} \\ &\leq \frac{1}{M_{\ell}} \sum_{i=1}^{M_{\ell}} \max\left(-\log \frac{f(x_i; y^*)}{p(y^*|\theta)}, 1\right)^{2t} \\ &\leq 1 + \frac{1}{M_{\ell}} \sum_{i=1}^{M_{\ell}} \left| \log \frac{f(x_i; y^*)}{p(y^*|\theta)} \right|^{2t}. \end{aligned}$$

As a result, as long as $2t \leq q$, we have

$$(3.6) \quad \begin{aligned} &\mathbb{E}[\max(-\log R, 1)^{2t} \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^{2t}] \\ &\leq \mathbb{E}\left[\left(1 + \left| \log \frac{f(x; y^*)}{p(y^*|\theta)} \right|^{2t}\right) \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^{2t}\right] < \infty. \end{aligned}$$

Particularly, we take $s = q/(q - 2)$, $t = q/2$ and $r = \min(p(q - 2)/(2q), 2)$. Since $(p - 2)(q - 2) > 4$, $r > 1$. Therefore, $\mathbb{E}[(\log R - R + 1)^2 \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^2] = O(M_{\ell}^{-r})$. This argument holds also by replacing R with $R^{(a)}$ or $R^{(b)}$. We thus have $\mathbb{E}[\Delta\psi_{\theta,\ell}^2 \|\nabla_{\lambda} \log q_{\lambda}(\theta)\|_2^2] = O(M_{\ell}^{-r}) = O(2^{-r\ell})$. \square

It should be noticed that [Theorem 3.1](#) requires $f(x; y^*) > 0$ w.p.1. If not, the inequalities in (3.6) do not hold. This implies that our result rules out the case of indicator functions in formulating likelihoods.

3.2. Re-parameterization gradient. Assume that there exists a transformation $x = \Lambda(\mathbf{v}; \theta) \sim p(x|\theta)$, where $\mathbf{v} \sim p_2(\mathbf{v})$ independently of θ and $\nabla_{\theta}\Lambda(\mathbf{v}; \theta)$ exists. Using $\theta = \Gamma(\mathbf{u}; \lambda)$ as before gives $x = \Lambda(\mathbf{v}; \Gamma(\mathbf{u}; \lambda))$. Allowing the interchange of expectation and differentiation, the gradient (2.4) is then rewritten as

$$\begin{aligned} \nabla_{\lambda} L(\lambda) &= \mathbb{E}_{\mathbf{u}}[\nabla_{\lambda}\Gamma(\mathbf{u}; \lambda) \cdot (\nabla_{\theta} \log \mathbb{E}_x[f(x; y^*)] + \nabla_{\theta} \log p(\theta) - \nabla_{\theta} \log q_{\lambda}(\theta))] \\ &= \mathbb{E}_{\mathbf{u}}[\nabla_{\lambda}\Gamma(\mathbf{u}; \lambda) \cdot (\nabla_{\theta} \log \mathbb{E}_{\mathbf{v}}[f(x; y^*)] + \nabla_{\theta} \log p(\theta) - \nabla_{\theta} \log q_{\lambda}(\theta))] \\ &= \mathbb{E}_{\mathbf{u}} \left[\nabla_{\lambda}\Gamma(\mathbf{u}; \lambda) \cdot \left(\frac{\mathbb{E}_{\mathbf{v}}[\nabla_{\theta} f(x; y^*)]}{\mathbb{E}_{\mathbf{v}}[f(x; y^*)]} + \nabla_{\theta} \log p(\theta) - \nabla_{\theta} \log q_{\lambda}(\theta) \right) \right] \\ &= \mathbb{E}_{\mathbf{u}} \left[\nabla_{\lambda}\Gamma(\mathbf{u}; \lambda) \cdot \left(\frac{\mathbb{E}_{\mathbf{v}}[\nabla_{\theta}\Lambda(\mathbf{v}; \theta)\nabla_x f(x; y^*)]}{\mathbb{E}_{\mathbf{v}}[f(x; y^*)]} + \nabla_{\theta} \log p(\theta) - \nabla_{\theta} \log q_{\lambda}(\theta) \right) \right], \end{aligned}$$

where $\nabla_{\lambda}\Gamma(\mathbf{u}; \lambda)$ is the Jacobian matrix with entries $[\nabla_{\lambda}\Gamma(\mathbf{u}; \lambda)]_{ij} = \partial\Gamma_j(\mathbf{u}; \lambda)/\partial\lambda_i$. Define

$$(3.7) \quad \text{RP}_N(\lambda) = \nabla_{\lambda}\Gamma(\mathbf{u}; \lambda) \cdot \left(\frac{\nabla_{\theta}\hat{p}_N(y^*|\theta)}{\hat{p}_N(y^*|\theta)} + \nabla_{\theta} \log p(\theta) - \nabla_{\theta} \log q_{\lambda}(\theta) \right),$$

where

$$\hat{p}_N(y^*|\theta) = \frac{1}{N} \sum_{i=1}^N f(x_i; y^*) \text{ with } x_i = \Lambda(\mathbf{v}_i; \theta),$$

$$\nabla_{\theta}\hat{p}_N(y^*|\theta) = \frac{1}{N} \sum_{i=1}^N \nabla_{\theta} f(x_i; y^*) = \frac{1}{N} \sum_{i=1}^N \nabla_{\theta}\Lambda(\mathbf{v}_i; \theta)\nabla_x f(x_i; y^*),$$

with $[\nabla_{\theta}\Lambda(\mathbf{v}; \theta)]_{ij} = \partial\Lambda_j(\mathbf{v}; \theta)/\partial\theta_i$ and $\mathbf{v}_i \sim p_2(\mathbf{v})$ independently. The estimator (3.7) is also biased. Now we take

$$\tilde{\psi}_{\theta, M_{\ell}} = \frac{\nabla_{\theta}\hat{p}_{M_{\ell}}(y^*|\theta)}{\hat{p}_{M_{\ell}}(y^*|\theta)} = \frac{\sum_{i=1}^{M_{\ell}} \nabla_{\theta}\Lambda(\mathbf{v}_i; \theta)\nabla_x f(x_i; y^*)}{\sum_{i=1}^{M_{\ell}} f(x_i; y^*)},$$

to differ from $\psi_{\theta, M_{\ell}}$ in the SF method. Analogously, we take $\Delta\tilde{\psi}_{\theta, 0} = \tilde{\psi}_{\theta, M_0}$. For $\ell \geq 1$, we use an antithetic coupling estimator again

$$(3.8) \quad \Delta\tilde{\psi}_{\theta, \ell} = \tilde{\psi}_{\theta, M_{\ell}} - \frac{1}{2} \left(\tilde{\psi}_{\theta, M_{\ell-1}}^{(a)} + \tilde{\psi}_{\theta, M_{\ell-1}}^{(b)} \right),$$

where

$$\begin{aligned} \tilde{\psi}_{\theta, M_{\ell-1}}^{(a)} &= \frac{\sum_{i=1}^{M_{\ell-1}} \nabla_{\theta}\Lambda(\mathbf{v}_i; \theta)\nabla_x f(x_i; y^*)}{\sum_{i=1}^{M_{\ell-1}} f(x_i; y^*)}, \\ \tilde{\psi}_{\theta, M_{\ell-1}}^{(b)} &= \frac{\sum_{i=M_{\ell-1}+1}^{M_{\ell}} \nabla_{\theta}\Lambda(\mathbf{v}_i; \theta)\nabla_x f(x_i; y^*)}{\sum_{i=M_{\ell-1}+1}^{M_{\ell}} f(x_i; y^*)}. \end{aligned}$$

Define

$$(3.9) \quad \text{RPM}_{\text{MLMC}}(\lambda) = \nabla_{\lambda}\Gamma(\mathbf{u}; \lambda) \cdot \left(\frac{\Delta\tilde{\psi}_{\theta, I}}{w_I} + \nabla_{\theta} \log p(\theta) - \nabla_{\theta} \log q_{\lambda}(\theta) \right),$$

where $\theta = \Gamma(\mathbf{u}; \lambda)$ and w_I is defined as in the SF method. For any number of outer samples $S \geq 1$, the gradient estimator

$$\widehat{\nabla_\lambda L}^{\text{RP}}(\lambda) = \frac{1}{S} \sum_{i=1}^S \text{RP}_{\text{MLMC}}^{(i)}(\lambda),$$

is unbiased, where $\text{RP}_{\text{MLMC}}^{(i)}(\lambda)$ are iid copy of $\text{RP}_{\text{MLMC}}(\lambda)$.

Similarly, to ensure a finite variance and finite expected computational cost of $\text{RP}_{\text{MLMC}}(\lambda)$, it suffices to show $\mathbb{E}[\|\nabla_\lambda \Gamma(\mathbf{u}; \lambda) \cdot \Delta \tilde{\psi}_{\theta, \ell}\|_2^2] = O(2^{-r\ell})$ for some $r > 1$. This can be achieved as shown in the following theorem.

THEOREM 3.2. *If*

$$\sup_x \|\nabla_\lambda \log f(x; y^*)\|_\infty < \infty,$$

where $x = \Lambda(\mathbf{v}; \Gamma(\mathbf{u}; \lambda))$, and assume that there exists $p > 2$ such that

$$\mathbb{E} \left[\left| \frac{f(x, y^*)}{p(y^*|\theta)} \right|^p \right] < \infty,$$

then

$$\mathbb{E}[\|\nabla_\lambda \Gamma(\mathbf{u}; \lambda) \Delta \tilde{\psi}_{\theta, \ell}\|_2^2] = O(2^{-r\ell}) \text{ with } r = \min(p/2, 2) \in (1, 2].$$

Proof. The proof follows an argument similar to Theorem 3.1 in [18], which considered a nested expectation involving a ratio of two inner conditional expectations. \square

4. Parameterizations in Gaussian variational family. Throughout this paper, we use the Gaussian family $N(\mu, \Sigma)$ as the variational family. For the SF method, we take the variational parameters as $\lambda = (\mu, \text{vech}(C))$, where C is the Cholesky decomposition (lower triangular) of Σ^{-1} and $\text{vech}(C)$ denotes a vector obtained by stacking the lower triangular elements of C . The number of variational parameters $d_\lambda = p + p(p+1)/2$. Since $\log q_\lambda(\theta) = \log |\det(C)| - \frac{1}{2}(\theta - \mu)^\top C C^\top (\theta - \mu)$, $\nabla_\lambda \log q_\lambda(\theta) = (\nabla_\mu \log q_\lambda(\theta), \nabla_{\text{vech}(C)} \log q_\lambda(\theta))$ with

$$\begin{aligned} \nabla_\mu \log q_\lambda(\theta) &= C C^\top (\theta - \mu), \\ \nabla_{\text{vech}(C)} \log q_\lambda(\theta) &= \text{vech}(\text{diag}(1/C) - (\theta - \mu)(\theta - \mu)^\top C), \end{aligned}$$

where $\text{diag}(1/C)$ denotes the diagonal matrix with the same dimensions as C with i th diagonal entry $1/C_{ii}$. Note that the score function $\nabla_\lambda \log q_\lambda(\theta)$ is model-free. The SF estimator $\text{SF}_N(\lambda)$ can be easily obtained by (3.1). It is common to use control variate (CV) to reduce the noise in estimating the gradient [24, 29]. Note that $\mathbb{E}[\nabla_\lambda \log q_\lambda(\theta)] = 0$. For any constant vector $c = (c_1, \dots, c_p) \in \mathbb{R}^p$, the estimator is also unbiased for the gradient,

$$\text{SF}_{\text{MLMC}}^{\text{CV}}(\lambda, c) = \nabla_\lambda \log q_\lambda(\theta) \left[\frac{\Delta \psi_{\theta, I}}{w_I} + \log p(\theta) - \log q_\lambda(\theta) - c \right].$$

We can take an optimal c_i to minimize the variance of the i th entry of $\text{SF}_{\text{MLMC}}^{\text{CV}}(\lambda, c)$. Solving

$$c_i^* = \arg \min_{c_i \in \mathbb{R}} \text{Var}(\text{SF}_{\text{MLMC}, i}^{\text{CV}}(\lambda, c_i))$$

gives

$$(4.1) \quad c_i^* = \frac{\mathbb{E}[(\nabla_{\lambda_i} \log q_\lambda(\theta))^2 \xi]}{\mathbb{E}[(\nabla_{\lambda_i} \log q_\lambda(\theta))^2]} = \frac{\text{Cov}(\nabla_{\lambda_i} \log q_\lambda(\theta), \nabla_{\lambda_i} \log q_\lambda(\theta) \xi)}{\text{Var}(\nabla_{\lambda_i} \log q_\lambda(\theta))},$$

where $\xi = \frac{\Delta\psi_{\theta,I}}{w_I} + \log p(\theta) - \log q_\lambda(\theta)$. In practice, c_i^* ($i = 1, \dots, p$) are estimated by using the samples in the previous iteration. The whole procedure is summarized in [Algorithm 4.1](#).

Algorithm 4.1 Unbiased MLMC with the SF gradient estimator

- 1: Initialize $\lambda^{(0)} = (\mu^{(0)}, \text{vech}(C^{(0)}))$, $t = 0$, M the number of outer samples, $\alpha \in (1, r)$ and $w_\ell \propto 2^{-\alpha\ell}$ such that $\sum_{\ell=0}^{\infty} w_\ell = 1$ and all $w_\ell > 0$.
- 2: Repeat
 - (a) Generate $\theta_1^{(t)}, \dots, \theta_m^{(t)} \sim N(\mu^{(t)}, (C^{(t)}C^{(t)\top})^{-1})$ independently and $I_1^{(t)}, \dots, I_m^{(t)}$ independently and randomly with probability w_ℓ .
 - (b) Let $n_i = M_0 2^{I_i^{(t)}}$. For $i = 1, \dots, m$, generate $x_{i1}^{(t)}, \dots, x_{in_i}^{(t)} \sim p(x|\theta_i^{(t)})$ independently. Compute the associated samples of the correction $\Delta\psi_{\theta,I}$, denoted by $\Delta\psi_i^{(t)}$, $i = 1, \dots, m$.
 - (c) Estimate c^* defined by (4.1) by the samples $\theta_i^{(t)}$, $I_i^{(t)}$, $\Delta\psi_i^{(t)}$, $i = 1, \dots, m$, resulting in $c^{(t)}$.
 - (d) If $t > 0$, compute the gradient estimator

$$\widehat{\nabla_\lambda L}^{\text{SF}}(\lambda^{(t)}) = \frac{1}{m} \sum_{i=1}^m \nabla_\lambda \log q_\lambda(\theta_i^{(t)}) \left[\frac{\Delta\psi_i^{(t)}}{w_{I_i^{(t)}}} + \log p(\theta_i^{(t)}) - \log q_{\lambda^{(t)}}(\theta_i^{(t)}) - c^{(t-1)} \right],$$

and the ELBO estimator

$$\text{LB}(\lambda^{(t)}) = \frac{1}{S} \sum_{i=1}^S \frac{\Delta\psi_i^{(t)}}{w_{I_i^{(t)}}} + \log p(\theta_i^{(t)}) - \log q_{\lambda^{(t)}}(\theta_i^{(t)}).$$

Update the VB parameter:

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t \widehat{\nabla_\lambda L}^{\text{SF}}(\lambda^{(t)}).$$

If $t = 0$, then set $\lambda^{(t+1)} = \lambda^{(t)}$. Note that this step is used to initialize c^* rather than updating the VB parameter.

(e) $t = t + 1$

until some stopping rule is satisfied.

Using the RP method, we take the variational parameter as $\lambda = (\mu, \text{vech}(L))$, where L is the Cholesky decomposition of Σ , which is different from the parameterizations in the SF method. For this case, $\theta = \Gamma(\mathbf{u}; \lambda) = \mu + L\mathbf{u} \sim N(\mu, \Sigma)$, where $\mathbf{u} \in \mathbb{R}^{p \times 1}$ is a standard normal. Let

$$G = \frac{\Delta\tilde{\psi}_{\theta,I}}{w_I} + \nabla_\theta \log p(\theta) - \nabla_\theta \log q_\lambda(\theta) \in \mathbb{R}^{p \times 1},$$

where $\Delta\tilde{\psi}_{\theta,\ell}$ is given by (3.8) and $\nabla_\theta \log q_\lambda(\theta) = -\Sigma^{-1}(\theta - \mu) = -(LL^\top)^{-1}(\theta - \mu)$. Then the RP estimator is given by

$$\text{RP}_{\text{MLMC}}(\lambda) = (G, \text{vech}(G\mathbf{u}^\top)) \in \mathbb{R}^{d_\lambda \times 1}.$$

The second term $\nabla_\theta \log p(\theta)$ in G depends on the prior. Particularly, if the prior is normally distributed, say, $N(\mu_0, \Sigma_0)$, then $\nabla_\theta \log p(\theta) = -\Sigma_0^{-1}(\theta - \mu_0)$. It is crucial to

work out the term $\nabla_{\theta}\Lambda(\mathbf{v};\theta)\nabla_x f(x;y^*)$ used in $\Delta\tilde{\psi}_{\theta,\ell}$, which is model-specific. The whole procedure for the RP method is summarized in [Algorithm 4.2](#).

Algorithm 4.2 Unbiased MLMC with the RP estimator

- 1: Initialize $\lambda^{(0)} = (\mu^{(0)}, \text{vech}(L^{(0)}))$, $t = 0$, M the number of outer samples, $\alpha \in (1, r)$ and $w_{\ell} \propto 2^{-\alpha\ell}$ such that $\sum_{\ell=0}^{\infty} w_{\ell} = 1$ and all $w_{\ell} > 0$.
- 2: Repeat
 - (a) Generate $\mathbf{u}_1^{(t)}, \dots, \mathbf{u}_m^{(t)} \sim N(0, I_p)$ independently and set $\theta_i^{(t)} = \mu^{(t)} + L^{(t)}\mathbf{u}_i^{(t)}$. Generate $I_1^{(t)}, \dots, I_m^{(t)}$ independently and randomly with probability w_{ℓ} .
 - (b) Let $n_i = M_0 2^{I_i^{(t)}}$. For $i = 1, \dots, m$, generate $\mathbf{v}_{i1}^{(t)}, \dots, \mathbf{v}_{in_i}^{(t)} \sim p_2(\mathbf{v})$ independently and set $x_{ij}^{(t)} = \Lambda(\mathbf{v}_{ij}^{(t)}; \theta_i^{(t)})$, $j = 1, \dots, n_i$. Compute the associated samples of the corrections $\Delta\psi_{\theta,I}$ and $\Delta\tilde{\psi}_{\theta,I}$, denoted by $\Delta\psi_i^{(t)}$ and $\Delta\tilde{\psi}_i^{(t)}$, respectively.
 - (c) Compute the gradient estimator

$$\widehat{\nabla_{\lambda} L}^{\text{RP}}(\lambda^{(t)}) = \frac{1}{m} \sum_{i=1}^m (G_i^{(t)}, \text{vech}(G_i^{(t)} \mathbf{u}_i^{\top})),$$

where

$$G_i^{(t)} = \frac{\Delta\tilde{\psi}_i^{(t)}}{w_{I_i^{(t)}}} + \nabla_{\theta} \log p(\theta_i^{(t)}) - \nabla_{\theta} \log q_{\lambda^{(t)}}(\theta_i^{(t)}),$$

and compute the ELBO estimator

$$\text{LB}(\lambda^{(t)}) = \frac{1}{S} \sum_{i=1}^S \frac{\Delta\psi_i^{(t)}}{w_{I_i^{(t)}}} + \log p(\theta_i^{(t)}) - \log q_{\lambda^{(t)}}(\theta_i^{(t)}).$$

Update the VB parameter:

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t \widehat{\nabla_{\lambda} L}^{\text{RP}}(\lambda^{(t)}).$$

(d) $t = t + 1$

until some stopping rule is satisfied.

Notice that not only the gradient estimators but also the ELBO estimators are unbiased in [Algorithms 4.1](#) and [4.2](#). The unbiased MLMC methods can be expected to estimate the ELBO more accurately.

5. Incorporating RQMC. We now incorporate RQMC sampling based scrambled (t, s) -sequences within the MLMC estimators. Quasi-Monte Carlo (QMC) is designed for computing expectations of $f(\mathbf{v})$ for $\mathbf{v} \sim U[0, 1]^s$. We should note that in our present context, the underlying distributions are not the form of uniforms. To fit QMC in practice, one must transform the base distribution $U[0, 1]^s$ to the underlying distributions. Suppose that there exists a transformation $\psi(\cdot)$ such $\psi(\mathbf{v}) \sim p$, where p is the underlying distribution. Below we subsume any such transformation $\psi(\cdot)$ into the definition of f .

To estimate $\mu = \int_{[0,1]^s} f(\mathbf{v})d\mathbf{v}$, QMC methods use a sample-mean estimator

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N f(\mathbf{v}_i),$$

where $\mathbf{v}_1, \dots, \mathbf{v}_N$ are the first N points of a low discrepancy sequence. By the Koksma-Hlawka inequality, we have

$$|\hat{\mu} - \mu| \leq V_{\text{HK}}(f)D^*(\mathbf{v}_1, \dots, \mathbf{v}_N),$$

where $V_{\text{HK}}(f)$ is the variation of the integrand $f(\cdot)$ in the sense of Hardy and Krause, and $D^*(\mathbf{v}_1, \dots, \mathbf{v}_N)$ is the star discrepancy of the point set $\{\mathbf{v}_1, \dots, \mathbf{v}_N\}$. For (t, s) -sequences, we have

$$D^*(\mathbf{v}_1, \dots, \mathbf{v}_N) = O(N^{-1}(\log N)^s) = O(N^{-1+\epsilon}),$$

where we use an arbitrarily small $\epsilon > 0$ for hiding the logarithm term throughout this paper. If f is of bounded variation in the sense of Hardy and Krause (BVHK), one gets a QMC error of $O(N^{-1+\epsilon})$. To get a practical error estimate, RQMC methods were introduced, see [22] for a review. In this paper, we use the scrambling technique proposed by [26] to randomize (t, s) -sequences. In RQMC, each $\mathbf{v}_i \sim U[0, 1]^s$ marginally, implying that $\hat{\mu}$ is unbiased for μ . More importantly, scrambled (t, s) -sequence retains a (t, s) -sequence w.p.1. This leads to

$$\text{Var}(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mu)^2] \leq V_{\text{HK}}(f)^2 D^*(\mathbf{v}_1, \dots, \mathbf{v}_N)^2,$$

where the expectation is taken with respect to the randomness of scrambling. Apparently, the RQMC variance is of $O(N^{-2+\epsilon})$ if f is of BVHK.

Now we focus on how to incorporate RQMC within the MLMC estimators. In fact, for both the SF and RP estimators, one needs to sample $\theta \sim q_\lambda(\theta)$, $x_1, \dots, x_{M_I} \sim p(x|\theta, I)$ and I from a discrete distribution with $P(I = i) = w_i$ as stated above. For each realization, the number of random variables depends on I , which takes values in \mathbb{N} . It is not possible to use a scrambled (t, s) -sequence to sample all random variables in a single run because we need determine the dimension s in advance. Instead, we use hybrid sequences within the MLMC estimators. Specifically, we still use MC to sample θ and I , but use RQMC in inner simulation. That is, x_1, \dots, x_{M_I} is based on a scrambled (t, s) -sequence. To this end, we assume that there exists a transformation Λ such that

$$x = \Lambda(\mathbf{v}; \theta) \sim p(x|\theta),$$

where $\mathbf{v} \sim U[0, 1]^s$. We then takes $x_i = \Lambda(\mathbf{v}_i; \theta)$ in the inner simulation, where $\mathbf{v}_1, \dots, \mathbf{v}_{M_I}$ are the first M_I points of a scrambled (t, s) -sequence. Since RQMC estimates are unbiased, the replacement of RQMC will not change the unbiasedness of the gradient estimators.

We are ready to establish an RQMC version of [Theorem 3.1](#) for the SF gradient. We should note that [Theorem 3.1](#) may not be extended to the RQMC setting since [Lemma 3.1](#) holds only for iid samples. Recently, for proving strong law of large numbers for scrambled net integration, [28] showed that $\mathbb{E}[|\bar{X}_N|^p] \leq C_p N^{1-p}$ for $p \in (1, 2)$ via the Riesz-Thorin interpolation theorem, where \bar{X}_N is an average of N RQMC samples of X with $\mathbb{E}[X] = 0$. However, this result is not for the case $p > 2$ required in [Lemma 3.1](#). It is not clear whether the RQMC version of [Lemma 3.1](#) holds. This is left for future research. The theorem we provide below is totally different from [Theorem 3.1](#), and the proof of which does not depend on [Lemma 3.1](#).

THEOREM 5.1. *Suppose that samples $x_i = \Lambda(\mathbf{v}_i; \theta)$, $i = 1, \dots, M_\ell$ are used in the SF estimator [subsection 3.1](#), where $\mathbf{v}_i \in [0, 1]^s$ are the first M_ℓ points of a scrambled (t, s) -sequence. If*

$$\mathbb{E} \left[\frac{V_{\text{HK}}(f_\theta)^2 \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2}{f_\theta(\mathbf{v})^2} \right] < \infty,$$

where $\mathbf{v} \sim U[0, 1]^s$, $f_\theta(\mathbf{v}) = f(\Lambda(\mathbf{v}; \theta); y^*)$, and $\Lambda(\mathbf{v}; \theta) \sim p(x|\theta)$, then we have

$$\mathbb{E}[\Delta\psi_{\theta, \ell}^2 | \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2] = O(2^{-r\ell}) \text{ with } r = 2 - \epsilon$$

for arbitrarily small $\epsilon > 0$.

Proof. Note that $p(y^*|\theta) = \mathbb{E}[f_\theta(\mathbf{v})|\theta]$. Let

$$P_\ell = \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} f(x_i; y^*) = \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} f_\theta(\mathbf{v}_i),$$

with $P_{\ell-1}^{(a)} = \frac{1}{M_{\ell-1}} \sum_{i=1}^{M_{\ell-1}} f_\theta(\mathbf{v}_i)$, and $P_{\ell-1}^{(b)} = \frac{1}{M_{\ell-1}} \sum_{i=M_{\ell-1}+1}^{M_\ell} f_\theta(\mathbf{v}_i)$. All of them are RQMC estimators for $p(y^*|\theta)$. We have

$$\Delta\psi_{\theta, \ell} = [\log P_\ell - \log p(y^*|\theta)] - \frac{1}{2}[(\log P_\ell^{(a)} - \log p(y^*|\theta)) + (\log P_\ell^{(b)} - \log p(y^*|\theta))].$$

Applying Jensen's inequality gives

$$\Delta\psi_{\theta, \ell}^2 \leq 2(\log P_\ell - \log p(y^*|\theta))^2 + (\log P_\ell^{(a)} - \log p(y^*|\theta))^2 + (\log P_\ell^{(b)} - \log p(y^*|\theta))^2.$$

Note that $|\log t| \leq \max(1, 1/t)|t - 1| \leq (1 + 1/t)|t - 1|$ for any $t > 0$. We thus have

$$|\log P_\ell - \log p(y^*|\theta)| \leq (1/p(y^*|\theta) + 1/P_\ell)|P_\ell - p(y^*|\theta)|.$$

By the Koksma-Hlawka inequality, we have

$$|P_\ell - p(y^*|\theta)| \leq V_{\text{HK}}(f_\theta)D_\ell,$$

where $D_\ell = D^*(\mathbf{v}_1, \dots, \mathbf{v}_{M_\ell})$. This implies that

$$(\log P_\ell - \log p(y^*|\theta))^2 \leq 2V_{\text{HK}}(f_\theta)^2 D_\ell^2 \left(\frac{1}{p(y^*|\theta)^2} + \frac{1}{P_\ell^2} \right).$$

Let $H(\theta) = V_{\text{HK}}(f_\theta) \|\nabla_\lambda \log q_\lambda(\theta)\|_2$. We then have

$$\mathbb{E}[(\log P_\ell - \log p(y^*|\theta))^2 | \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2] \leq 2D_\ell^2 \left(\mathbb{E} \left[\frac{H(\theta)^2}{p(y^*|\theta)^2} \right] + \mathbb{E} \left[\frac{H(\theta)^2}{P_\ell^2} \right] \right).$$

By Jensen's inequality, we have

$$(5.1) \quad \frac{1}{P_\ell^2} = \left(\frac{1}{\frac{1}{M_\ell} \sum_{i=1}^{M_\ell} f_\theta(\mathbf{v}_i)} \right)^2 \leq \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} \frac{1}{f_\theta(\mathbf{v}_i)^2}.$$

By the unbiasedness of RQMC estimators and the law of total expectation,

$$\begin{aligned} \mathbb{E} \left[\frac{H(\theta)^2}{P_\ell^2} \right] &\leq \mathbb{E} \left[\frac{H(\theta)^2}{M_\ell} \sum_{i=1}^{M_\ell} \frac{1}{f(x_i; y^*)^2} \right] \\ &= \mathbb{E} \left[H(\theta)^2 \mathbb{E} \left[\frac{1}{M_\ell} \sum_{i=1}^{M_\ell} \frac{1}{f(x_i; y^*)^2} \middle| \theta \right] \right] \\ &= \mathbb{E} \left[H(\theta)^2 \mathbb{E} \left[\frac{1}{f(x; y^*)^2} \middle| \theta \right] \right] = \mathbb{E} \left[\frac{H(\theta)^2}{f(x; y^*)^2} \right] < \infty. \end{aligned}$$

On the other hand, by using Jensen's inequality and the law of total expectation again,

$$\begin{aligned} \mathbb{E} \left[\frac{H(\theta)^2}{p(y^*|\theta)^2} \right] &= \mathbb{E} \left[\frac{H(\theta)^2}{\mathbb{E}[f(x; y^*)|\theta]^2} \right] \\ &\leq \mathbb{E} \left[H(\theta)^2 \mathbb{E} \left[\frac{1}{f(x; y^*)^2} \middle| \theta \right] \right] = \mathbb{E} \left[\frac{H(\theta)^2}{f(x; y^*)^2} \right] < \infty. \end{aligned}$$

We therefore have

$$\mathbb{E}[(\log P_\ell - \log p(y^*|\theta))^2 \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2] = O(D_\ell^2) = O(M_\ell^{-2+\epsilon}) = O(2^{-r\ell})$$

for $r = 2 - \epsilon$ and any $\epsilon > 0$. This argument holds also by replacing P_ℓ with $P_\ell^{(a)}$ or $P_\ell^{(b)}$. We thus have $\mathbb{E}[\Delta\psi_{\theta,\ell}^2 \|\nabla_\lambda \log q_\lambda(\theta)\|_2^2] = O(2^{-r\ell})$. \square

We next establish an RQMC version of [Theorem 3.2](#) for the RP gradient. [Theorem 3.2](#) cannot be extended to the RQMC setting since its proof depends on [Lemma 3.1](#) as well.

THEOREM 5.2. *Suppose that samples $x_i = \Lambda(\mathbf{v}_i; \theta)$, $i = 1, \dots, M_\ell$ in the RP estimator [\(3.9\)](#), where $\mathbf{v}_i \in [0, 1]^s$ are the first M_ℓ points of a scrambled (t, s) -sequence. If*

$$\mathbb{E} \left[\frac{\|\nabla_\lambda \Gamma(\mathbf{u}; \lambda)\|_{\max}^2}{p(y^*|\theta)^2} \left(\frac{(\|\nabla p(y^*|\theta)\|_2^2 + \|V_{\text{HK}}(\nabla_\theta f_\theta)\|_2^2) V_{\text{HK}}(f_\theta)^2}{f_\theta(\mathbf{v})^2} + \|V_{\text{HK}}(\nabla_\theta f_\theta)\|_2^2 \right) \right]$$

is finite, where $\mathbf{v} \sim U[0, 1]^s$, $f_\theta(\mathbf{v}) = f(\Lambda(\mathbf{v}; \theta); y^*)$, $\Lambda(\mathbf{v}; \theta) \sim p(x|\theta)$, $\theta = \Gamma(\mathbf{u}; \lambda) \sim q_\lambda(\theta)$, $V_{\text{HK}}(\nabla_\theta f_\theta)$ denotes a vector of $V_{\text{HK}}(\partial_{\theta_i} f_\theta)$, and $\|A\|_{\max}$ denotes the largest absolute value of the entries of the matrix A , we have

$$\mathbb{E}[\|\nabla_\lambda \Gamma(\mathbf{u}; \lambda) \Delta\tilde{\psi}_{\theta,\ell}\|_2^2] = O(2^{-r\ell}) \text{ with } r = 2 - \epsilon$$

for arbitrarily small $\epsilon > 0$.

Proof. We use the notations P_ℓ , $P_{\ell-1}^{(a)}$ and $P_{\ell-1}^{(b)}$ defined in the proof of [Theorem 5.1](#), and define

$$\mathcal{N}_\ell = \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} \nabla_\theta \Lambda(\mathbf{v}_i; \theta) \nabla_x f(x_i; y^*) = \frac{1}{M_\ell} \sum_{i=1}^{M_\ell} \nabla_\theta f_\theta(\mathbf{v}_i),$$

with $\mathcal{N}_{\ell-1}^{(a)} = \frac{1}{M_{\ell-1}} \sum_{i=1}^{M_{\ell-1}} \nabla_\theta f_\theta(\mathbf{v}_i)$, and $\mathcal{N}_{\ell-1}^{(b)} = \frac{1}{M_{\ell-1}} \sum_{i=M_{\ell-1}+1}^{M_\ell} \nabla_\theta f_\theta(\mathbf{v}_i)$. It is clear that $\mathbb{E}[\mathcal{N}_\ell|\theta] = \mathbb{E}[\mathcal{N}_{\ell-1}^{(a)}|\theta] = \mathbb{E}[\mathcal{N}_{\ell-1}^{(b)}|\theta] = \nabla_\theta p(y^*|\theta)$, and $\mathbb{E}[P_\ell|\theta] = \mathbb{E}[P_{\ell-1}^{(a)}|\theta] = \mathbb{E}[P_{\ell-1}^{(b)}|\theta] = p(y^*|\theta)$. Note that

$$\begin{aligned} \Delta\tilde{\psi}_{\theta,\ell} &= \frac{\mathcal{N}_\ell}{P_\ell} - \frac{1}{2} \left(\frac{\mathcal{N}_{\ell-1}^{(a)}}{P_{\ell-1}^{(a)}} + \frac{\mathcal{N}_{\ell-1}^{(b)}}{P_{\ell-1}^{(b)}} \right) \\ &= \left[\frac{\mathcal{N}_\ell}{P_\ell} - \frac{\nabla_\theta p(y^*|\theta)}{p(y^*|\theta)} \right] - \frac{1}{2} \left[\frac{\mathcal{N}_{\ell-1}^{(a)}}{P_{\ell-1}^{(a)}} - \frac{\nabla_\theta p(y^*|\theta)}{p(y^*|\theta)} \right] - \frac{1}{2} \left[\frac{\mathcal{N}_{\ell-1}^{(b)}}{P_{\ell-1}^{(b)}} - \frac{\nabla_\theta p(y^*|\theta)}{p(y^*|\theta)} \right]. \end{aligned}$$

Let $\mathcal{N}_{\ell,i}$ be the i th entry of \mathcal{N}_ℓ , which is an unbiased estimator for $\partial_{\theta_i} p(y^*|\theta)$. By the triangle inequality, we find that

$$(5.2) \quad \begin{aligned} \left(\frac{\mathcal{N}_{\ell,i}}{P_\ell} - \frac{\partial_{\theta_i} p(y^*|\theta)}{p(y^*|\theta)} \right)^2 &= \left(\frac{\mathcal{N}_{\ell,i}}{P_\ell} - \frac{\mathcal{N}_{\ell,i}}{p(y^*|\theta)} + \frac{\mathcal{N}_{\ell,i}}{p(y^*|\theta)} - \frac{\partial_{\theta_i} p(y^*|\theta)}{p(y^*|\theta)} \right)^2 \\ &\leq \frac{2}{p(y^*|\theta)^2} \left[\frac{\mathcal{N}_{\ell,i}^2}{P_\ell^2} (P_\ell - p(y^*|\theta))^2 + (\mathcal{N}_{\ell,i} - \partial_{\theta_i} p(y^*|\theta))^2 \right]. \end{aligned}$$

By the Koksma-Hlawka inequality, we have

$$\begin{aligned} |P_\ell - p(y^*|\theta)| &\leq V_{\text{HK}}(f_\theta) D_\ell, \\ |\mathcal{N}_{\ell,i} - \partial_{\theta_i} p(y^*|\theta)| &\leq V_{\text{HK}}(\partial_{\theta_i} f_\theta) D_\ell, \end{aligned}$$

where $D_\ell = D^*(\mathbf{v}_1, \dots, \mathbf{v}_{M_\ell})$.

For large enough ℓ , it is reasonable to assume that $D_\ell < 1$. Together with (5.1) and (5.2), we then have

$$\begin{aligned} &\left(\frac{\mathcal{N}_{\ell,i}}{P_\ell} - \frac{\partial_{\theta_i} p(y^*|\theta)}{p(y^*|\theta)} \right)^2 \\ &\leq \frac{2D_\ell^2}{p(y^*|\theta)^2} \left[\frac{\mathcal{N}_{\ell,i}^2}{P_\ell^2} V_{\text{HK}}(f_\theta)^2 + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2 \right] \\ &\leq \frac{4D_\ell^2}{p(y^*|\theta)^2} \left[\frac{\partial_{\theta_i} p(y^*|\theta)^2 + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2}{P_\ell^2} V_{\text{HK}}(f_\theta)^2 + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2 \right] \\ &\leq \frac{4D_\ell^2}{p(y^*|\theta)^2} \left[\frac{(\partial_{\theta_i} p(y^*|\theta)^2 + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2) V_{\text{HK}}(f_\theta)^2}{M_\ell} \sum_{i=1}^{M_\ell} \frac{1}{f_\theta(\mathbf{v}_i)^2} + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2 \right]. \end{aligned}$$

Let n_r and n_c be the number of rows and columns of the Jacobian matrix $\nabla_\lambda \Gamma(\mathbf{u}; \lambda)$, respectively, and $M_\lambda = \|\nabla_\lambda \Gamma(\mathbf{u}; \lambda)\|_{\max}$. As a result,

$$\begin{aligned} &\mathbb{E} \left[\left\| \nabla_\lambda \Gamma(\mathbf{u}; \lambda) \cdot \left(\frac{\mathcal{N}_\ell}{P_\ell} - \frac{\nabla_\theta p(y^*|\theta)}{p(y^*|\theta)} \right) \right\|_2^2 \right] \\ &\leq n_c n_r \mathbb{E} \left[\sum_{i=1}^{n_r} M_\lambda^2 \left(\frac{\mathcal{N}_{\ell,i}}{P_\ell} - \frac{\partial_{\theta_i} p(y^*|\theta)}{p(y^*|\theta)} \right)^2 \right] \\ &\leq C_\ell \mathbb{E} \left[\frac{M_\lambda^2}{p(y^*|\theta)^2} \sum_{i=1}^{n_r} \left(\frac{(\partial_{\theta_i} p(y^*|\theta)^2 + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2) V_{\text{HK}}(f_\theta)^2}{f_\theta(\mathbf{v})^2} + V_{\text{HK}}(\partial_{\theta_i} f_\theta)^2 \right) \right] \\ &= C_\ell \mathbb{E} \left[\frac{M_\lambda^2}{p(y^*|\theta)^2} \left(\frac{(\|\nabla p(y^*|\theta)\|_2^2 + \|\mathbf{V}_{\text{HK}}(\nabla_\theta f_\theta)\|_2^2) V_{\text{HK}}(f_\theta)^2}{f_\theta(\mathbf{v})^2} + \|\mathbf{V}_{\text{HK}}(\nabla_\theta f_\theta)\|_2^2 \right) \right] \\ &= O(M_\ell^{-2+\epsilon}) = O(2^{-r\ell}) \end{aligned}$$

with $C_\ell = 4n_c n_r D_\ell^2$ for $r = 2 - \epsilon$ and any $\epsilon > 0$. By a similar argument in the proof of Theorem 5.1, we have $\mathbb{E}[\|\nabla_\lambda \Gamma(\mathbf{u}; \lambda) \Delta \tilde{\psi}_{\theta, \ell}\|_2^2] = O(2^{-r\ell})$. \square

In Theorems 5.1 and 5.2, the integrands in RQMC quadratures need to be BVHK. For practical problems, it may be very hard to verify such a condition. Particularly, if the integrands are not smooth enough, the BVHK condition does not hold. For

such cases, one may get a lower rate r . For any integrand in $L^2[0, 1]^s$, scrambled nets have variance $o(1/N)$ without requiring the BVHK condition [27]. Additionally, for any fixed N , the scrambled nets variance is no worse than a constant times the MC variance. From this point of view, under the same conditions in [Theorems 3.1](#) and [3.2](#), we can expect that the rate r for RQMC is no worse than that of MC. Finally, we should note that the rates established in [Theorems 5.1](#) and [5.2](#) do not benefit from the antithetic coupling, implying that the results also hold for the usual way of coupling. One might get a better rate by taking account for the form of antithetic coupling.

There are some other ways to incorporate RQMC in MLMC. For example, one can use RQMC in the outer simulation. That is, the samples of θ are based on a scrambled (t, s') -sequence while the inner samples x_i and the samples of I are based on MC. To this end, assuming $\theta = \Gamma_\lambda(\mathbf{u}) \sim q_\lambda(\theta)$ with $\mathbf{u} \sim U[0, 1]^{s'}$, we take

$$\theta_i = \Gamma_\lambda(\mathbf{u}_i), \quad i = 1, \dots, S,$$

where $\mathbf{u}_1, \dots, \mathbf{u}_S$ are the first S points of a scrambled (t, s') sequence. Taking the SF gradient estimator [\(3.3\)](#) for instance, we have

$$\begin{aligned} \text{Var} \left(\widehat{\nabla_\lambda L}^{\text{SF}}(\lambda) \right) &= \mathbb{E} \left[\text{Var} \left(\widehat{\nabla_\lambda L}^{\text{SF}}(\lambda) | \theta_{\{1:S\}} \right) \right] + \text{Var} \left(\mathbb{E}[\widehat{\nabla_\lambda L}^{\text{SF}}(\lambda) | \theta_{\{1:S\}}] \right) \\ &= \frac{1}{S} \mathbb{E}[\text{Var}(\text{SF}_{\text{MLMC}}(\lambda) | \theta)] + \text{Var} \left(\frac{1}{S} \sum_{i=1}^S \mathbb{E}[\text{SF}_{\text{MLMC}}^{(i)}(\lambda) | \theta_i] \right) \\ (5.3) \quad &= \frac{1}{S} \mathbb{E}[\text{Var}(\text{SF}_{\text{MLMC}}(\lambda) | \theta)] + \text{Var} \left(\frac{1}{S} \sum_{i=1}^S H(\theta_i) \right), \end{aligned}$$

where $H(\theta) := \nabla_\lambda \log q_\lambda(\theta) [\log p(y^* | \theta) + \log p(\theta) - \log q_\lambda(\theta)]$, $\theta_{\{1:S\}} = \{\theta_1, \dots, \theta_S\}$ and $\text{Var}(\cdot)$ and $\mathbb{E}[\cdot]$ are applied component-wisely. The second term in [\(5.3\)](#) is $O(1/S)$ when the θ_i 's are generated using MC, while it should be $o(1/S)$ when the θ_i 's are generated using RQMC, or even better $O(S^{-2+\epsilon})$ if $H \circ \Gamma_\lambda$ is of BVHK. The first term in [\(5.3\)](#) is $O(1/S)$ for both cases. As a result, this strategy helps to reduce the variance in the outer sampling. Buchholz and Chopin [3] applied this strategy in ABC. They found that the resulting ABC estimate has a lower variance than the MC counter-part. However, the rate of convergence cannot be improved due to the first term in [\(5.3\)](#). This strategy cannot improve the rates r in [Theorems 3.1](#) and [3.2](#) either.

On the other hand, we can also use a two-stage RQMC strategy. In the outer samples, we use a scrambled (t, s') -sequence to simulate θ ; while in each inner simulation, we use another independent branch of scrambled (t, s) -sequence to sample x_i . This two-stage RQMC strategy helps to reduce the noise in both inner and outer simulations. In our numerical experiments, we shall compare the effects of the three ways of using RQMC in MLMC.

6. Numerical experiments.

6.1. Approximate Bayesian computation. ABC method is a generic tool in likelihood-free inference provided that it is easy to generate $y \sim p(y|\theta)$. However, ABC methods do not target the exact posterior, but an approximation to some extent. More specially, let $\mathcal{S}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^d$ be a vector of summary statistics, and $K_h(\cdot, \cdot)$ be a d -dimensional kernel density with bandwidth $h > 0$. ABC posterior density of θ is

given by

$$p_{\text{ABC}}(\theta|y^*) \propto p(\theta)\tilde{p}(y^*|\theta),$$

where the intractable likelihood is given by

$$(6.1) \quad \tilde{p}(\theta|y^*) = \int K_h(\mathcal{S}(y), \mathcal{S}(y^*))p(y|\theta)dy = \mathbb{E}_{p(y|\theta)}[K_h(\mathcal{S}(y), \mathcal{S}(y^*))].$$

To fit the form (2.2), one gets $f(y; y^*) := K_h(\mathcal{S}(y), \mathcal{S}(y^*))$, in which the latent variable x is replaced by y . To ensure $f(y; y^*) > 0$, we particularly take the Gaussian kernel

$$K_h(s, s^*) = (2\pi h)^{-d/2} \exp\left\{-\frac{(s - s^*)^\top (s - s^*)}{2h}\right\},$$

where d denotes the dimension of the summary statistics $\mathcal{S}(y)$. If $\mathcal{S}(y^*)$ is a sufficient statistic, then $p_{\text{ABC}}(\theta|y^*)$ converges to the exact posterior $p(\theta|y^*)$ as $h \rightarrow 0$. Otherwise, $p_{\text{ABC}}(\theta|y^*)$ converges to the posterior $p(\theta|\mathcal{S}(y^*))$ as $h \rightarrow 0$, where is a gap between $p(\theta|\mathcal{S}(y^*))$ and $p(\theta|y^*)$.

To apply the SF method, it suffices to provide the sample-mean likelihood estimator

$$\hat{p}_N(y^*|\theta) = \frac{1}{N} \sum_{i=1}^N K_h(\mathcal{S}(y^{[i]}), \mathcal{S}(y^*)),$$

where $y^{[i]}$ are iid sample of $p(y|\theta)$. To apply the RP methods, we need to find the mappings such that

$$\theta = \Gamma(\mathbf{u}; \lambda) \sim q_\lambda(\theta) \text{ and } y = \Lambda(\mathbf{v}; \theta) \sim p(y|\theta),$$

where the distributions of \mathbf{u} , \mathbf{v} do not depend on λ and θ , respectively. We also require the closed forms of $\nabla_y f(y; y^*)$, $\nabla_\theta \Lambda(\mathbf{v}; \theta)$, and $\nabla_\lambda \Gamma(\mathbf{u}; \lambda)$. Note that

$$\begin{aligned} \frac{\partial f(y; y^*)}{\partial y_i} &= \sum_{j=1}^d \frac{\partial K_h(\mathcal{S}(y), \mathcal{S}(y^*))}{\partial \mathcal{S}_j} \frac{\partial \mathcal{S}_j(y)}{\partial y_i} \\ &= \frac{K_h(\mathcal{S}(y), \mathcal{S}(y^*))}{h} \sum_{j=1}^d [\mathcal{S}_j(y^*) - \mathcal{S}_j(y)] \frac{\partial \mathcal{S}_j(y)}{\partial y_i}. \end{aligned}$$

As a result,

$$\nabla_y f(y; y^*) = \frac{K_h(\mathcal{S}(y), \mathcal{S}(y^*)) \nabla_y \mathcal{S}(y) [\mathcal{S}(y^*) - \mathcal{S}(y)]}{h}.$$

It reduces to verify and then compute the Jacobian matrix $\nabla_y \mathcal{S}(y)$. If we take the entire data as the summary statistics, then $\nabla_y \mathcal{S}(y)$ is an identity matrix. If the summary statistics $\mathcal{S}(y)$ are sample moments, $\nabla_y \mathcal{S}(y)$ can be easily computed. However, if the summary statistics $\mathcal{S}(y)$ are functions of sample quantiles, $\nabla_y \mathcal{S}(y)$ does not exist. So the SF method has a wider scope than the RP method.

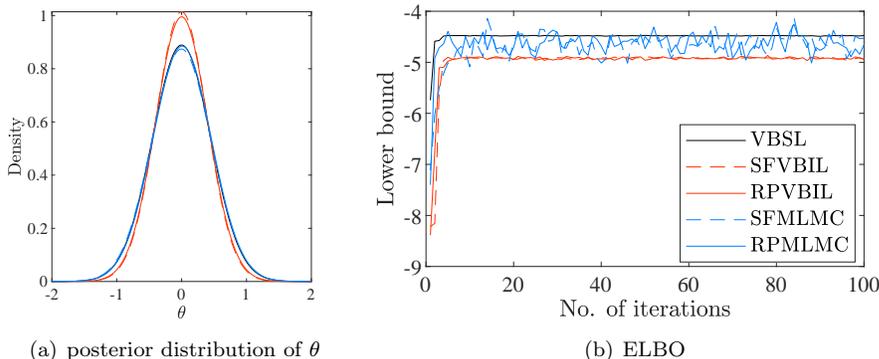
6.1.1. A toy example. To show the unbiasedness of our methods visually, we consider a toy example of ABC which is investigated in [25]. Let the data y_1, \dots, y_n be from a Gaussian distribution with unknown mean θ and unit variance. We assume further the prior of θ is a standard normal distribution $N(0, 1)$. Under this setting, the posterior distribution is tractable actually, which is $\theta|y^* \sim N(n/(1+n)\bar{y}^*, 1/(1+n))$, where \bar{y}^* is the sample mean, but we still approximate the posterior distribution by VB methods for comparisons. Naturally, we take variational distribution $q(\theta)$ to be a normal $N(\mu, \sigma^2)$.

We take the entire data set y^* as the summary statistics (i.e., $\mathcal{S}(y) = y$) to compare the VBIL, VBSL and MLMC methods. The distribution of the summary statistic is normal, and so VBSL renders an unbiased estimator acting as a benchmark. With the Gaussian kernel, the ABC likelihood (6.1) can be calculated analytically actually, which gives a guidance to choose a proper h . The details have been stated in [25]. We take $h = 0.1$ for the kernel function K_h to guarantee the accuracy of the kernel approximation to the true posterior.

We test the SF and RP methods under the MC framework, respectively. In the all simulations, we consider $d = n = 4$ and set the number of outer samples $S = 100$ and the number of inner samples $N = 100$ for all of the methods. We set the learning rate $\rho_t = 1/(5 + t)$. And $\alpha = 1.3$ is taken for the SF methods while $\alpha = 1.1$ is taken for the RP methods. We initialize the starting points for $q(\theta)$ to be $N(\bar{y}^*, 1)$ and $y^* = (0, \dots, 0)$.

Figure 1 illustrates the variational posterior approximations of θ and corresponding ELBOs of VBSL, VBIL and unbiased MLMC method under the SF and RP frameworks respectively. Observed from the left panel of Figure 1, the estimated densities of the MLMC methods and the benchmark method (VBSL) overlap considerably. On the contrary, the VBIL methods yield inaccurate densities and lower ELBOs. The ELBO of unbiased MLMC methods has more volatility than the other methods. A possible explanation is that, although the MLMC method eliminates bias, it may introduce more randomness. Nevertheless, it is apparent that MLMC methods find better variational parameters which benefit from the unbiasedness of the gradient estimators.

FIG. 1. Comparison of VBIL, VBSL and unbiased MLMC.



6.1.2. The g -and- k model. The univariate g -and- k distribution is a flexible unimodal distribution that is able to describe data with significant amounts of skew-

ness and kurtosis [32]. Its density function has no closed form, but is alternatively defined through its quantile function as:

$$Q(q|\theta) = A + B \left[1 + 0.8 \frac{1 - \exp\{-gz(q)\}}{1 + \exp\{-gz(q)\}} \right] (1 + z(q)^2)^k z(q),$$

where $\theta = (A, B, g, k)$, $B > 0$, $k > -1/2$, and $z(q) = \Phi^{-1}(q)$ denotes the inverse CDF of $N(0, 1)$. If $g = k = 0$, it reduces to a normal distribution. As shown in [1], ABC is a good candidate for handling this model.

Suppose that the observations y^* of length $T = 1000$ are independently generated from the g -and- k distribution with parameter $\theta_0 = (3, 1, 2, 0.5)$. We use the unconstrained parameter $\tilde{\theta} = (A, \log B, g, \log(k + 1/2))$ in the VB and take the prior density for $\tilde{\theta}$ as $N(0, 4 \cdot I_4)$. As suggested in [6], we take the summary statistics $\mathcal{S}(y) = (\mathcal{S}_A, \mathcal{S}_B, \mathcal{S}_g, \mathcal{S}_k)$ with

$$\begin{aligned} \mathcal{S}_A &= E_4, \\ \mathcal{S}_B &= E_6 - E_2, \\ \mathcal{S}_g &= (E_6 + E_2 - 2E_4)/S_B, \\ \mathcal{S}_k &= (E_7 - E_5 + E_3 - E_1)/S_B, \end{aligned}$$

where $E_1 \leq E_2 \leq \dots \leq E_7$ are the octiles of y . Note that $\mathcal{S}(y)$ is not differentiable, and thus the RP method cannot be applied. The observed summary statistics $\mathcal{S}(y^*) = (3.05, 1.63, 1.58, 0.42)$.

We compare MLMC and VBIL for a large bandwidth ($h = 5$) and a small bandwidth ($h = 0.5$), and look at the effect of bandwidth. The benchmark is the ABC acceptance-rejection (ABC-AR) samples of size 10^4 . When $h = 5$, the acceptance rate of ABC sampling is about 18%, while $h = 0.5$, the acceptance rate reduces to 1%. We take $\alpha = 1.3$ when $h = 5$ while $\alpha = 1.1$ when $h = 0.5$ for the minor h has effect on the smoothness of the inner function.

Figure 2 shows the variational posterior distributions of VBIL and unbiased MLMC. As we can see, unbiased MLMC-based VB approximates the ABC posterior well, particularly for the marginal distributions of A and g . Again, as shown in Figure 3, unbiased MLMC leads to a larger ELBO.

Using RQMC in MLMC is minor for this example (the results are similar to Figures 2 and 3, and are thus omitted for saving space). The reason is two-fold. First, it is required 1000-dimensional RQMC points in the inner simulation, which is quite large. On the other hand, the summary statistics are functions of sampling quantiles, which are not smooth enough. Due to the high-dimensionality and the absence of smoothness in the integrands, RQMC may not perform well as expected. To overcome this, one may design some dimension reduction techniques for handling the integrand in (6.1).

6.2. Generalized linear mixed models. Generalized linear mixed models (GLMM) use a vector of random effects α_i to account for the dependence between the observations $y_i = \{y_{ij}, j = 1, \dots, n_i\}$ which are measured on the same individual i . The joint likelihood function of the model parameters θ and the random effects $\alpha = (\alpha_1, \dots, \alpha_n)$ is $p(y^*, \alpha|\theta) = \prod_{i=1}^n p(\alpha_i|\theta)p(y_i|\theta, \alpha_i)$ which is tractable. However, the likelihood function $p(y^*|\theta) = \prod_{i=1}^n p(y_i|\theta)$ with

$$p(y_i|\theta) = \int p(y_i|\theta, \alpha_i)p(\alpha_i|\theta)d\alpha_i$$

FIG. 2. Comparison of marginal posterior distributions.

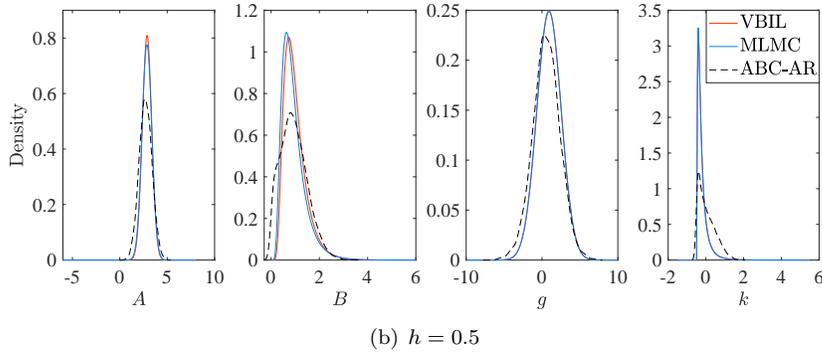
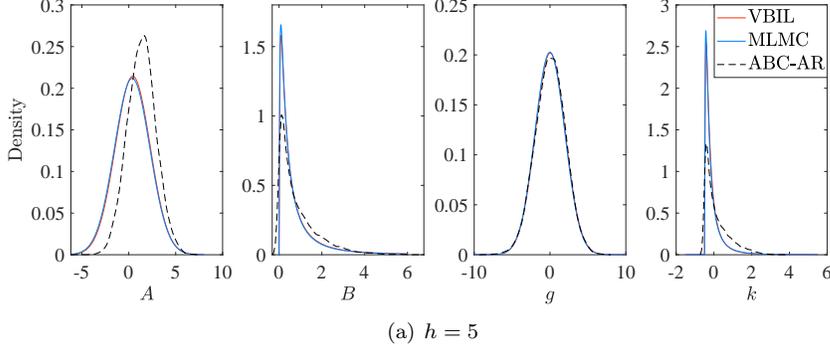
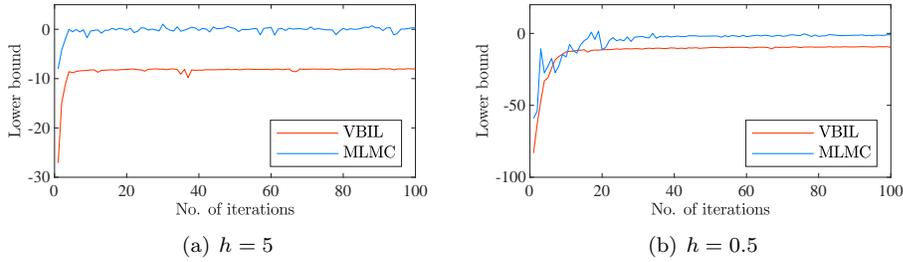


FIG. 3. Comparison of ELBOs.



is analytically intractable in most cases, while it can be easily estimated unbiasedly with importance sampling. Suppose $h_i(\alpha_i|y^*, \theta)$ is an importance density for α_i , then the likelihood $p(y_i|\theta)$ is estimated unbiasedly by

$$\hat{p}_{N_i}(y_i|\theta) = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{p(y_i|\alpha_i^{(j)}, \theta)p(\alpha_i^{(j)}|\theta)}{h_i(\alpha_i^{(j)}|y^*, \theta)},$$

with $\alpha_i^{(j)} \stackrel{iid}{\sim} h_i(\cdot|y^*, \theta)$.

We now compare the VBIL method and the unbiased MLMC methods using the Six City data in [11]. The data consist of binary responses y_{ij} which is the wheezing

status (1 if wheezing, 0 if not wheezing) of the i th child at time-point j , where $i = 1, \dots, 537$ which represent 537 children and $j = 1, 2, 3, 4$ which denote 7, 8, 9, 10 year-old centered at 9 years correspondingly. Covariates are A_{ij} , the age of the i th child at time-point j and S_i the i th maternal smoking status (0 or 1). We consider the logistic regression model with a random intercept $y_{ij}|\beta, \alpha \sim \text{Binomial}(1, p_{ij})$, where $\text{logit}(p_{ij}) = \beta_1 + \beta_2 A_{ij} + \beta_3 S_i + \alpha_i$ with $\alpha_i \sim N(0, \tau^2)$. The parameters of this model are $\theta = (\beta, \tau^2)$. Then the likelihood function is given by

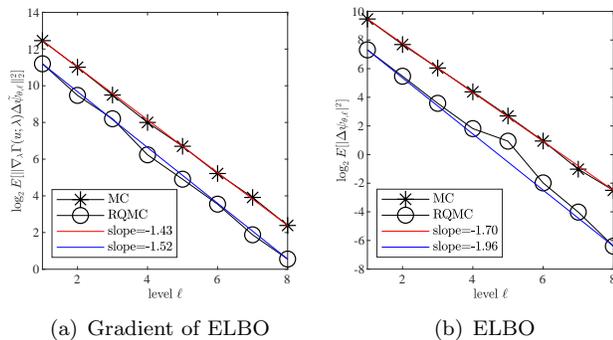
$$p(y^*|\theta) = \prod_{i=1}^{537} \int \prod_{j=1}^4 \frac{\exp\{y_{ij}(\beta_1 + \beta_2 A_{ij} + \beta_3 S_i + \alpha_i)\}}{1 + \exp\{\beta_1 + \beta_2 A_{ij} + \beta_3 S_i + \alpha_i\}} \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\{-\frac{\alpha_i^2}{2\tau^2}\} d\alpha_i.$$

A normal prior $N(0, 50I_3)$ is taken for β with a Gamma(1, 0.1) prior for τ , the square root of τ^2 . We set the variational distribution $q_\lambda(\theta)$ to be a 4-dimensional normal $N(\mu, \Sigma)$, where we let $(\theta_1, \theta_2, \theta_3, \theta_4)$ denote $(\beta_1, \beta_2, \beta_3, \log \tau^2)$, which means the variational distribution of β is a 3-dimensional normal distribution and τ^2 is a log-normal distribution. This example was also investigated in [35]. We focus on the RP method in this example because there is overwhelming empirical evidence in the literature showing the superiority of RP than SF. Some theoretical explanation can be found in [36].

In the RP method, we take $\theta = (\beta, \log \tau^2) = \mu + L\mathbf{u}$, where $\mathbf{u} \sim N(0, I_4)$. In the inner simulation, we take $x_i = (x_{i1}, \dots, x_{i4}) = (z_{i1}, \dots, z_{i4}) + \sqrt{\tau^2} \mathbf{v}_i \cdot \mathbf{1}_4$, where $z_{ij} = \beta_1 + \beta_2 A_{ij} + \beta_3 S_i$, $\mathbf{v}_i \sim N(0, 1)$ and $\mathbf{1}_4$ denotes the vector (1, 1, 1, 1).

Firstly, we test the decreasing rates of $\mathbb{E}[\|\nabla_\lambda \Gamma(\mathbf{u}; \lambda) \Delta \tilde{\psi}_{\theta, \ell}\|_2^2]$ for testing MLMC-based gradient estimation and $\mathbb{E}[\|\Delta \psi_{\theta, \ell}\|^2]$ for testing MLMC-based ELBO estimation. We run the algorithms starting with $\mu = (0, 0, 0, 0)^T$, $\Sigma = I_4$ and $M_0 = 16$. We compare the cases of using MC and RQMC in the inner simulation. To get accurate estimates of these quantities, we use RQMC in the outer sampling. As shown in Figure 4, we find that $r = 1.52$ for the gradient estimator when RQMC is used in the inner simulation while $r = 1.43$ for MC in the inner. Also, RQMC leads to a larger $r = 1.96$ for the ELBO estimator. When MC is used in the inner, we take $\alpha = 1.4$ to finalize the probability distribution of w_ℓ . While $\alpha = 1.5$ when RQMC is used in the inner. A large α speeds up the VB algorithm. According to (3.5), RQMC reduces the cost by a factor of 16% compared to MC.

FIG. 4. Tests of the decrease rates.



The results in Figure 4 show that RQMC can improve the sampling accuracy in the inner simulation with a large r , but the effect of RQMC used in the outer

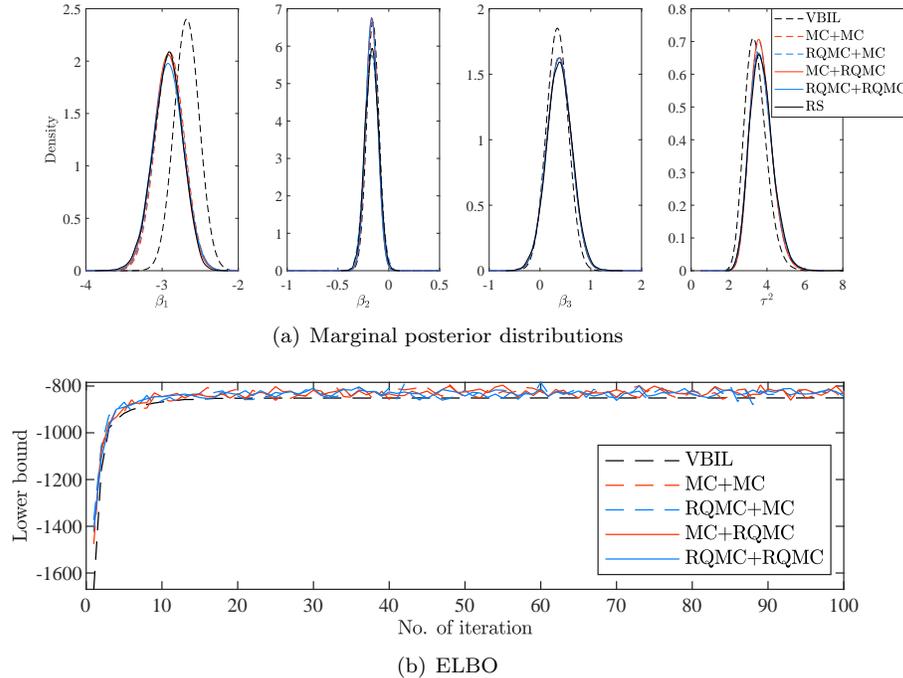
TABLE 1

Variations of unbiased MLMC-based gradient estimators for the initial variational parameters. 'I' is short for 'Inner', 'O' for 'Outer', 'M' for 'MC' and 'Q' for 'RQMC'.

I/O	β_1	β_2	β_3	τ^2	L_{11}	L_{21}	L_{31}	L_{41}	L_{22}	L_{23}	L_{24}	L_{33}	L_{34}	L_{44}
M/M	152	164	30	41	253	182	54	118	226	28	74	55	99	32
M/Q	69	97	11	30	171	142	26	99	215	18	89	29	87	30
Q/M	111	84	17	22	260	148	34	49	162	31	47	27	33	17
Q/Q	82	80	13	40	170	146	30	93	161	24	86	20	30	11

simulation is still unclear. To this end, we estimate the variance of the unbiased MLMC-based gradient estimator for the initial variational parameters by 50 repetitions. The empirical variances are shown in Table 1. It can be seen that using RQMC in either inner or outer simulation reduce the variances for most parameters. Variance reduction of gradient estimates should help to improve VB.

FIG. 5. Comparison of VBIL and four unbiased MLMC methods: MC+MC, MC+RQMC, RQMC+MC and RQMC+RQMC.



Finally, we compare VBIL with four unbiased MLMC methods: MC+MC, MC+RQMC, RQMC+MC and RQMC+RQMC, where for example, MC+RQMC means the MC method is used in the outer while the RQMC method is used in the inner and so on. We take $M_0 = 8$ for the unbiased MLMC methods and $N = 16$ for VBIL. The RStan package 'rstanarm' is used to sample from $p(\theta|y^*)$ as a benchmark, which performs posterior analysis for models with dependent data such as GLMMs. As shown in Figure 5, unbiased MLMC-based methods show great consistency with the benchmark distribution (labeled as RS). On the other hand, all unbiased MLMC methods lead to larger ELBOs than VBIL.

7. Concluding remarks. In this paper, we developed a general method to deal with VB problems with intractable likelihoods. The central point is to find an unbiased gradient estimator in stochastic gradient-based optimization. We achieve this goal by designing unbiased nested MLMC estimators for both the SF and RP gradients. Compared to VBIL, our proposed methods find a better fitting of the posterior distribution and a tighter estimate of the marginal likelihood. Compared to VBSL, our methods work with general distributions of summary statistics. To improve the sampling efficiency, we incorporated RQMC in the inner and the outer simulations. Using RQMC in the inner simulation can reduce the average cost of unbiased MLMC. Using RQMC in the outer simulation can reduce the variance of the gradient estimator. Both aspects speed up the VB algorithm.

REFERENCES

- [1] D. ALLINGHAM, R. A. KING, AND K. L. MENGERSEN, *Bayesian estimation of quantile distributions*, Stat. Comput., 19 (2009), pp. 189–201, <https://doi.org/10.1007/s11222-008-9083-x>.
- [2] S. BARTHELMÉ AND N. CHOPIN, *Expectation propagation for likelihood-free inference*, J. Amer. Statist. Assoc., 109 (2014), pp. 315–333, <https://doi.org/10.1080/01621459.2013.864178>.
- [3] A. BUCHHOLZ AND N. CHOPIN, *Improving approximate Bayesian computation via quasi-Monte Carlo*, J. Comput. Graph. Statist., 28 (2019), pp. 205–219, <https://doi.org/10.1080/10618600.2018.1497511>.
- [4] A. BUCHHOLZ, F. WENZEL, AND S. MANDT, *Quasi-Monte Carlo variational inference*, in International Conference on Machine Learning, 2018, pp. 668–677, <https://arxiv.org/abs/1807.01604>.
- [5] K. BUJOK, B. M. HAMBLY, AND C. REISINGER, *Multilevel simulation of functionals of Bernoulli random variables with application to basket credit derivatives*, Methodol. Comput. Appl. Probab., 17 (2015), pp. 579–604, <https://doi.org/10.1007/s11009-013-9380-5>.
- [6] C. C. DROVANDI AND A. N. PETTITT, *Likelihood-free Bayesian estimation of multivariate quantile distributions*, Comput. Statist. Data Anal., 55 (2011), pp. 2541–2556, <https://doi.org/10.1016/j.csda.2011.03.019>.
- [7] J. DUCHI, E. HAZAN, AND Y. SINGER, *Adaptive subgradient methods for online learning and stochastic optimization.*, J. Mach. Learn. Res., 12 (2011), p. 2121–2159, <https://dl.acm.org/doi/10.5555/1953048.2021068>.
- [8] J. C. DUCHI, *Introductory lectures on stochastic optimization*, The Mathematics of Data, 25 (2018), pp. 99–185.
- [9] J. DURBIN AND S. J. KOOPMAN, *Time Series Analysis by State Space Methods*, Oxford : Oxford University Press, 2nd ed.
- [10] J. DURBIN AND S. J. KOOPMAN, *Monte Carlo maximum likelihood estimation for non-Gaussian state space models*, Biometrika, 84 (1997), pp. 669–684, <http://www.jstor.org/stable/2337587>.
- [11] G. M. FITZMAURICE AND N. M. LAIRD, *A likelihood-based method for analysing longitudinal binary responses*, Biometrika, 80 (1993), pp. 141–151, <https://doi.org/10.1093/biomet/80.1.141>.
- [12] M. B. GILES, *Multilevel Monte Carlo path simulation*, Oper. Res., 56 (2008), pp. 607–617, <https://doi.org/10.1287/opre.1070.0496>.
- [13] M. B. GILES, *Multilevel Monte Carlo methods*, Acta Numer., 24 (2015), pp. 259–328, <https://doi.org/10.1017/S096249291500001X>.
- [14] M. B. GILES, *MLMC for nested expectations*, in Contemporary Computational Mathematics-A Celebration of the 80th Birthday of Ian Sloan, 2018, pp. 425–442.
- [15] M. B. GILES AND T. GODA, *Decision-making under uncertainty: Using MLMC for efficient estimation of EVPPI*, Stat. Comput., 29 (2019), pp. 739–751, <https://doi.org/10.1007/s11222-018-9835-1>.
- [16] M. B. GILES AND L. SZPRUCH, *Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation*, Ann. Appl. Probab., 24 (2014), pp. 1585–1620, <https://doi.org/10.1214/13-AAP957>.
- [17] T. GODA, T. HIRONAKA, AND T. IWAMOTO, *Multilevel Monte Carlo estimation of expected information gains*, Stoch. Anal. Appl., 38 (2020), pp. 581–600, <https://doi.org/10.1080/07362994.2019.1705168>.
- [18] T. GODA, T. HIRONAKA, AND W. KITADE, *Unbiased MLMC stochastic gradient-based optimiza-*

- tion of Bayesian experimental designs*, arXiv preprint arXiv:2005.08414, (2020), <https://arxiv.org/abs/2005.08414>.
- [19] S. HEINRICH, *Monte Carlo complexity of global solution of integral equations*, J. Complexity, 14 (1998), pp. 151–175, <https://doi.org/10.1006/jcom.1998.0471>.
- [20] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014), <https://arxiv.org/abs/1412.6980>.
- [21] D. P. KINGMA AND M. WELLING, *Auto-encoding variational Bayes*, arXiv preprint arXiv:1312.6114, (2013), <https://arxiv.org/abs/1312.6114>.
- [22] P. L'ECUYER AND C. LEMIEUX, *Recent advances in randomized quasi-Monte Carlo methods*, in Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications, M. Dror, P. L'Ecuyer, and F. Szidarovszky, eds., Kluwer Academic Publishers, 2005, pp. 419–474, https://doi.org/10.1007/0-306-48102-2_20.
- [23] S. LIU AND A. B. OWEN, *Quasi-Newton quasi-Monte Carlo for variational Bayes*, arXiv preprint arXiv:2104.02865, (2021), <https://arxiv.org/abs/2104.02865>.
- [24] A. C. MILLER, N. J. FOTI, A. D'AMOUR, AND R. P. ADAMS, *Reducing reparameterization gradient variance*, in Advances in Neural Information Processing Systems, 2017, <https://arxiv.org/abs/1705.07880>.
- [25] V. M. ONG, D. J. NOTT, M.-N. TRAN, S. A. SISSON, AND C. C. DROVANDI, *Variational Bayes with synthetic likelihood*, Stat. Comput., 28 (2018), pp. 971–988, <https://doi.org/10.1007/s11222-017-9773-3>.
- [26] A. B. OWEN, *Randomly permuted (t, m, s)-nets and (t, s)-sequences*, in Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, H. Niederreiter and P. J.-S. Shiue, eds., Springer, 1995, pp. 299–317.
- [27] A. B. OWEN, *Scrambled net variance for integrals of smooth functions*, Ann. Statist., 25 (1997), pp. 1541–1562, <https://doi.org/10.1214/aos/1031594731>.
- [28] A. B. OWEN AND D. RUDOLF, *A strong law of large numbers for scrambled net integration*, SIAM Rev., 63 (2021), pp. 360–372, <https://doi.org/10.1137/20M1320535>.
- [29] J. PAISLEY, D. BLEI, AND M. JORDAN, *Variational Bayesian inference with stochastic search*, in Proceedings of the 29th International Conference on International Conference on Machine Learning, 2012, pp. 1363–1370.
- [30] G. PETERS, S. SISSON, AND Y. FAN, *Likelihood-free Bayesian inference for α -stable models*, Comput. Statist. Data Anal., 56 (2012), pp. 3743–3756, <https://doi.org/10.1016/j.csda.2010.10.004>.
- [31] M. K. PITT, R. S. SILVA, P. GIORDANI, AND R. KOHN, *On some properties of Markov chain Monte Carlo simulation methods based on the particle filter*, J. Econometrics, 171 (2012), pp. 134–151, <https://doi.org/10.1016/j.jeconom.2012.06.004>.
- [32] G. RAYNER AND H. MACGILLIVRAY, *Weighted quantile-based estimation for a class of transformation distributions*, Comput. Statist. Data Anal., 39 (2002), pp. 401–433, [https://doi.org/10.1016/S0167-9473\(01\)00090-1](https://doi.org/10.1016/S0167-9473(01)00090-1).
- [33] C.-H. RHEE AND P. W. GLYNN, *Unbiased estimation with square root convergence for SDE models*, Oper. Res., 63 (2015), pp. 1026–1043, <https://doi.org/10.1287/opre.2015.1404>.
- [34] S. TAVARE, D. J. BALDING, R. C. GRIFFITHS, AND P. DONNELLY, *Inferring coalescence times from DNA sequence data*, Genetics, 145 (1997), pp. 505–518, <https://doi.org/10.1093/genetics/145.2.505>.
- [35] M.-N. TRAN, D. J. NOTT, AND R. KOHN, *Variational Bayes with intractable likelihood*, J. Comput. Graph. Statist., 26 (2017), pp. 873–882, <https://doi.org/10.1080/10618600.2017.1330205>.
- [36] M. XU, M. QUIROZ, R. KOHN, AND S. A. SISSON, *Variance reduction properties of the reparameterization trick*, in The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 2711–2720, <https://arxiv.org/abs/1809.10330>.