On Multilevel Monte Carlo Unbiased Gradient Estimation For Deep Latent Variable Models

Yuyang Shi
University of Oxford

Rob Cornish
University of Oxford

Abstract

Standard variational schemes for training deep latent variable models rely on biased gradient estimates of the target objective. Techniques based on the Evidence Lower Bound (ELBO), and tighter variants obtained via importance sampling, produce biased gradient estimates of the true log-likelihood. The family of Reweighted Wake-Sleep (RWS) methods further relies on a biased estimator of the inference objective, which biases training of the encoder also. In this work, we show how multilevel Monte Carlo (MLMC) can provide a natural framework for debiasing these methods with two different estimators. We prove rigorously that this approach yields unbiased gradient estimators with finite variance under reasonable conditions. Furthermore, we investigate methods that can reduce variance and ensure finite variance in practice. Finally, we show empirically that the proposed unbiased estimators outperform IWAE and other debiasing method on a variety of applications at the same expected cost.

1 INTRODUCTION

Latent Variable Models (LVM) are ubiquitous in machine learning and statistics, but performing inference and learning for such models are typically challenging because exact likelihoods are rarely tractable. LVMs model an observation $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ as the marginal of a joint probability density $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ parameterized by $\boldsymbol{\theta}$, where $\boldsymbol{z} \in \boldsymbol{\mathcal{Z}}$ denotes some latent variables. A standard approach to learn $\boldsymbol{\theta}$ involves maximizing the log marginal likelihood

$$\ell(\boldsymbol{\theta}) \coloneqq \log p_{\boldsymbol{\theta}}(\boldsymbol{x}), \text{ where } p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int_{\boldsymbol{\mathcal{Z}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}, \quad (1)$$

averaged across observations of $\boldsymbol{x}$ from the ground truth model $p_{\boldsymbol{\theta}^*}(\boldsymbol{x})$. However, the integral in (1) is intractable for all but simple models. This is the case for example with deep LVMs such as Variational Autoencoders (VAEs) (Kingma and Welling, 2014), which parameterize $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ using a neural network.

Common techniques for learning $\boldsymbol{\theta}$ in this setting involve the use of biased gradient estimates. A standard approach introduces a variational approximation $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ of the true posterior and optimizes the Evidence Lower Bound (ELBO), or more generally, the Importance Weighted Autoencoder (IWAE) objective (Burda et al., 2016)[1]

$$\ell^{(k)}_{\mathrm{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}_{1:k}}\left[\log\left(\frac{1}{k}\sum_{i=1}^{k} w_i\right)\right] \leq \ell(\boldsymbol{\theta}), \quad (2)$$

where $\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} q_{\boldsymbol{\phi}}(\cdot|\boldsymbol{x})$ and $w_i \coloneqq \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}_i)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_i|\boldsymbol{x})}$. Optimizing $\ell^{(k)}_{\mathrm{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ is performed jointly in $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. However, unless $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ exactly matches $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, then in general $\nabla_{\boldsymbol{\theta}} \ell^{(k)}_{\mathrm{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \neq \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, and hence SGD can optimize $\boldsymbol{\theta}$ far from the true maximizer of $\ell(\boldsymbol{\theta})$ unless k is large, as observed by e.g. Yacoby et al. (2020).

An alternative to IWAE is the method of Reweighted Wake-Sleep (RWS) (Bornschein and Bengio, 2015), which maintains $\ell^{(k)}_{\mathrm{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ as the objective for $\boldsymbol{\theta}$, but for $\boldsymbol{\phi}$ proposes to optimize the forward Kullback-Leibler (KL) divergence $\mathrm{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$. However, $\boldsymbol{\phi}$ gradients of the forward KL are also usually intractable and are approximated using Self-Normalized Importance Sampling (SNIS), which introduces an additional source of bias in the $\boldsymbol{\phi}$ gradients.

In light of these problems, it is useful to consider how to optimize $\ell(\boldsymbol{\theta})$ without relying on biased gradient

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] We assume throughout that $p_\theta(\boldsymbol{z}|\boldsymbol{x}) > 0$ if and only if $q_\phi(\boldsymbol{z}|\boldsymbol{x}) > 0$, so that $w_i > 0$ almost surely.

updates. Recently, Luo et al. (2020) proposed SUMO, which yields unbiased estimates of $\ell(\boldsymbol{\theta})$ via a Russian Roulette mechanism (McLeish, 2011). However, the variance of the estimator this produces is potentially infinite, as noted therein. To address this, in this paper we show how the multilevel Monte Carlo (MLMC) method described by Blanchet et al. (2019) may be used to debias $\ell(\boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$, as well as other quantities useful for the learning of deep LVMs with provable control over the variance under weak assumptions.

Concurrently to us, Goda and Ishikawa (2019) and Ishikawa and Goda (2020) independently proposed one of the log-likelihood estimators $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta})$ for LVMs that we consider here, and established a result similar to Theorem 2 under slightly different assumptions. In this work, we go further by considering theoretical and practical aspects of training with unbiased MLMC estimators. In Section 4, we consider theory for debiasing general SNIS estimators, and establish weak sufficient conditions to admit finite variance in finite expected time that do not require the evaluated function to be bounded. This ensures finite variance for $\nabla_{\boldsymbol{\theta}}\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ under reasonable assumptions, as well as various gradient estimates for the variational distribution $q_{\boldsymbol{\phi}}$, which is key to the training of LVMs. In Section 5, we provide further analysis of the variance of the unbiased estimators, which provides some insight into variance reduction and suitable training methods for $q_{\boldsymbol{\phi}}$. When the required finite variance assumptions are violated, we further relate to thermodynamic integration (Gelman and Meng, 1998) and consider an extension of the method that can still ensure finite variance in Section 6. Finally, we provide novel experimental results that the proposed MLMC estimators outperform IWAE and SUMO on density and image modeling tasks at the same expected cost.

2 BACKGROUND AND RELATED WORK

2.1 Setup and Problem Statement

Consider the joint probability density $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$ defined on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{Z}}$, where the normalizing constant $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is intractable, and define $\pi(\boldsymbol{z}) \coloneqq p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}$ to be the posterior. Given a proposal density $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ whose support is equal to the support of $\pi(\boldsymbol{z})$, we have the following importance sampling identities

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}) = q_{\boldsymbol{\phi}}[w], \;\; \pi[\psi] = \frac{q_{\boldsymbol{\phi}}[w\psi]}{q_{\boldsymbol{\phi}}[w]}, \;\; \text{for } w(\boldsymbol{z}) \coloneqq \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})},$$

where we use the notation $\mu[f] \coloneqq \mathbb{E}_{\mu}[f(\boldsymbol{z})]$ for any density μ and test function f. Using $\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} q_{\boldsymbol{\phi}}(\cdot|\boldsymbol{x})$ for $i = 1, ..., k$, we obtain the following *biased* estimator of the log marginal likelihood $\ell(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$

$$\hat{\ell}^{(k)}(\boldsymbol{\theta}) = \log\left(\frac{1}{k}\sum_{i=1}^{k} w_i\right), \quad w_i \coloneqq w(\boldsymbol{z}_i), \tag{3}$$

and the Self-Normalized Importance Sampling (SNIS) estimator of $\pi[\psi]$ for test function ψ

$$\hat{\pi}^{(k)}[\psi] = \sum_{i=1}^{k} \bar{w}_i \psi(\boldsymbol{z}_i), \quad \bar{w}_i \coloneqq \frac{w(\boldsymbol{z}_i)}{\sum_{j=1}^{k} w(\boldsymbol{z}_j)}. \tag{4}$$

The IWAE objective (Burda et al., 2016) is precisely $\ell^{(k)}_{\text{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{z}_{1:k}}[\hat{\ell}^{(k)}(\boldsymbol{\theta})]$. The estimators $\hat{\ell}^{(k)}(\boldsymbol{\theta})$ and $\hat{\pi}^{(k)}[\psi]$ are both consistent, but *biased* for any finite k. For IWAE, we further have that $\ell^{(k)}_{\text{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ is a lower bound for $\ell(\boldsymbol{\theta})$ (Burda et al., 2016).

2.2 Unbiased Estimators of Approximable Quantities

We now review a general framework for *debiasing* estimators: that is, for producing unbiased estimators of some quantity given only biased estimates. Denote a quantity of interest by I_∞, which may correspond to $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ or $\pi[\psi]$ in our context. Suppose I_∞ can be written as

$$I_\infty = \mathbb{E}[I_0] + \sum_{k=0}^{\infty} \mathbb{E}[\Delta_k]$$

for random variables I_0 and $(\Delta_k)_{k\geq 0}$. We can estimate I_∞ unbiasedly via the following "single sample" (SS) or "Russian Roulette" (RR) estimators (McLeish, 2011; Rhee and Glynn, 2015; Blanchet et al., 2019):

$$\text{SS} = I_0 + \frac{\Delta_K}{p(K)}, \quad \text{RR} = I_0 + \sum_{k=0}^{K} \frac{\Delta_k}{\mathbb{P}\left(K \geq k\right)}, \tag{5}$$

where $K \sim p(\cdot)$ is a non-negative integer sampled independently. Unbiasedness follows directly when K has full support, e.g.

$$\mathbb{E}[\text{SS}] = \mathbb{E}[I_0] + \sum_{k=0}^{\infty} \underbrace{\mathbb{E}\left[\frac{\Delta_K}{p(K)} \mid K = k\right]}_{=\mathbb{E}[\Delta_k]/p(k)} p(k) = I_\infty.$$

However, the variance of SS and RR may be very large or even infinite (McLeish, 2011; Rhee and Glynn, 2015; Beatson and Adams, 2019) and depends crucially on the choice of $p(K)$ as well as Δ_k. In order to guarantee both finite variance and finite expected computation time, we make use of the following result (Rhee and Glynn, 2015; Blanchet et al., 2019) for SS and RR.

Theorem 1. *If the following conditions hold:*

(a) $\mathbb{E}\left[I_0\right] + \sum_{k=0}^{\infty} \mathbb{E}\left[\Delta_k\right] = I_\infty$*;*

(b) I_0 *has finite variance, and there exists* $\alpha, c > 0$ *such that* $\mathbb{E}\left[\Delta_k^2\right] \leq c \cdot 2^{-(1+\alpha)k}$ *for all* k*;*[2]

(c) $\mathbb{E}\left[C_k\right] \leq c' \cdot 2^k$ *for some* $c' > 0$*, where* C_k *is the sampling cost of* Δ_k*;*

then for $K \sim \text{Geom}(r)$ *where* $r \in (\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}})$*,* SS *and* RR *are unbiased estimators*[3] *of* I_∞*, whose variance and expected sampling cost are both finite.*

2.3 SUMO (Luo et al., 2020)

To estimate the log-likelihood $I_\infty = \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, it appears natural to consider $\Delta_k^{\text{SUMO}} := \hat{\ell}^{(k+2)}(\boldsymbol{\theta}) - \hat{\ell}^{(k+1)}(\boldsymbol{\theta})$. The RR estimator with Δ_k^{SUMO} is precisely the SUMO estimator $\hat{\ell}^{\text{SUMO}}(\boldsymbol{\theta})$ (Luo et al., 2020). However, SUMO is not guaranteed to have finite variance in finite expected time, unless

$$\begin{aligned} \mathbb{E}[(\Delta_k^{\text{SUMO}})^2] &= O(k^{-(2+\alpha)}) \\ \textstyle\sum_{j>k} \mathbb{E}[\Delta_j^{\text{SUMO}} \Delta_k^{\text{SUMO}}] &= O(k^{-(2+\alpha)}) \end{aligned} \tag{6}$$

for some $\alpha > 0$. Luo et al. (2020) only establish that $\mathbb{E}[(\Delta_k^{\text{SUMO}})^2] = O(k^{-2})$ (and indeed under the strong assumption that all moments of $w(\boldsymbol{z})$ exist), resulting in potentially unbounded variance as reported therein.

Note that the sampling cost of Δ_k^{SUMO} grows linearly with k, and so falls outside of the context of Theorem 1, which assumes an exponentially increasing sample cost in k. In order to obtain guarantees of finite variance, we might attempt to bring SUMO closer to this setting by considering $\Delta_k^{\text{SUMO}'} := \hat{\ell}^{(2^{k+1})}(\boldsymbol{\theta}) - \hat{\ell}^{(2^k)}(\boldsymbol{\theta})$. However, this construction does not address the limitations of SUMO in general: Figure 1 shows an example in which $\mathbb{E}[(\Delta_k^{\text{SUMO}'})^2] \approx O(2^{-k})$ so that SUMO′ also appears to have infinite variance.

We next show how one can instead apply the multilevel Monte Carlo (MLMC) methodology to obtain unbiased estimators of $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and $\pi[\psi]$, and provide sufficient conditions to verify Theorem 1.

3 UNBIASED MLMC LOG-LIKELIHOOD ESTIMATOR

3.1 Definition

Multilevel Monte Carlo methodology of Blanchet and Glynn (2015); Giles (2015); Blanchet et al. (2019) relies on an alternative clever scheme to construct Δ_k,

[2]Intuitively, since $p(k)$ or $\mathbb{P}(K \geq k)$ is $O\left((1-r)^k\right)$ for $K \sim \text{Geom}(r)$, we need $\mathbb{E}[\Delta_k^2]$ to decay at least as fast to ensure finite variance of the $\frac{\Delta_K}{p(K)}$ and $\sum_{k=0}^K \frac{\Delta_k}{\mathbb{P}(K \geq k)}$ terms.

[3]Precisely, $\mathbb{E}\left[\text{SS}\right] = \mathbb{E}\left[\text{RR}\right] = I_\infty$, where the expectations are taken over K, I_0, and $(\Delta_k)_{k \geq 1}$.

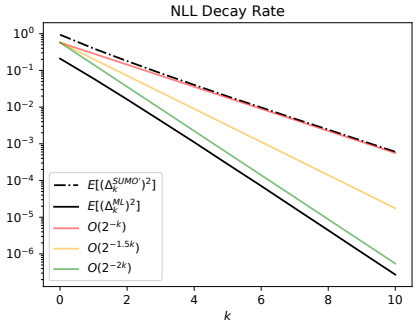


Figure 1: On the linear Gaussian example from Rainforth et al. (2018), while the importance weights $w(\boldsymbol{z})$ are suitably well-behaved, $\mathbb{E}[(\Delta_k^{\text{SUMO}'})^2]$ only converges at approximately $O(2^{-k})$, which results in infinite variance. On the other hand, $\mathbb{E}[(\Delta_k^{\text{ML}})^2] \approx O(2^{-2k})$, and Δ_k^{ML} also has smaller variance than $\Delta_k^{\text{SUMO}'}$.

which can ensure the construction of unbiased estimators of $\ell(\boldsymbol{\theta})$ that admit finite variance and can be computed in finite expected time.

We denote by $\boldsymbol{z}_i^O, \boldsymbol{z}_i^E$ two independent sequences of i.i.d. samples of $q_{\boldsymbol{\phi}}$, where O, E denotes odd, even respectively, and w_i^O, w_i^E the corresponding importance weights. We then let $I_0 = \hat{\ell}^{(1)}(\boldsymbol{\theta})$ and

$$\Delta_k^{\text{ML}} = \hat{\ell}_{O \cup E}^{(2^{k+1})}(\boldsymbol{\theta}) - \frac{1}{2}\left(\hat{\ell}_O^{(2^k)}(\boldsymbol{\theta}) + \hat{\ell}_E^{(2^k)}(\boldsymbol{\theta})\right), \tag{7}$$

where $\hat{\ell}_O^{(2^k)}(\boldsymbol{\theta})$ is computed as in (3) using the odd samples $\{\boldsymbol{z}_i^O\}_{i=1}^{2^k}$, $\hat{\ell}_E^{(2^k)}(\boldsymbol{\theta})$ using the even samples $\{\boldsymbol{z}_i^E\}_{i=1}^{2^k}$, and $\hat{\ell}_{O \cup E}^{(2^{k+1})}(\boldsymbol{\theta})$ using $\{\boldsymbol{z}_i^O\}_{i=1}^{2^k} \cup \{\boldsymbol{z}_i^E\}_{i=1}^{2^k}$. We denote the corresponding multilevel SS/RR estimators of $\ell(\boldsymbol{\theta})$ as $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta})$ and $\hat{\ell}^{\text{ML-RR}}(\boldsymbol{\theta})$, and collectively as $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$.

3.1.1 Implementation Details

Several design choices are available in terms of practical implementation. For I_0, it is possible to choose any I_0 such that $\mathbb{E}[I_0] = \mathbb{E}[\hat{\ell}^{(1)}(\boldsymbol{\theta})]$. Therefore, we can compute an average of $\hat{\ell}^{(1)}(\boldsymbol{\theta})$ using all available samples $\{\boldsymbol{z}_i^O\}_{i=1}^{2^k} \cup \{\boldsymbol{z}_i^E\}_{i=1}^{2^k}$.

It is also possible to start at a higher level $l \geq 1$, i.e. taking $I_0 = \hat{\ell}^{(2^l)}(\boldsymbol{\theta})$ and $\Delta_k^{\text{ML}} = \hat{\ell}_{O \cup E}^{(2^{k+l+1})}(\boldsymbol{\theta}) - \frac{1}{2}\left(\hat{\ell}_O^{(2^{k+l})}(\boldsymbol{\theta}) + \hat{\ell}_E^{(2^{k+l})}(\boldsymbol{\theta})\right)$. Using more samples in each Δ_k reduces the variance of the estimator, at the cost of an increased computational cost. It can be checked that the expected sampling cost is thus $\sum_{k=0}^\infty 2^{k+l+1} p(k) = \frac{r \cdot 2^{l+1}}{1-2(1-r)}$.

Algorithm 1 Unbiased multilevel estimator $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta})$ of $\ell(\boldsymbol{\theta})$.

Input: Proposal $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, unnormalized target $p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})$, probability mass function $p(k)$.

Sample $K \sim p(k)$.

Sample $\{\boldsymbol{z}_i\}_{i=1}^{2^{K+1}} = \{\boldsymbol{z}_i^O\}_{i=1}^{2^K} \cup \{\boldsymbol{z}_i^E\}_{i=1}^{2^K} \overset{\text{i.i.d.}}{\sim} q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, compute $\log w_i \leftarrow \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}_i)}{q_{\boldsymbol{\phi}}(\boldsymbol{z}_i|\boldsymbol{x})}$.

Compute $I_0 = \text{mean}\left(\{\log w_i\}_{i=1}^{2^{K+1}}\right)$.

Compute $\hat{\ell}_O^{(2^K)}(\boldsymbol{\theta}) \leftarrow \text{logmeanexp}\left(\{\log w_i^O\}_{i=1}^{2^K}\right)$, $\hat{\ell}_E^{(2^K)}(\boldsymbol{\theta}) \leftarrow \text{logmeanexp}\left(\{\log w_i^E\}_{i=1}^{2^K}\right)$.

Compute $\hat{\ell}_{O\cup E}^{(2^{K+1})}(\boldsymbol{\theta}) \leftarrow \text{logmeanexp}\left(\{\log w_i\}_{i=1}^{2^{K+1}}\right) = \text{logmeanexp}\left(\hat{\ell}_O^{(2^K)}(\boldsymbol{\theta}), \hat{\ell}_E^{(2^K)}(\boldsymbol{\theta})\right)$.

Compute $\Delta_K = \hat{\ell}_{O\cup E}^{(2^{K+1})}(\boldsymbol{\theta}) - \frac{1}{2}\left(\hat{\ell}_O^{(2^K)}(\boldsymbol{\theta}) + \hat{\ell}_E^{(2^K)}(\boldsymbol{\theta})\right)$.

return $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta}) = I_0 + \frac{\Delta_K}{p(K)}$.

3.2 Theoretical Guarantees

Under the following weak assumptions on the moments of the importance weights, Theorem 1 is satisfied by $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$, i.e. $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ is an unbiased estimator of $\ell(\boldsymbol{\theta})$, admits finite variance and can be computed in finite expected time:

Theorem 2. *Assume there exists $\epsilon, \delta > 0$ such that $\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[w(\boldsymbol{z})^{2+\epsilon} + w(\boldsymbol{z})^{-\delta}\right] < \infty$. Then $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ satisfies Theorem 1 for $r \in \left(\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}}\right)$, where $\alpha = \min\left(\frac{\epsilon}{2}, 1\right) > 0$.*

We note that the finite $2+\epsilon$ moment condition on $w(\boldsymbol{z})$ are weaker than Blanchet et al. (2019), and are assumed in general by other existing works such as Domke and Sheldon (2018).

Intuitively, the key to this construction is that $\mathbb{E}[\Delta_k^{\text{ML}}] = \mathbb{E}[\Delta_k^{\text{SUMO}'}]$, but when performing a Taylor expansion the first order error in Δ_k^{ML} are canceled. Specifically, if we define $O_k = \frac{1}{2^k}\sum_{i=1}^{2^k} w_i^O$, and Taylor expand $\hat{\ell}_O^{(2^k)}(\boldsymbol{\theta}) = \log O_k$ at $\mathbb{E}[O_k] = p_{\boldsymbol{\theta}}(\boldsymbol{x})$, we have $\hat{\ell}_O^{(2^k)}(\boldsymbol{\theta}) = \log p_\theta(\boldsymbol{x}) + \frac{O_k - p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} + O(\frac{O_k - p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})})^2$ Similar expressions hold for $\hat{\ell}_E^{(2^k)}(\boldsymbol{\theta}) = \log E_k$ and $\hat{\ell}_{O\cup E}^{(2^{k+1})}(\boldsymbol{\theta}) = \log \frac{O_k + E_k}{2}$. Therefore, the 1st order error in Δ_k^{ML} equals to

$$\frac{\frac{1}{2}(O_k + E_k) - p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} - \frac{1}{2}\left(\frac{O_k - p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})} + \frac{E_k - p_{\boldsymbol{\theta}}(\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{x})}\right) = 0. \quad (8)$$

This leads to $\mathbb{E}[(\Delta_k^{\text{ML}})^2]$ converging at a faster rate $O(2^{-(1+\alpha)k})$, which satisfies Theorem 1 and guarantees finite variance for $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$.

4 UNBIASED MLMC ESTIMATOR OF EXPECTATIONS

4.1 Definition

Similarly, in order to estimate $I_\infty = \pi[\psi]$ unbiasedly, we can consider

$$\Delta_k^{\text{ML}} = \hat{\pi}_{O\cup E}^{(2^{k+1})}[\psi] - \frac{1}{2}\left(\hat{\pi}_O^{(2^k)}[\psi] + \hat{\pi}_E^{(2^k)}[\psi]\right), \quad (9)$$

where $\hat{\pi}_O^{(2^k)}[\psi], \hat{\pi}_E^{(2^k)}[\psi], \hat{\pi}_{O\cup E}^{(2^{k+1})}[\psi]$ are similarly computed using odd, even, and all samples of $\boldsymbol{z}$ using equation (4). The corresponding SS/RR estimators of $\pi[\psi]$ are denoted $\hat{\pi}^{\text{ML-SS}}[\psi]$ and $\hat{\pi}^{\text{ML-RR}}[\psi]$, and collectively as $\hat{\pi}^{\text{ML}}[\psi]$.

In particular, if $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ exactly, it can be shown easily that $\Delta_k^{\text{ML}} \equiv 0$, but such property does not hold for SUMO. This shows that when we have an unbiased quantity I_0 to start with, $\hat{\pi}^{\text{ML}}[\psi]$ does not introduce extra variance into the estimator.

4.2 Unbiased Estimation of $\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$

We now apply the introduced method to the problem of learning LVMs. We are here interested in estimating unbiasedly the gradient $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$. Fisher's identity states that

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}) \mathrm{d}\boldsymbol{z}. \quad (10)$$

This identity shows that estimating $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$ corresponds to estimating $\pi[\psi]$ for $\psi(\boldsymbol{z}) \coloneqq \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})$. For IWAE, $\pi[\psi]$ is estimated using biased SNIS estimates

$$\hat{\pi}^{(k)}[\psi] = \nabla_{\boldsymbol{\theta}}\hat{\ell}^{(k)}(\boldsymbol{\theta}) = \sum_{i=1}^{k} \bar{w}_i \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}_i). \quad (11)$$

Using MLMC, unbiased estimators $\hat{\pi}^{\text{ML}}[\psi]$ can instead be constructed with SS/RR. Since $\hat{\pi}^{(k)}[\psi] = \nabla_{\boldsymbol{\theta}}\hat{\ell}^{(k)}(\boldsymbol{\theta})$,

similarly $\hat{\pi}^{\text{ML}}[\psi] = \nabla_{\boldsymbol{\theta}} \hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$. This is not surprising as $\mathbb{E}[\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})] = \ell(\boldsymbol{\theta})$, so under regularity conditions allowing to swap differentiation and expectation, $\nabla_{\boldsymbol{\theta}} \hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ is indeed an unbiased estimator of $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$.

4.2.1 Implementation Details

Since $\hat{\pi}^{\text{ML}}[\psi] = \nabla_{\boldsymbol{\theta}} \hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$, it suffices to directly differentiate the unbiased log-likelihood estimator $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$. Compared to IWAE with the same average sampling cost, $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ uses a stochastic number of particles, but only performs elementary operations on $\log w$ in the final layer of the computational graph. Therefore, differentiating $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ has roughly the same average cost as differentiating IWAE.

4.3 Theoretical Guarantees

We now establish theoretical guarantees of finite variance for the gradient estimator $\hat{\pi}^{\text{ML}}[\psi]$, which are key for the unbiased training of LVMs. By Luo et al. (2020, Appendix A5), SUMO also has unbounded variance in this setting under the strong assumption that $w(\boldsymbol{z})$ is bounded. However, we show that $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ instead admits finite variance and can be computed in finite expected time under the following conditions:

Theorem 3. *Assume there exists $a > 4$, $b \geq \frac{2a}{a-4}$ such that $\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[w(\boldsymbol{z})^a + |\psi(\boldsymbol{z})|^b\right] < \infty$. Then $\hat{\pi}^{\text{ML}}[\psi]$ satisfies Theorem 1 for $r \in \left(\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}}\right)$, where $\alpha = 1 - \frac{2}{b} > 0$.*

Alternatively, if $2 < a \leq 4$, $b > \frac{2a+4}{a-2}$, or $a > 4$, $\frac{2a+4}{a-2} < b \leq \frac{2a}{a-4}$, then $\hat{\pi}^{\text{ML}}[\psi]$ satisfies Theorem 1 for $r \in \left(\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}}\right)$, where $\alpha = \frac{a}{2} - \frac{a+2}{b} - 1 > 0$.

Observe that if ψ is bounded, then the condition on b holds trivially and the theorem applies whenever $a > 2$, resulting in $\alpha = \min(\frac{a}{2} - 1, 1)$, just as in Theorem 2. This result for bounded ψ was also obtained by Hironaka et al. (2020), as well as Blanchet et al. (2019) under the stronger assumption that $a > 3$. However, the assumption of a bounded ψ is in general unrealistic in our context, and so Theorem 3 is of interest insofar as it applies also to unbounded ψ, unlike these other results. However, note that, for unbounded ψ, the condition on $w(\boldsymbol{z})$ is in general stronger than in Theorem 2 and depends on the specific function $\psi(\boldsymbol{z})$.

4.4 Other Applications of $\hat{\pi}^{\text{ML}}[\psi]$

To minimize the forward KL, the Reweighted Wake-Sleep (RWS) method (Bornschein and Bengio, 2015; Le et al., 2019) makes use of the following identity

$$\nabla_{\boldsymbol{\phi}} \text{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) || q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})) = -\int p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \mathrm{d}\boldsymbol{z}, \quad (12)$$

which corresponds to $\pi[\psi]$ with $\psi(\boldsymbol{z}) \coloneqq -\nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. RWS uses biased SNIS estimates $\hat{\pi}^{(k)}[\psi] = \nabla_{\boldsymbol{\phi}} \hat{\ell}^{(k)}(\boldsymbol{\theta}) = -\sum_{i=1}^k \bar{w}_i \nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_i|\boldsymbol{x})$. However, unlike the ELBO or IWAE, which can be viewed as lower bounds of the true log-likelihood, this bias is not readily interpretable.

For reparametrizable $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$,[4] Finke and Thiery (2019) show that $\nabla_{\boldsymbol{\phi}} \text{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) || q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$ can alternatively be written as

$$\nabla_{\boldsymbol{\phi}} \text{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) || q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})) = -\int p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) \frac{\partial \log w(\boldsymbol{z})}{\partial \boldsymbol{z}} \frac{\partial h_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}}|_{\boldsymbol{\epsilon} = h_{\boldsymbol{\phi}}^{-1}(\boldsymbol{z})} \mathrm{d}\boldsymbol{z}. \quad (13)$$

which also corresponds to $\pi[\psi]$ where $\psi(\boldsymbol{z}) \coloneqq -\frac{\partial \log w(\boldsymbol{z})}{\partial \boldsymbol{z}} \frac{\partial h_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})}{\partial \boldsymbol{\phi}}|_{\boldsymbol{\epsilon} = h_{\boldsymbol{\phi}}^{-1}(\boldsymbol{z})}$. Similar to RWS, the RHS can be approximated using an SNIS estimator $\hat{\pi}^{(k)}[\psi]$, which coincides with the "sticking the landing" IWAE (IWAE-STL) gradient estimator (Roeder et al., 2017).

To avoid using biased gradient estimates, MLMC can be used to construct the unbiased estimator $\hat{\pi}^{\text{ML}}[\psi]$ in both cases, which admits finite variance under Theorem 3. Comparing the two methods, the reparameterized unbiased estimator has the advantage of achieving zero variance when $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ as $\psi \equiv 0$.

5 VARIANCE ANALYSIS

In general, the unbiased estimators suffer from larger variance than their biased counterparts, which can hinder training in practice. In this section, we further investigate the variance of the multilevel estimators and propose variance reduction techniques.

5.1 Comparing Variance of SS and RR

The SS and RR estimates require the same number of samples when each Δ_k in RR is computed using the same samples, and this number is approximately doubled when one uses different samples. We illustrate here that both estimators present advantages in different scenarios. For simplicity, we consider the first term I_0 and the telescoping part $\tilde{\text{SS}} = \frac{\Delta_K}{p(K)}$, $\tilde{\text{RR}} = \sum_{k=0}^K \frac{\Delta_k}{\mathbb{P}(K \geq k)}$ separately.

Theorem 4. *The variance of $\tilde{\text{SS}}$ and $\tilde{\text{RR}}$ can be decomposed into a sum of two parts, where ② measures the inherent variance when estimating a deterministic series $\sum_{k=0}^\infty \mathbb{E}[\Delta_k]$, and ① measures the extra variance caused by the variance and covariance of Δ_k.*

[4] Precisely, if there exists invertible $h_{\boldsymbol{\phi}}$ and a distribution $q(\boldsymbol{\epsilon})$ such that $h_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}) \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ when $\boldsymbol{\epsilon} \sim q(\boldsymbol{\epsilon})$.

For the $\tilde{\text{SS}}$ *estimator,*

$$\begin{aligned}
① &= \sum_{k=0}^{\infty} \frac{\operatorname{Var} \Delta_k}{p\left(k\right)}, \\
② &= \sum_{k=0}^{\infty} \left(\frac{1}{p\left(k\right)} - 1 \right) \mathbb{E}\left[\Delta_k\right]^2 - 2 \sum_{i,j:i<j} \mathbb{E}\left[\Delta_i\right] \mathbb{E}\left[\Delta_j\right].
\end{aligned} \tag{14}$$

For the $\tilde{\text{RR}}$ *estimator,*

$$\begin{aligned}
① &= \sum_{k=0}^{\infty} \frac{\operatorname{Var} \Delta_k}{\mathbb{P}\left(K \geq k\right)} + 2 \sum_{i,j:i<j} \frac{\operatorname{Cov}(\Delta_i, \Delta_j)}{\mathbb{P}\left(K \geq i\right)}, \\
② &= \sum_{k=0}^{\infty} \left(\frac{1}{\mathbb{P}\left(K \geq k\right)} - 1 \right) \mathbb{E}\left[\Delta_k\right]^2 \\
&\quad + 2 \sum_{i,j:i<j} \left(\frac{1}{\mathbb{P}\left(K \geq i\right)} - 1 \right) \mathbb{E}\left[\Delta_i\right] \mathbb{E}\left[\Delta_j\right].
\end{aligned} \tag{15}$$

We notice that the term ① for $\tilde{\text{RR}}$ is clearly smaller than $\tilde{\text{SS}}$ in the case Δ_k are pairwise independent, and can achieve almost half the magnitude as $\tilde{\text{SS}}$ by choosing $K \sim \text{Geom}\left(r\right)$ with $r \to \frac{1}{2}$. However, the term ② for $\tilde{\text{RR}}$ is potentially larger. In particular, we observe the $\tilde{\text{RR}}$ estimator can be advantageous if Δ_k have high variances but mean close to 0.

For the two considered debiasing problems, the bias of $\hat{\ell}^{(k)}(\boldsymbol{\theta})$ and $\hat{\pi}^{(k)}[\psi]$ decay at asymptotic rate $O(k^{-1})$ under the conditions of Theorem 2 and 3; see e.g. Domke and Sheldon (2018); Nowozin (2018). This shows that $\mathbb{E}[\Delta_k]^2 = O(2^{-2k})$ and $\sum_{j=k+1}^{\infty} \mathbb{E}[\Delta_k]\mathbb{E}[\Delta_j] = O(2^{-2k})$. Using the multilevel construction, we show in the proofs that $\mathbb{E}[\Delta_k^2] = O(2^{-(1+\alpha)k})$, where $\alpha \leq 1$ is defined as in Theorem 2 or 3. Therefore, in the case $\alpha < 1$, asymptotically the term ① dominates the term ②, so it suffices to compare the term ① between the two estimators. In the case that Δ_k are almost pairwise independent, i.e. $\mathbb{E}[\Delta_i\Delta_j] \approx \mathbb{E}[\Delta_i]\mathbb{E}[\Delta_j]$, the RR estimator should be chosen asymptotically (if we start with $I_0 = \hat{\ell}^{(2^l)}(\boldsymbol{\theta})$ or $\hat{\pi}^{(2^l)}[\psi]$ for l large enough). On the other hand, if the covariance terms have large magnitude, e.g. in the worst case $\mathbb{E}[\Delta_i\Delta_j] \approx \mathbb{E}[\Delta_i^2]^{1/2}\mathbb{E}[\Delta_j^2]^{1/2}$, the SS estimator should be chosen asymptotically.

5.2 Reducing Work-Variance Product by Optimal Distribution of K

For SS, Blanchet et al. (2019) give that the optimal choice of K to minimize the work-variance product should be geometrically distributed with $r = 1 - 2^{-(1+\alpha/2)}$. In the case $\alpha = 1$, $r \approx 0.6464$, resulting in an average sampling complexity of approximately 4.41 samples per evaluation. For RR, it can be similarly shown that the optimal $r \approx 0.6340$ for $\alpha = 1$.

5.3 Training Objective for $q_{\boldsymbol{\phi}}$

For unbiased estimators $\hat{\ell}^{\text{SUMO}}(\boldsymbol{\theta})$ and $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$, the variational distribution $q_{\boldsymbol{\phi}}$ does not affect their unbiasedness under reasonable conditions that $w(\boldsymbol{z})$ has finite moment, so $\hat{\ell}^{\text{SUMO}}(\boldsymbol{\theta})$ and $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$ also provide no signal for learning $q_{\boldsymbol{\phi}}$ (i.e. $\nabla_{\boldsymbol{\phi}}\mathbb{E}[\hat{\ell}^{\text{SUMO}}(\boldsymbol{\theta})] = \nabla_{\boldsymbol{\phi}}\mathbb{E}[\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})] = 0$). Instead, the variational distribution $q_{\boldsymbol{\phi}}$ is useful in our context through the variance of $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$.

Goda and Ishikawa (2019); Ishikawa and Goda (2020) propose to learn $\boldsymbol{\phi}$ by maximizing the ELBO. i.e. minimizing $\text{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}))$. However, it is well-known that the reverse KL is mode-seeking, which typically leads to $q_{\boldsymbol{\phi}}$ admitting thinner tails than $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ with the usual Gaussian parameterization of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. Thus, the weights $w(\boldsymbol{z})$ are more prone to have infinite higher moments, violating the conditions of Theorem 2 and 3. Instead, maximizing $\ell_{\text{IWAE}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ would provide a better training objective, as it can be interpreted as a proxy to minimize the variance of $w(\boldsymbol{z})$ (Domke and Sheldon, 2018). Minimizing the forward $\text{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$, which is mean-seeking, also typically leads to distributions $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ having thicker tails than $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$. This can be interpreted as an unbiased version of RWS.

In Luo et al. (2020), it is proposed to train $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ by directly minimizing the variance of the SUMO estimator with $\nabla_{\boldsymbol{\phi}} \operatorname{Var}(\hat{\ell}^{\text{SUMO}}(\boldsymbol{\theta})) = \nabla_{\boldsymbol{\phi}}\mathbb{E}[\hat{\ell}^{\text{SUMO}}(\boldsymbol{\theta})^2]$. However, the theoretical foundation behind this approach is somewhat unclear, since the variance of SUMO is possibly infinite. A similar strategy can be applied to the multilevel estimators $\hat{\ell}^{\text{ML}}(\boldsymbol{\theta})$, where the estimators now have finite variance.

5.4 Trading off Bias and Variance

While having an unbiased objective is desirable, when a large $K = k$ is sampled from $p(k)$ (with a small probability), the inverse weight $p(k)^{-1}$ or $\mathbb{P}(K \geq k)^{-1}$ in SS or RR can be very large in magnitude. To reduce variance at the cost of introducing an interpretable bias, we can simply restrict the support of the distribution of K to $\{0, \ldots, k_{\max}\}$, thus ensuring $K \leq k_{\max}$ almost surely and providing upper bounds on $p(k)^{-1}$ or $\mathbb{P}(K \geq k)^{-1}$. In this case, SS and RR for Δ_k given by (7) will be unbiased estimators of the IWAE given in (2) using $2^{k_{\max}+1}$ importance samples but computed, on average, using much less than $2^{k_{\max}+1}$ samples.

6 USING THERMODYNAMIC INTEGRATION

When $\mathcal{Z}$ is high-dimensional, simple IS estimators will typically perform poorly, so the importance weights

$w(\boldsymbol{z})$ can violate the assumptions in Theorem 2 and 3, and the variance of the IS-based unbiased estimators can become infinite.

A popular way to mitigate high-dimensionality utilizes Thermodynamic Integration (TI) (Gelman and Meng, 1998), which transforms the estimation of $\ell(\boldsymbol{\theta})$ into an one-dimensional integral

$$\ell(\boldsymbol{\theta}) = \int_0^1 \pi_\beta[\log w]\mathrm{d}\beta, \tag{16}$$

where $\pi_\beta(\boldsymbol{z})$ is the normalized density of $\tilde{\pi}_\beta(\boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})^\beta q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})^{1-\beta}$. We can thus easily obtain an unbiased estimate of $\ell(\boldsymbol{\theta})$ by sampling $\beta \sim \mathrm{Unif}[0,1]$, and then debias the SNIS estimate of $\pi_\beta[\psi]$ given by

$$\hat{\pi}_\beta^{(k)}[\psi] = \sum_{i=1}^k \overline{w_i^\beta}\psi(\boldsymbol{z}_i), \quad \overline{w_i^\beta} \coloneqq \frac{w_i^\beta}{\sum_{j=1}^k w_j^\beta}. \tag{17}$$

We call $\hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})$ the resulting estimator. However, although seemingly attractive, this method is actually not of practical interest as the following result shows:

Proposition 5. *The following identity holds*

$$\hat{\ell}^{(k)}(\boldsymbol{\theta}) = \int_0^1 \hat{\pi}_\beta^{(k)}[\log w]\mathrm{d}\beta. \tag{18}$$

It follows that $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta}) = \int_0^1 \hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})\mathrm{d}\beta$, *i.e.* $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$ *is a Rao-Blackwellized version of* $\hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})$ *and thus* $\mathrm{Var}(\hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})) \geq \mathrm{Var}(\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta}))$.

However, thermodynamic integration can still be useful in high-dimensional scenarios. By leveraging (16), it is proposed by Masrani et al. (2019) for $0 = \beta_0 < \beta_1 < \cdots < \beta_{T-1} < \beta_T = 1$ the TVO objective

$$\ell_{\mathrm{TVO}}(\boldsymbol{\theta}) \coloneqq \sum_{t=0}^{T-1}(\beta_{t+1} - \beta_t)\pi_{\beta_t}[\log w] \leq \ell(\boldsymbol{\theta}). \tag{19}$$

This "thermodynamic" evidence lower bound is tighter than the standard ELBO and can be seen as a left Riemann sum approximation of the integral, and empirically Masrani et al. (2019) show that the values of β_t should be chosen after a point of maximum curvature to obtain tight approximations.

A limitation of this attractive approach is that one cannot estimate expectations in the form $\pi_\beta[\psi]$ unbiasedly, so Masrani et al. (2019) rely on SNIS approximations therein[5]. Plugging (17) into (19) and reusing the w_i samples for all β_t, the resulting log-likelihood estimator $\hat{\ell}_{\mathrm{TVO}}^{(k)}(\boldsymbol{\theta}) = \sum_{t=0}^{T-1}(\beta_{t+1} - \beta_t)\hat{\pi}_{\beta_t}^{(k)}[\log w]$ has roughly the same computation cost as IWAE. However, although seemingly attractive, $\hat{\ell}_{\mathrm{TVO}}^{(k)}(\boldsymbol{\theta})$ is a biased estimator of $\ell_{\mathrm{TVO}}(\boldsymbol{\theta})$, and we show that the SNIS approximation actually biases $\hat{\ell}_{\mathrm{TVO}}^{(k)}(\boldsymbol{\theta})$ to be *less tight* than IWAE:

Proposition 6. $\mathbb{E}_{\boldsymbol{z}_{1:k}}[\hat{\ell}_{\mathrm{TVO}}^{(k)}(\boldsymbol{\theta})] \leq \mathbb{E}_{\boldsymbol{z}_{1:k}}[\hat{\ell}^{(k)}(\boldsymbol{\theta})]$, *regardless of the placement of all* β_t, *i.e.* $\hat{\ell}_{\mathrm{TVO}}^{(k)}(\boldsymbol{\theta})$ *is less tight than* IWAE *due to the SNIS bias.*

Using MLMC, we can instead unbiasedly estimate $\ell_{\mathrm{TVO}}(\boldsymbol{\theta})$ without the SNIS bias by debiasing (17). The resulting estimator

$$\hat{\ell}_{\mathrm{TVO}}^{\mathrm{ML}}(\boldsymbol{\theta}) \coloneqq \sum_{t=0}^{T-1}(\beta_{t+1} - \beta_t)\hat{\pi}_{\beta_t}^{\mathrm{ML}}[\log w] \tag{20}$$

is thus an unbiased estimator of $\ell_{\mathrm{TVO}}(\boldsymbol{\theta})$. We can then obtain an unbiased estimate of $\nabla_{\boldsymbol{\theta}}\ell_{\mathrm{TVO}}(\boldsymbol{\theta})$ by directly differentiating $\hat{\ell}_{\mathrm{TVO}}^{\mathrm{ML}}(\boldsymbol{\theta})$. Moreover, since the importance weights $w_i^{\beta_t}$ in $\hat{\pi}_{\beta_t}^{\mathrm{ML}}[\log w]$ are more well-behaved than w_i, the finite variance condition on w for $\hat{\ell}_{\mathrm{TVO}}^{\mathrm{ML}}(\boldsymbol{\theta})$ becomes more relaxed than $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$, by an application of Theorem 3:

Theorem 7. *Assume there exists* $\epsilon, \delta > 0$ *such that* $\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[w(\boldsymbol{z})^{\beta_{T-1}(2+\epsilon)} + w(\boldsymbol{z})^{-\delta}\right] < \infty$. *Then* $\hat{\ell}_{\mathrm{TVO}}^{\mathrm{ML}}(\boldsymbol{\theta})$ *satisfies Theorem 1 for* $r \in \left(\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}}\right)$, *where* $\alpha = \min\left(\frac{\epsilon}{2}, 1\right)$.

7 EMPIRICAL RESULTS

7.1 Linear Gaussian Experiment

We first consider the linear Gaussian example from Rainforth et al. (2018); Tucker et al. (2019), where we can analytically calculate the true log-likelihood to quantify the bias and variance of all estimators. The generative model is given by $p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{I})\mathcal{N}(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{I})$, where both $\boldsymbol{x},\boldsymbol{z} \in \mathbb{R}^{20}$, so that $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta},2\boldsymbol{I})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\frac{\boldsymbol{\theta}+\boldsymbol{x}}{2}, \frac{1}{2}\boldsymbol{I}\right)$. The encoder distribution is $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{b}, \frac{2}{3}\boldsymbol{I}\right)$, where $\boldsymbol{\phi} = (\boldsymbol{A},\boldsymbol{b})$. Following Rainforth et al. (2018), we consider random perturbations of the parameters near the optimal value by a zero-mean Gaussian with standard deviation 0.01.

We evaluate the performance of ML-SS and ML-RR with $r = 0.6$, and compare against IWAE and SUMO. Figure 2 displays the empirical bias and variance of $\ell(\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$ as the expected computational cost increases from 6 to 384 terms for all methods (computed by averaging of 1000 realizations of the estimators for 10 random perturbations). As expected, the empirical biases of ML-SS and ML-RR are much lower than

[5] In more detail, Masrani et al. (2019) directly considers the covariance gradient estimator, which verifies for $\boldsymbol{\lambda} = \{\boldsymbol{\theta},\boldsymbol{\phi}\}$

$$\nabla_{\boldsymbol{\lambda}}\pi_\beta[\log w] = \pi_\beta[\nabla_{\boldsymbol{\lambda}}\log w] + \mathrm{Cov}_{\pi_\beta}[\nabla_{\boldsymbol{\lambda}}\log\tilde{\pi}_\beta, \log w].$$

SNIS is applied for the RHS in Masrani et al. (2019), but *for the* $\boldsymbol{\theta}$ *gradients* this is identical to directly differentiating $\hat{\pi}_\beta^{(k)}[\log w]$. See Appendix for details.

IWAE at the cost of an increased variance. SUMO also has reduced bias compared to IWAE when the overall computation cost is low, but its empirical variance does not decay monotonically as the computation increases, which is strongly suggestive that SUMO admits infinite variance. This results in the empirical bias of SUMO actually underperforming IWAE at high computational cost. A study of different estimators of $\nabla_{\boldsymbol{\phi}}\mathrm{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$ is also provided in the Appendix.

7.2 2D Density Modeling Experiment

We next consider the 2D density dataset and neural network model (3 hidden layers each with 50 leaky ReLU units) considered by Yacoby et al. (2020). We investigate the effect of training with biased and unbiased $p_{\boldsymbol{\theta}}$ objectives and different $q_{\boldsymbol{\phi}}$ objectives as described in Section 5.3. We perform three experiments fixing the expected computational cost k to be 5, 10, 20 with $r = 0.625$ and choose a learning rate $2.5 \cdot 10^{-4}$ with the Adam optimizer without gradient clipping (so as not to introduce any bias). For the $q_{\boldsymbol{\phi}}$ objective, we find that the variance and unbiased RWS-STL objectives obtained the best results, and both objectives are able to reuse the samples in the $p_{\boldsymbol{\theta}}$ objective. Interestingly, the variance objective for $q_{\boldsymbol{\phi}}$ achieves the best results especially as k increases. The results are presented in Table 1. Compared to IWAE, the multilevel unbiased estimators are able to attain higher test log-likelihoods at the same expected sampling cost. A visualization of the density the models produce is also provided in Figure 3. Further results on the effect of using different $q_{\boldsymbol{\phi}}$ objectives on the multilevel estimators can be found in the Appendix.

7.3 Image Modeling Experiment

We now compare the performance on a standard VAE example as in Burda et al. (2016); Luo et al. (2020). We use the same network architecture and the dynamically binarized MNIST (LeCun et al., 2010), OMNIGLOT (Lake et al., 2015) and Fashion MNIST (Xiao et al., 2017) benchmark datasets following previous works. We follow the training scheme by Luo et al. (2020) closely, which makes use of gradient clipping for excessively large gradients. We also modify the tail of K so that ML-RR is an unbiased estimator of IWAE with $k = 128$ to limit the memory usage and reduce variance, similar to SUMO which softly truncates the tail of K after $k = 80$ terms. Unlike Luo et al. (2020), however, in order to accurately compare different estimators under the same budget, we fix the number of training epochs to 3280 for all estimators, and set IWAE as the training objective for $q_{\boldsymbol{\phi}}$, so we can test the effect of debiasing the $p_{\boldsymbol{\theta}}$ objective using different estimators.

For this task, we observe in Table 2 while SUMO achieves slightly better performance than IWAE on two datasets, ML-RR outperforms SUMO at the same computational budget. Nevertheless, the advantage of debiased estimators (SUMO and ML-RR) seems to be diminishing with increasing k as the variational bound tightens for IWAE, especially for the MNIST dataset.

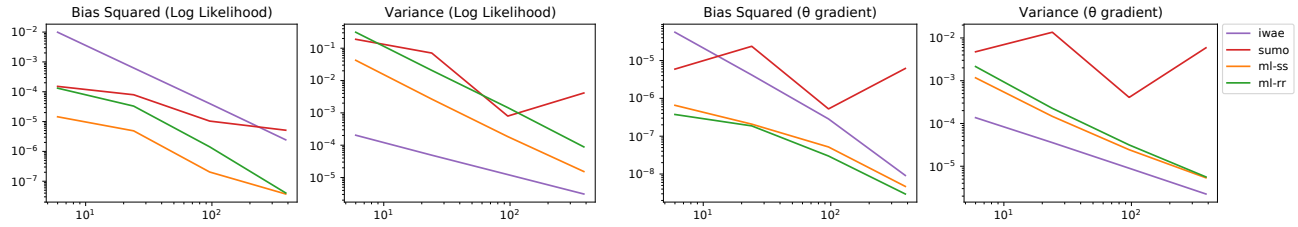
In addition, while we observe TVO underperforms IWAE in this example in line with Masrani et al. (2019), we find that models trained with ML-TVO-RR obtain better test log-likelihood compared to IWAE. As discussed in Section 6, ML-TVO relaxes the finite variance conditions on $w(\boldsymbol{z})$ while introducing a bias relative to ML-RR as a tradeoff due to the discrete integral approximation. However, $\ell_{\mathrm{TVO}}(\boldsymbol{\theta})$ is still guaranteed to lower bound the log-likelihood, and we observe from the experiment that $\ell_{\mathrm{TVO}}(\boldsymbol{\theta})$ (with the unbiased estimator ML-TVO) can be a more preferable family of evidence lower bounds than IWAE.

8 CONCLUSION

We have shown how the multilevel Monte Carlo methodology of Blanchet et al. (2019) can be used to produce various unbiased estimators with finite variance useful for evaluating and training LVMs, and demonstrated their advantages on several models involving deep neural networks. As interesting future work, it may be fruitful to consider variance reduction techniques such as control variates, as well as alternative unbiased estimates of the log-evidence and its gradients that do not rely on importance sampling. Promising progress has been made recently by Ruiz et al. (2020) in this direction, which may provide another possibility for unbiased, finite variance training of VAEs.

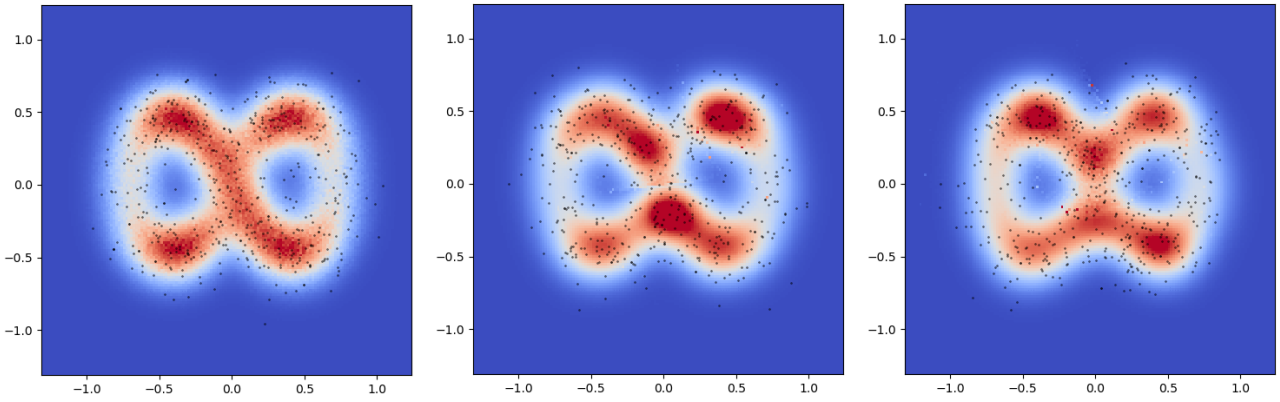
9 ACKNOWLEDGEMENTS

The authors thank Tom Rainforth for his helpful comments and suggestions. Rob Cornish is supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme Grant EP/R018561/1.



(a) (b)

Figure 2: Empirical bias squared and variance of estimators of (a) $\ell(\boldsymbol{\theta})$ and (b) $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$, both plotted against expected computational cost.



(a) (b) (c)

Figure 3: Marginal density plot of (a) the ground truth model; (b) trained model using IWAE with $k = 5$; (c) trained model using ML-SS with $k = 5$.

Table 1: Test negative log-likelihood of the trained models for the 2D dataset over 3 runs, estimated using IWAE5000. All results that are statistically insignificant from the best result are highlighted in bold.

$p_{\boldsymbol{\theta}}$ objective	IWAE	ML-SS		ML-RR	
$q_{\boldsymbol{\phi}}$ objective	IWAE	Var ML-SS	Unbiased RWS-STL	Var ML-RR	Unbiased RWS-STL
$k = 5$	0.9325±0.0054	**0.9228±0.0076**	**0.9216±0.0041**	**0.9222±0.0071**	**0.9202±0.0048**
$k = 10$	0.9248±0.0032	**0.9191±0.0015**	**0.9206±0.0026**	**0.9183±0.0024**	**0.9193±0.0015**
$k = 20$	0.9198±0.0005	**0.9153±0.0016**	0.9194±0.0011	**0.9158±0.0005**	0.9216±0.0011

Table 2: Test negative log-likelihood of the trained models for the image datasets over 3 runs, estimated using IWAE5000. All results that are statistically insignificant from the best result are highlighted in bold.

	MNIST		OMNIGLOT		Fashion MNIST	
$p_{\boldsymbol{\theta}}$ objective	$k = 5$	$k = 15$	$k = 5$	$k = 15$	$k = 5$	$k = 15$
IWAE (Our impl.)	85.11±0.02	**84.62±0.05**	105.40±0.07	104.53±0.10	230.13±0.14	229.77±0.04
SUMO (Our impl.)	85.24±0.05	84.78±0.08	105.19±0.03	104.51±0.06	229.97±0.05	229.61±0.08
ML-RR (Ours)	**85.06±0.07**	**84.64±0.03**	**104.98±0.02**	**104.26±0.05**	**229.78±0.13**	**229.36±0.13**
ML-TVO-RR (Ours)	**85.04±0.01**	**84.59±0.03**	**104.95±0.06**	**104.22±0.11**	**229.78±0.05**	**229.40±0.12**

References

Alex Beatson and Ryan P Adams. Efficient optimization of loops and limits with randomized telescoping sums. In *International Conference on Machine Learning*, volume 97, pages 534–543, 2019.

J.H. Blanchet and P.W. Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multilevel randomization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

Jose H Blanchet, Peter W Glynn, and Yanan Pei. Unbiased multilevel Monte Carlo: Stochastic optimization, steady-state simulation, quantiles, and other applications. *arXiv preprint arXiv:1904.09929*, 2019.

Jörg Bornschein and Yoshua Bengio. Reweighted wake-sleep. In *3rd International Conference on Learning Representations*, 2015.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations*, 2016.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems 31*, pages 4470–4479. 2018.

Axel Finke and Alexandre H Thiery. On the relationship between variational inference and adaptive importance sampling. *arXiv preprint arXiv:1907.10477*, 2019.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

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

Takashi Goda and Kei Ishikawa. Multilevel Monte Carlo estimation of log marginal likelihood. *arXiv preprint arXiv:1912.10636*, 2019.

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, Jan 2020.

Kei Ishikawa and Takashi Goda. Efficient debiased variational Bayes by multilevel Monte Carlo methods. *arXiv preprint arXiv:2001.04676*, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *2nd International Conference on Learning Representations*, 2014.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Tuan Anh Le, Adam R. Kosiorek, N. Siddharth, Yee Whye Teh, and Frank Wood. Revisiting reweighted wake-sleep for models with stochastic control flow. In *Uncertainty in Artificial Intelligence*, 2019.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

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

Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. In *Advances in Neural Information Processing Systems*, pages 11525–11534, 2019.

Don McLeish. A general method for debiasing a Monte Carlo estimator. *Monte Carlo Methods and Applications*, 17(4):301–315, 2011.

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

Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4274–4282, 2018.

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

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pages 6925–6934, 2017.

Francisco J. R. 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.

George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J. Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *International Conference on Learning Representations*, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Yaniv Yacoby, Weiwei Pan, and Finale Doshi-Velez. Failure modes of variational autoencoders and their effects on downstream tasks. *arXiv preprint arXiv:2007.07124*, 2020.