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

1 Preliminary Lemmas

In this section we first establish some preliminary results which we use to prove Theorems 2, 3 and 7.

Lemma 1. *The following inequalities hold.*

(i) $\forall 0 < p \leq 1, \exists c_p > 0$ *such that* $\forall x \geq 0, \log(1+x) \leq c_p x^p$.

(ii) $\forall p \geq 1, \forall j, k > 0, \exists c_{p,j,k} > 0$ *such that* $\forall x > 0, |\log x|^p \leq c_{p,j,k}\left(x^j + \frac{1}{x^k}\right)$.

Proof. (i) For $p = 1$ we have the standard inequality $\log(1+x) \leq x$. For $0 < p < 1$, we have $\lim_{x\to\infty} \frac{\log(1+x)}{x^p} = 0$ and $\lim_{x\to 0^+} \frac{\log(1+x)}{x^p} = \lim_{x\to 0^+} \frac{1}{1+x}\frac{1}{px^{p-1}} = 0$. By continuity, $\frac{\log(1+x)}{x^p}$ attains its maximum and is bounded above on $(0,\infty)$.

(ii) Similarly, this is due to $|\log x|^p$ increasing slower than x^j as $x \to \infty$ and slower than x^{-k} as $x \to 0^+$. More rigorously, we have by standard results $0 \leq \lim_{x\to\infty} \frac{(\log x)^p}{x^j + x^{-k}} \leq \lim_{x\to\infty} \frac{(\log x)^p}{x^j} = 0$ and $0 \leq \lim_{x\to 0^+} \frac{(-\log x)^p}{x^j + x^{-k}} \leq \lim_{x\to 0^+} \frac{(-\log x)^p}{x^{-k}} = 0$. Hence by continuity $\frac{|\log x|^p}{x^j + x^{-k}}$ attains its maximum and is bounded above on $(0,\infty)$. □

Lemma 2. *Let $X_1, \ldots, X_n$ be i.i.d. zero mean random variables, whose p-th moments are finite for some $p \geq 2$. Then there exists a constant C_p, independent of n and the distribution of X_i, such that*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right|^p\right] \leq C_p n^{-p/2}\mathbb{E}\left[|X_i|^p\right] \tag{1}$$

Proof. This follows from a direct application of Marcienkiewicz-Zygmund inequality then Jensen inequality. □

Lemma 3. *Let $X_1, \ldots, X_n$ be positive i.i.d. random variables, then $\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i\right)^{-p}\right] \leq \mathbb{E}\left[X_1^{-p}\right]$ for all $p \geq 0$.*

Proof. By Jensen's inequality,

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-p} \leq \frac{1}{n}\sum_{i=1}^{n} X_i^{-p}.$$

Taking expectation over both sides yields the result. □

Lemma 4. *Suppose $\{X_k\}_{k=1}^\infty$, $\{Y_k\}_{k=1}^\infty$ are two sequences of random variables satisfying $\mathbb{E}\left[|X_k|^p\right] = O\left(2^{-ck}\right)$, $\mathbb{E}\left[|Y_k|^p\right] = O\left(2^{-ck}\right)$, where $p \geq 1$, then $\mathbb{E}\left[|X_k + Y_k|^p\right] = O\left(2^{-ck}\right)$.*

Proof. By the Cp-inequality, we have for $p \geq 1$

$$\begin{aligned}\mathbb{E}\left[|X_k + Y_k|^p\right] &\leq 2^{p-1}\cdot\left(\mathbb{E}\left[|X_k|^p\right] + \mathbb{E}\left[|Y_k|^p\right]\right).\\ &= O\left(2^{-ck}\right)\end{aligned}$$

□

2 Proof of Theorem 2

We now give a proof of Theorem 2. We note that concurrently with our work, Goda and Ishikawa (2019); Ishikawa and Goda (2020) independently proposed the estimator $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta})$ and a related theoretical result under slightly different assumptions (requiring finite moments for $\log w(\boldsymbol{z})$ when $\boldsymbol{z} \sim q_{\boldsymbol{\phi}}$).

Theorem. *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 $\alpha = \min\left(\frac{\epsilon}{2}, 1\right) > 0$.*

Proof. To prove this Theorem, we are going to check that the assumptions of Theorem 1 hold. We first start proving Assumption (a) and Assumption (c) and then finally Assumption (b) whose proof is more involved.

Assumption (a): We first show that $\hat{\ell}^{(k)}(\boldsymbol{\theta})$ is uniformly bounded in L^2. Observe that by Lemma 1, $\mathbb{E}\left[|\hat{\ell}^{(k)}(\boldsymbol{\theta})|^2\right] = \mathbb{E}\left[\left|\log\left(\frac{1}{k}\sum_{i=1}^k w_i\right)\right|^2\right] \leq c \cdot \mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^k w_i\right) + \left(\frac{1}{k}\sum_{i=1}^k w_i\right)^{-\delta}\right]$. But now $\mathbb{E}[|w_i|] < \infty$, and $\mathbb{E}\left[\left(\frac{1}{k}\sum_{i=1}^k w_i\right)^{-\delta}\right]$ is bounded for all k by Lemma 3. It follows that $\sup_k \mathbb{E}\left[|\hat{\ell}^{(k)}(\boldsymbol{\theta})|^2\right] < \infty$. Therefore

$$
\begin{aligned}
\mathbb{E}\left[I_0\right] + \sum_{k=0}^{\infty} \mathbb{E}\left[\Delta_k\right] &= \mathbb{E}\left[I_0\right] + \sum_{k=0}^{\infty} \mathbb{E}\left[\log\left(\frac{O_k + E_k}{2}\right) - \frac{1}{2}\left(\log O_k + \log E_k\right)\right] \\
&= \mathbb{E}[I_0] + \sum_{k=0}^{\infty} \mathbb{E}\left[\log\left(\frac{O_k + E_k}{2}\right)\right] - \frac{1}{2}\left(\mathbb{E}[\log O_k] + \mathbb{E}[\log E_k]\right) \\
&= \mathbb{E}\left[\hat{\ell}^{(1)}(\boldsymbol{\theta})\right] + \sum_{k=0}^{\infty}\left(\mathbb{E}\left[\hat{\ell}^{(2^{k+1})}(\boldsymbol{\theta})\right] - \mathbb{E}\left[\hat{\ell}^{(2^k)}(\boldsymbol{\theta})\right]\right) \\
&= \lim_{k\to\infty} \mathbb{E}\left[\hat{\ell}^{(2^{k+1})}(\boldsymbol{\theta})\right] \\
&= \mathbb{E}\left[\lim_{k\to\infty} \hat{\ell}^{(2^{k+1})}(\boldsymbol{\theta})\right] \\
&= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}).
\end{aligned}
$$

Here the integrability requirements of the second equality follow from boundedness of $\hat{\ell}^{(k)}(\boldsymbol{\theta})$ in L^2. The second last equality holds due to uniform integrability, which likewise follows from boundedness in L^2. Finally, the last equality holds because $\hat{\ell}^{(k)}(\boldsymbol{\theta}) = \log\left(\frac{1}{k}\sum_{i=1}^k w_i\right) \to \log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ almost surely by the SLLN (which is applicable due to our moment conditions, which entail $\mathbb{E}[|w_i|] < \infty$) and the continuous mapping theorem. This shows that Assumption (a) in Theorem 1 is satisfied.

Assumption (c): This trivially holds.

Assumption (b): We divide our analysis into two cases where $O_k E_k$ is large and $O_k E_k$ is small, specifically into $O_k E_k \geq 2^{-2\kappa} Z^2$ and $O_k E_k < 2^{-2\kappa} Z^2$, where $\mathbb{E}\left[O_k\right] = \mathbb{E}\left[E_k\right] = Z = p_{\boldsymbol{\theta}}(\boldsymbol{x})$, and $\kappa > 0$ is a constant, independent of k, taken such that $2^{-\kappa} Z < 1$.

1. Case $O_k E_k$ large, i.e. $O_k E_k \geq 2^{-2\kappa} Z^2$:

We have $\Delta_k = \log\left(\frac{O_k+E_k}{2}\right) - \frac{1}{2}\left(\log O_k + \log E_k\right) = \log\left(\frac{O_k+E_k}{2\sqrt{O_k E_k}}\right) = \log\left(1 + \frac{(O_k - E_k)^2}{2\sqrt{O_k E_k}\left(\sqrt{O_k}+\sqrt{E_k}\right)^2}\right)$. By Lemma 1, $\forall 0 < p \leq 1, \exists c_p > 0$ such that

$$
\Delta_k \leq c_p \left(\frac{(O_k - E_k)^2}{2\sqrt{O_k E_k}\left(\sqrt{O_k}+\sqrt{E_k}\right)^2}\right)^p \qquad \text{and} \qquad \Delta_k^2 \leq c_p^2 \left(\frac{(O_k - E_k)^4}{4 O_k E_k \left(\sqrt{O_k}+\sqrt{E_k}\right)^4}\right)^p. \tag{2}
$$

The denominator in (2) is clearly bounded from below in this case, since $\left(\sqrt{O_k}+\sqrt{E_k}\right)^4 \geq O_k E_k \geq 2^{-2\kappa} Z^2$.

For the numerator, we can take $p = \min\left(\frac{2+\epsilon}{4}, 1\right)$. Hence by Lemma 4,

$$
\mathbb{E}\left[|O_k - E_k|^{4p}\right] = O\left(2^{-2pk}\right),
$$

since $\mathbb{E}\left[|O_k - Z|^{4p}\right], \mathbb{E}\left[|E_k - Z|^{4p}\right]$ are both $O\left(2^{-2pk}\right)$ by Lemma 2. Hence $\mathbb{E}\left[\Delta_k^2 \mathbb{I}\left(O_k E_k \geq 2^{-2\kappa} Z^2\right)\right] = O\left(2^{-(1+\alpha)k}\right)$, where $\alpha = \min\left(\frac{\epsilon}{2}, 1\right)$.

2. Case $O_k E_k$ small, i.e. $O_k E_k < 2^{-2\kappa} Z^2$: In this case, we must have either $O_k < 2^{-\kappa} Z < 1$ or $E_k < 2^{-\kappa} Z < 1$. Suppose that $O_k < 2^{-\kappa} Z < 1$, then since $\log(1+x) \leq \max(1 + \log x, 1)$, we have

$$\begin{aligned}
\Delta_k &= \log\left(1 + \frac{\left(\sqrt{O_k} - \sqrt{E_k}\right)^2}{2\sqrt{O_k E_k}}\right) \\
&\leq \max\left(1 + \log\left(\frac{\left(\sqrt{O_k} - \sqrt{E_k}\right)^2}{2\sqrt{O_k E_k}}\right), 1\right) \\
&= \max\left(1 + 2\log\left|\sqrt{O_k} - \sqrt{E_k}\right| - \frac{1}{2}\left(\log O_k + \log E_k\right) - \log 2, 1\right).
\end{aligned}$$

For $2\log\left|\sqrt{O_k} - \sqrt{E_k}\right|$, we know that if $O_k \leq E_k$, then $2\log\left|\sqrt{O_k} - \sqrt{E_k}\right| \leq \log E_k \leq \log E_k \mathbb{I}(E_k \geq 1)$. If $O_k > E_k$, then $2\log\left|\sqrt{O_k} - \sqrt{E_k}\right| \leq \log O_k < 0 \leq \log E_k \mathbb{I}(E_k \geq 1)$. Hence we have

$$\Delta_k \leq \max\left(1 + \log E_k \mathbb{I}(E_k \geq 1) - \frac{1}{2}\left(\log O_k + \log E_k\right), 1\right).$$

We also have $-\frac{1}{2}\log E_k \leq -\frac{1}{2}\log E_k \mathbb{I}(E_k < 1)$, and $-\frac{1}{2}\log O_k > 0$. Overall, we thus have

$$\begin{aligned}
\Delta_k &\leq \max\left(1 + \log E_k \mathbb{I}(E_k \geq 1) - \frac{1}{2}\log E_k \mathbb{I}(E_k < 1) - \frac{1}{2}\log O_k, 1\right) \\
&= 1 + \log E_k \mathbb{I}(E_k \geq 1) - \frac{1}{2}\log E_k \mathbb{I}(E_k < 1) - \frac{1}{2}\log O_k.
\end{aligned}$$

Let $Y_k = 1 + \log E_k \mathbb{I}(E_k \geq 1) - \frac{1}{2}\log E_k \mathbb{I}(E_k < 1)$. Then we have

$$\Delta_k^2 \leq Y_k^2 - Y_k \log O_k + \frac{1}{4}\left(\log O_k\right)^2,$$

$$Y_k^2 = 1 + \left(\left(\log E_k\right)^2 + 2\log E_k\right)\mathbb{I}(E_k \geq 1) + \left(\frac{1}{4}\left(\log E_k\right)^2 - \log E_k\right)\mathbb{I}(E_k < 1).$$

By independence of O_k and E_k, we have

$$\mathbb{E}\left[\Delta_k^2 \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right] \leq \mathbb{E}\left[Y_k^2\right] \cdot \mathbb{P}\left(O_k < 2^{-\kappa} Z\right) - \mathbb{E}[Y_k]\,\mathbb{E}\left[\log O_k \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right] + \frac{1}{4}\mathbb{E}\left[\left(\log O_k\right)^2 \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right].$$

We first prove that $\mathbb{E}[Y_k]$ and $\mathbb{E}\left[Y_k^2\right]$ are $O(1)$. By Hölder's inequality, we have for any event A $\mathbb{E}\left[|\log E_k|^p \mathbb{I}(A)\right] \leq \mathbb{E}\left[|\log E_k|^{2p}\right]^{\frac{1}{2}}$. However, $\forall p \geq 1, \exists c_p > 0$ s.t. $|\log x|^p \leq c_p\left(x^{-\delta} + x\right)$ by Lemma 1. By Lemma 3, $\mathbb{E}\left[E_k^{-\delta}\right]$ is bounded above. Therefore, $\mathbb{E}[Y_k]$ and $\mathbb{E}\left[Y_k^2\right]$ are both $O(1)$.

Next, we prove that $\mathbb{P}\left(O_k < 2^{-\kappa} Z\right)$ and $\mathbb{E}\left[|\log O_k|^p \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right]$ are both arbitrarily close to the convergence speed $O(2^{-(1+\alpha)k})$. By Hölder's inequality, $\mathbb{E}\left[|\log O_k|^p \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right] \leq \mathbb{E}\left[|\log O_k|^{pq}\right]^{\frac{1}{q}} \mathbb{P}\left(O_k < 2^{-\kappa} Z\right)^{\left(1-\frac{1}{q}\right)}$. We have shown above that $\mathbb{E}\left[|\log O_k|^{pq}\right]^{\frac{1}{q}}$ is $O(1)$ for fixed p, q, and $\mathbb{P}\left(O_k < 2^{-\kappa} Z\right) \leq \mathbb{P}\left(|O_k - Z| > \left(1 - 2^{-\kappa}\right) Z\right) \leq \frac{\mathbb{E}\left[|O_k - Z|^{2+\epsilon}\right]}{\left(\left(1-2^{-\kappa}\right) Z\right)^{2+\epsilon}} = O\left(2^{-(1+\epsilon/2)k}\right)$, by Markov's inequality and Lemma 2. Take q to be sufficiently large to obtain $\mathbb{P}\left(O_k < 2^{-\kappa} Z\right)^{\left(1-\frac{1}{q}\right)} = O\left(2^{-(1+\alpha)k}\right)$, where α can be arbitrarily close to $\frac{\epsilon}{2}$.

Hence $\mathbb{P}\left(O_k < 2^{-\kappa} Z\right)$ and $\mathbb{E}\left[|\log O_k|^p \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right]$ are both $O\left(2^{-(1+\alpha)k}\right)$, where α can be arbitrarily close to $\frac{\epsilon}{2}$. Therefore, $\mathbb{E}\left[\Delta_k^2 \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right] = O\left(2^{-(1+\alpha)k}\right)$, and by symmetry the case is similar for $E_k < 2^{-\kappa} Z$.

Combining all the results, since $\mathbb{E}\left[\Delta_k^2\right] \leq \mathbb{E}\left[\Delta_k^2 \mathbb{I}\left(O_k E_k \geq 2^{-2\kappa} Z^2\right)\right] + \mathbb{E}\left[\Delta_k^2 \mathbb{I}\left(O_k < 2^{-\kappa} Z\right)\right] + \mathbb{E}\left[\Delta_k^2 \mathbb{I}\left(E_k < 2^{-\kappa} Z\right)\right]$, we obtain $\mathbb{E}\left[\Delta_k^2\right] = O\left(2^{-(1+\alpha)k}\right)$, where α can be taken arbitrarily close to $\min\left(\frac{\epsilon}{2}, 1\right)$. Therefore, all values of r strictly in $\left(\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}}\right)$ admit finite variance for $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$ by Theorem 1, where $\alpha = \min\left(\frac{\epsilon}{2}, 1\right)$. □

3 Proof of Theorem 3

We first provide here a general result for the SNIS estimator

$$\hat{\pi}^{(k)}[\psi] = \sum_{i=1}^{k} \bar{w}_i \psi(\boldsymbol{z}_i), \qquad \bar{w}_i = \frac{w_i}{\sum_{j=1}^{k} w_j}$$

introduced in Section 2.1. We will use the following notation for the standard Monte Carlo estimator of the expectation of a function ψ with respect to a measure μ, $\hat{\mu}^{(k)}_{\mathrm{MC}}[\psi] := \frac{1}{k}\sum_{i=1}^{k} \psi(\boldsymbol{z}_i)$ for $\boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} \mu$.

Lemma 5. *Assume that the SNIS estimator $\hat{\pi}^{(k)}[\psi]$, with unnormalized target density $\gamma(\boldsymbol{z})$, normalizing constant Z, normalized density $\pi(\boldsymbol{z})$, proposal density $q(\boldsymbol{z})$ and importance weight $w(\boldsymbol{z}) = \frac{\gamma(\boldsymbol{z})}{q(\boldsymbol{z})}$, satisfies $\mathbb{E}_q[w(\boldsymbol{z})^{2p}] < \infty$ for some $p \geq 1$. Then for any bounded test function ψ, $\mathbb{E}\left[\left|\hat{\pi}^{(k)}[\psi] - \pi[\psi]\right|^{2p}\right] = O(k^{-p})$.*

Proof. We note that

$$\pi[\psi] = \frac{q[w\psi]}{q[w]} = \frac{q[w\psi]}{Z}, \qquad \hat{\pi}^{(k)}[\psi] = \frac{\hat{q}^{(k)}_{\mathrm{MC}}[w\psi]}{\hat{q}^{(k)}_{\mathrm{MC}}[w]}.$$

We have the following decomposition

$$\begin{aligned}
\hat{\pi}^{(k)}[\psi] - \pi[\psi] &= \hat{\pi}^{(k)}[\psi] - \frac{1}{Z}\hat{q}^{(k)}_{\mathrm{MC}}[w\psi] + \frac{1}{Z}\hat{q}^{(k)}_{\mathrm{MC}}[w\psi] - \pi[\psi] \\
&= \frac{1}{Z}\left[\left(q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right)\hat{\pi}^{(k)}[\psi] + \left(\hat{q}^{(k)}_{\mathrm{MC}}[w\psi] - q[w\psi]\right)\right]
\end{aligned}$$

By the Cp-inequality, we have for $p \geq 1/2$

$$\begin{aligned}
\left|\hat{\pi}^{(k)}[\psi] - \pi[\psi]\right|^{2p} &\leq \frac{1}{Z^{2p}}\left(\left|\left(q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right)\hat{\pi}^{(k)}[\psi]\right| + \left|\hat{q}^{(k)}_{\mathrm{MC}}[w\psi] - q[w\psi]\right|\right)^{2p} \\
&\leq \frac{2^{2p-1}}{Z^{2p}}\left(\left|\left(q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right)\hat{\pi}^{(k)}[\psi]\right|^{2p} + \left|\hat{q}^{(k)}_{\mathrm{MC}}[w\psi] - q[w\psi]\right|^{2p}\right). \qquad (3)
\end{aligned}$$

For $|\psi| \leq B$,

$$\left|\hat{\pi}^{(k)}[\psi] - \pi[\psi]\right|^{2p} \leq \frac{2^{2p-1}}{Z^{2p}}\left(B^{2p}\left|q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right|^{2p} + \left|\hat{q}^{(k)}_{\mathrm{MC}}[w\psi] - q[w\psi]\right|^{2p}\right).$$

Hence we can conclude by by Lemma 2. □

Lemma 6. *With the same notation as in Lemma 5, suppose instead $\mathbb{E}_q[w(\boldsymbol{z})^{2ps}] < \infty$ and $\mathbb{E}_q\left[|\psi(\boldsymbol{z})|^{2pt}\right] < \infty$, where $s, t \geq 1$ and $\frac{1}{s} + \frac{1}{t} = 1$, then $\mathbb{E}\left[\left|\hat{\pi}^{(k)}[\psi] - \pi[\psi]\right|^{2p}\right] = O(k^{-p+1/t})$.*

Proof. Following (3), it is sufficient to study the term $\mathbb{E}\left[\left|\left(q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right)\hat{\pi}^{(k)}[\psi]\right|^{2p}\right]$. By Hölder's inequality, we have

$$\mathbb{E}\left[\left|\left(q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right)\hat{\pi}^{(k)}[\psi]\right|^{2p}\right] \leq \mathbb{E}\left[\left|q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right|^{2ps}\right]^{1/s} \mathbb{E}\left[\left|\hat{\pi}^{(k)}[\psi]\right|^{2pt}\right]^{1/t}$$

where $s, t \geq 1$ such that $\frac{1}{s} + \frac{1}{t} = 1$. The first term $\mathbb{E}\left[\left|q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right|^{2ps}\right]^{1/s}$ decays at $O\left(k^{-p}\right)$ by Lemma 2.

For the second term $\mathbb{E}\left[\left|\hat{\pi}^{(k)}[\psi]\right|^{2pt}\right]^{1/t} = \mathbb{E}\left[\left|\sum_{i=1}^{k} \bar{w}_i \psi(\boldsymbol{z}_i)\right|^{2pt}\right]^{1/t}$, we have $\mathbb{E}\left[\left|\sum_{i=1}^{k} \bar{w}_i \psi(\boldsymbol{z}_i)\right|^{2pt}\right]^{1/t} \leq \mathbb{E}\left[\left|\sum_{i=1}^{k} \bar{w}_i \left|\psi(\boldsymbol{z}_i)\right|\right|^{2pt}\right]^{1/t} \leq \mathbb{E}\left[\sum_{i=1}^{k} \left|\psi(\boldsymbol{z}_i)\right|^{2pt}\right]^{1/t} \leq k^{1/t}\mathbb{E}\left[\left|\psi(\boldsymbol{z}_i)\right|^{2pt}\right]^{1/t}$ where the second inequality follows from $\bar{w}_i \geq 0, \sum_{i=1}^{k} \bar{w}_i = 1$. Hence $\mathbb{E}\left[\left|\left(q[w] - \hat{q}^{(k)}_{\mathrm{MC}}[w]\right)\hat{\pi}^{(k)}[\psi]\right|^{2p}\right]$ converges to 0 at a $O(k^{-p+1/t})$ rate. □

We recall below the statement of Theorem 3 and give its proof.

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

Proof. We check that the assumptions of Theorem 1 are satisfied.

Assumption (a): $\hat{\pi}^{(k)}[\psi] \to \pi[\psi]$ in L^{2q} by Lemma 6 (choice of q, s, t established later), so $\hat{\pi}^{(k)}[\psi]$ is uniformly integrable, hence similarly to the proof of Theorem 2 $\hat{\pi}^{\text{ML}}[\psi]$ is an unbiased estimator of $\pi[\psi]$.

Assumption (c): This trivially holds.

Assumption (b): We have

$$\begin{aligned}\Delta_k &= \hat{\pi}_{O\cup E}^{(2^{(k+1)})}[\psi] - \frac{1}{2}\left(\hat{\pi}_O^{(2^k)}[\psi] + \hat{\pi}_E^{(2^k)}[\psi]\right) \\ &= \sum_{i=1}^{2^k} \bar{w}_i^{O/O\cup E}\psi(\boldsymbol{z}_i^O) + \sum_{i=1}^{2^k} \bar{w}_i^{E/O\cup E}\psi(\boldsymbol{z}_i^E) - \frac{1}{2}\left(\sum_{i=1}^{2^k} \bar{w}_i^O\psi(\boldsymbol{z}_i^O) + \sum_{i=1}^{2^k} \bar{w}_i^E\psi(\boldsymbol{z}_i^E)\right)\end{aligned}$$

where

$$\bar{w}_i^{O/O\cup E} = \frac{w_i^O}{\sum_{j=1}^{2^k} w_j^O + \sum_{j=1}^{2^k} w_j^E}, \qquad \bar{w}_i^{E/O\cup E} = \frac{w_i^E}{\sum_{j=1}^{2^k} w_j^O + \sum_{j=1}^{2^k} w_j^E}.$$

Consider the coefficients c_i^O, c_i^E for each $\psi(\boldsymbol{z}_i^O)$ and $\psi(\boldsymbol{z}_i^E)$. Rearranging the terms gives that

$$\begin{aligned}c_i^O &= \bar{w}_i^{O/O\cup E} - \frac{1}{2}\bar{w}_i^O = \frac{w_i^O}{\sum_{j=1}^{2^k} w_j^O + \sum_{j=1}^{2^k} w_j^E} - \frac{1}{2}\frac{w_i^O}{\sum_{j=1}^{2^k} w_j^O} \\ &= \frac{w_i^O\left(\sum_{j=1}^{2^k} w_j^O - \sum_{j=1}^{2^k} w_j^E\right)}{2\left(\sum_{j=1}^{2^k} w_j^O\right)\left(\sum_{j=1}^{2^k} w_j^O + \sum_{j=1}^{2^k} w_j^E\right)} = \frac{w_i^O\left(O_k - E_k\right)}{2\sum_{j=1}^{2^k} w_j^O\left(O_k + E_k\right)}\end{aligned}$$

and similarly $c_i^E = \frac{w_i^E(E_k - O_k)}{2\sum_{i=j}^{2^k} w_j^E(O_k + E_k)}$. Hence,

$$\Delta_k = \sum_{i=1}^{2^k} c_i^O\psi(\boldsymbol{z}_i^O) + \sum_{i=1}^{2^k} c_i^E\psi(\boldsymbol{z}_i^E) = \frac{O_k - E_k}{2\left(O_k + E_k\right)}\left(\sum_{i=1}^{2^k} \bar{w}_i^O\psi(\boldsymbol{z}_i^O) - \sum_{i=1}^{2^k} \bar{w}_i^E\psi(\boldsymbol{z}_i^E)\right).$$

Consequently, we have

$$\mathbb{E}\left[\Delta_k^2\right] = \mathbb{E}\left[\frac{\left(O_k - E_k\right)^2}{4\left(O_k + E_k\right)^2}\left(\hat{\pi}_O^{(2^k)}[\psi] - \hat{\pi}_E^{(2^k)}[\psi]\right)^2\right] \leq \mathbb{E}\left[\frac{\left|O_k - E_k\right|^{2p}}{4^p\left(O_k + E_k\right)^{2p}}\right]^{1/p} \mathbb{E}\left[\left|\hat{\pi}_O^{(2^k)}[\psi] - \hat{\pi}_E^{(2^k)}[\psi]\right|^{2q}\right]^{1/q}$$

for $\frac{1}{p} + \frac{1}{q} = 1$ by Hölder's inequality.

By Lemma 6, $\mathbb{E}\left[\left|\hat{\pi}_O^{(2^k)}[\psi] - \pi[\psi]\right|^{2q}\right]$ decays at rate $O(2^{-(q-1/t)k})$ if $\mathbb{E}_q[w(\boldsymbol{z})^{2qs}] < \infty$, $\mathbb{E}_q\left[|\psi(\boldsymbol{z})|^{2qt}\right] < \infty$, and $\frac{1}{s} + \frac{1}{t} = 1$. By Lemma 4, we conclude that if $a \geq 2qs$, $b \geq 2qt$, $\mathbb{E}\left[\left|\hat{\pi}_O^{(2^k)}[\psi] - \hat{\pi}_E^{(2^k)}[\psi]\right|^{2q}\right]^{1/q} = O\left(2^{-\left(1-\frac{1}{qt}\right)k}\right)$.

For $\mathbb{E}\left[\frac{|O_k - E_k|^{2p}}{4^p(O_k + E_k)^{2p}}\right]$, we can similarly consider two cases.

1. Case O_k, E_k large, i.e. $O_k \geq 2^{-\kappa}Z$ and $E_k \geq 2^{-\kappa}Z$, where $\kappa > 0$ is a constant:

Since w_i has finite a moments, if we choose $2p \leq a$, we have for the numerator $\mathbb{E}\left[|O_k - E_k|^{2p}\right] = O(2^{-pk})$ using Lemma 2 and 4, and the denominator is bounded below. Hence $\mathbb{E}\left[\frac{|O_k-E_k|^{2p}}{(O_k+E_k)^{2p}}\mathbb{I}\left\{O_k \geq 2^{-\kappa}Z, E_k \geq 2^{-\kappa}Z\right\}\right] = O(2^{-pk})$.

If we choose $2p \geq a$, we have that

$$\frac{|O_k - E_k|^{2p}}{(O_k + E_k)^{2p}} \leq \frac{|O_k - E_k|^{a}}{(O_k + E_k)^{a}},$$

since $\frac{|O_k-E_k|^{2p}}{(O_k+E_k)^{2p}}$ is bounded above by 1, and similarly $\mathbb{E}\left[\frac{|O_k-E_k|^{2p}}{(O_k+E_k)^{2p}}\mathbb{I}\left\{O_k \geq 2^{-\kappa}Z, E_k \geq 2^{-\kappa}Z\right\}\right] = O(2^{-ak/2})$.

2. Case O_k or E_k small, i.e. $O_k < 2^{-\kappa}Z$ or $E_k < 2^{-\kappa}Z$:

We have that $\frac{|O_k-E_k|^{2p}}{(O_k+E_k)^{2p}}$ is bounded above by 1. Hence, we have

$$\begin{aligned}\mathbb{E}\left[\frac{|O_k - E_k|^{2p}}{(O_k + E_k)^{2p}}\mathbb{I}\left(O_k < 2^{-\kappa}Z\right)\right] &\leq \mathbb{P}\left(O_k < 2^{-\kappa}Z\right) \\ &= O(2^{-ak/2}),\end{aligned}$$

where the last line is by the proof of Theorem 2. The case is similar for $E_k < 2^{-\kappa}Z$.

Therefore, we have that if we choose $2p \leq a$, $\mathbb{E}\left[\frac{|O_k-E_k|^{2p}}{4^p(O_k+E_k)^{2p}}\right]$ is $O(2^{-pk})$, hence $\mathbb{E}\left[\frac{|O_k-E_k|^{2p}}{4^p(O_k+E_k)^{2p}}\right]^{1/p}$ is $O\left(2^{-k}\right)$. Aggregating the results, we have that $\mathbb{E}\left[\Delta_k^2\right] = O\left(2^{-k} \cdot 2^{-\left(1-\frac{1}{qt}\right)k}\right) = O\left(2^{-\left(2-\frac{1}{qt}\right)k}\right)$. Collecting together the conditions, we must have that

$$\begin{aligned}2qs &\leq a \\ 2qt &\leq b \\ 2p &\leq a\end{aligned}$$

as well as the Hölder conjugate conditions $\frac{1}{p} + \frac{1}{q} = 1$, $\frac{1}{s} + \frac{1}{t} = 1$, $p, q, s, t \geq 1$.

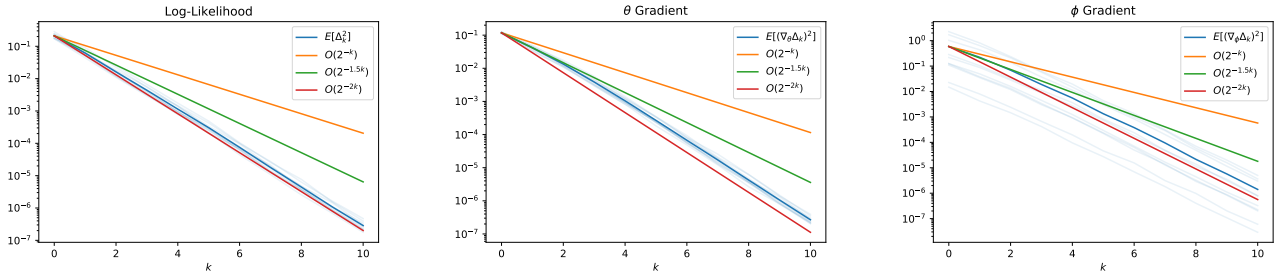
Let $u = qt$, which we would like to maximize. Then solving the equation we have equivalently that

$$\begin{aligned}u &\leq \frac{a}{2}t - \frac{a}{2} \\ u &\leq \frac{b}{2} \\ u &\geq \frac{a}{a-2}t.\end{aligned}$$

In order for this set to be valid, one can check that we must have $\frac{a+b}{a} \leq \frac{b(a-2)}{2a}$, i.e. $a > 4$, $b \geq \frac{2a}{a-4}$. Solving for the maximum u, we have that $u = \frac{b}{2}$, and the rest can be taken as $t = \frac{b(a-2)}{2a}$, $s = \frac{t}{t-1}$, $p = \frac{u}{u-t} = \frac{a}{2}$ and $q = \frac{u}{t} = \frac{a}{a-2}$. Thus we have overall a convergence rate $O\left(2^{-\left(2-\frac{2}{b}\right)k}\right)$. For multilevel Monte Carlo to work, we must have the strict inequality that $\frac{2}{b} < 1$, i.e. $b > 2$, so that $\alpha = 1 - \frac{2}{b} > 0$, but this is implied by $b \geq \frac{2a}{a-4}$.

If we choose $2p \geq a$, $\mathbb{E}\left[\frac{|O_k-E_k|^{2p}}{4^p(O_k+E_k)^{2p}}\right]$ is $O(2^{-ak/2})$, hence $\mathbb{E}\left[\frac{|O_k-E_k|^{2p}}{4^p(O_k+E_k)^{2p}}\right]^{1/p}$ is $O(2^{-\frac{ak}{2p}})$. Aggregating the results, we have that $\mathbb{E}\left[\Delta_k^2\right] = O\left(2^{-\frac{ak}{2p}} \cdot 2^{-\left(1-\frac{1}{qt}\right)k}\right) = O\left(2^{-\left(1+\frac{a}{2p}-\frac{1}{qt}\right)k}\right)$. Collecting together the conditions, we must have that

$$\begin{aligned}2qs &\leq a \\ 2qt &\leq b \\ 2p &> a\end{aligned}$$



(a) (b) (c)

Figure 1: Empirical confirmation of the convergence rate of $\mathbb{E}[\Delta_k^2]$ for (a) $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$; (b) $\hat{\pi}^{\mathrm{ML}}[\psi]$ with $\psi(\boldsymbol{z}) := \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$; and (c) $\hat{\pi}^{\mathrm{ML}}[\psi]$ with $\psi(\boldsymbol{z}) := -\nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. For all three estimators, the convergence rate of $\mathbb{E}[\Delta_k^2]$ is close to the theoretical upper bound $O(2^{-2k})$.

as well as the Hölder conjugate conditions $\frac{1}{p} + \frac{1}{q} = 1$, $\frac{1}{s} + \frac{1}{t} = 1$, $p, q, s, t \geq 1$. Notice here that $\frac{a}{2p} \leq 1$ by our assumption, and that $2qt \leq b$, so the overall convergence rate $O\left(2^{-\left(1+\frac{a}{2p}-\frac{1}{qt}\right)k}\right)$ is slower than $O\left(2^{-\left(2-\frac{2}{b}\right)k}\right)$. Therefore, when $a > 4$, $b \geq \frac{2a}{a-4}$ holds, we already have the best convergence rate $O\left(2^{-\left(2-\frac{2}{b}\right)k}\right)$.

We would like to maximize $1 + \frac{a}{2p} - \frac{1}{qt} = 1 + \frac{a}{2} - \frac{a}{2q} - \frac{1}{qt}$, so we would like to minimize $\frac{a}{2q} + \frac{1}{qt} = \frac{at+2}{2qt}$. Again letting $u = qt$, we have equivalently that

$$\begin{aligned} u &\leq \frac{a}{2}t - \frac{a}{2} \\ u &\leq \frac{b}{2} \\ u &\leq \frac{a}{a-2}t \\ u &\geq t. \end{aligned}$$

(In the case $a \leq 4$ or $a > 4, b \leq \frac{2a}{a-4}$, the third inequality is slack.) Solving again, the minimum $\frac{at+2}{2qt} = 1 + \frac{a+2}{b}$ is attained at $u = \frac{b}{2}$, $t = \frac{a+b}{a}$, $s = \frac{a+b}{b}$, $p = \frac{u}{u-t} = \frac{ab}{ab-2a-2b}$ and $q = \frac{u}{t} = \frac{ab}{2(a+b)}$. Hence the overall convergence rate $O\left(2^{-\left(1+\frac{a}{2p}-\frac{1}{qt}\right)k}\right) = O\left(2^{-\left(\frac{a}{2}-\frac{a+2}{b}\right)k}\right)$. For multilevel Monte Carlo to work, we must have the strict inequality that $\frac{a}{2} - \frac{a+2}{b} > 1$, i.e. $a > 2$, $b > \frac{2a+4}{a-2}$, so that $\alpha = \frac{a}{2} - \frac{a+2}{b} - 1 > 0$. □

Corollary 7. *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 $\hat{\pi}^{\mathrm{ML}}[\log w]$ satisfies Theorem 1 for $r \in \left(\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}}\right)$, where $\alpha = \min\left(\frac{\epsilon}{2}, 1\right) > 0$.*

Proof. By Lemma 1, all moments of $\log w$ are finite under $q_{\boldsymbol{\phi}}$. Therefore, taking b arbitrarily large in Theorem 3 gives us the result. □

4 Empirical Confirmation of the Convergence Rate of $\mathbb{E}[\Delta_k^2]$

We confirm our theoretical results and verify the convergence rate of $\mathbb{E}[\Delta_k^2]$ using the same example as in Section 7.1 for the three estimators: $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$, $\hat{\pi}^{\mathrm{ML}}[\psi]$ with $\psi(\boldsymbol{z}) := \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})$, and $\hat{\pi}^{\mathrm{ML}}[\psi]$ with $\psi(\boldsymbol{z}) := -\nabla_{\boldsymbol{\phi}} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$. The experiments are conducted with 10 random perturbations and 1000 estimator samples per each perturbation, and the values of $\mathbb{E}[\Delta_k^2]$ are taken as the mean of the 10 trials. We observe that $\mathbb{E}[\Delta_k^2]$ for all estimators converges much faster than $O\left(2^{-1.5k}\right)$ (corresponding to $\alpha = 0.5$) and very close to our theoretical limit $O(2^{-2k})$ (corresponding to $\alpha = 1$) when k is large. Therefore, Theorem 1 justifies the use of our approach with approximately $r \in (0.5, 0.75)$.

5 Comparing Single Sample (SS) and Russian Roulette (RR) Estimators

For simplicity, we consider the first term I_0 and the telescoping part $\tilde{\text{SS}} = \frac{\Delta_K}{p(K)}$, $\tilde{\text{RR}} = \sum_{k=0}^{K} \frac{\Delta_k}{\mathbb{P}(K \geq k)}$ in SS and RR separately. If I_0 is sampled independently, the variance of SS or RR is simply the sum of the variance of the two parts.

5.1 Variance of Single Sample (SS) Estimator

It is straightforward to show that

$$\mathbb{E}\left[\tilde{\text{SS}}\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[\Delta_k\right], \qquad \mathbb{E}\left[\tilde{\text{SS}}^2\right] = \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[\Delta_k^2\right]}{p(k)}. \tag{4}$$

Hence we have

$$\begin{aligned}
\text{Var}\left(\tilde{\text{SS}}\right) &= \mathbb{E}\left[\tilde{\text{SS}}^2\right] - \mathbb{E}\left[\tilde{\text{SS}}\right]^2 && (5)\\
&= \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[\Delta_k^2\right]}{p(k)} - \left(\sum_{k=0}^{\infty} \mathbb{E}\left[\Delta_k\right]\right)^2 && (6)\\
&= \underbrace{\sum_{k=0}^{\infty} \frac{\text{Var}\left(\Delta_k\right)}{p(k)}}_{①} + \underbrace{\left(\sum_{k=0}^{\infty} \frac{\mathbb{E}\left[\Delta_k\right]^2}{p(k)} - \sum_{k=0}^{\infty} \mathbb{E}\left[\Delta_k\right]^2 - 2 \sum_{i,j:i<j} \mathbb{E}\left[\Delta_i\right]\mathbb{E}\left[\Delta_j\right]\right)}_{②}. && (7)
\end{aligned}$$

5.2 Variance of Russian Roulette (RR) Estimator

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

5.3 Proof of Theorem 1 for Russian Roulette (RR)

In Blanchet et al. (2019), only Theorem 1 for SS is mentioned. Here we show that the theorem similarly holds for RR.

Theorem. *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 $\mathbb{E}\left[\Delta_k^2\right] \le c \cdot 2^{-(1+\alpha)k}$ for all k;

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

then for $K \sim \mathrm{Geom}(r)$ where $r \in (\frac{1}{2}, 1 - \frac{1}{2^{1+\alpha}})$, RR is an unbiased estimator of I_∞, whose variance and expected sampling cost are both finite.

Proof. The unbiasedness of RR is standard. Finite expected sampling cost follows directly from Theorem 1 for SS, since the sampling cost of SS and RR are the same. To show finite variance, we show that

$$\mathbb{E}\left[\tilde{\mathrm{RR}}^2\right] = \sum_{k=0}^{\infty} \frac{\mathbb{E}\left[\Delta_k^2\right]}{\mathbb{P}(K \ge k)} + 2 \sum_{i,j:i<j} \frac{\mathbb{E}[\Delta_i \Delta_j]}{\mathbb{P}(K \ge i)}$$

is finite. Since $\mathbb{E}\left[\Delta_k^2\right] \le c \cdot 2^{-(1+\alpha)k}$, for fixed i

$$\begin{aligned}\sum_{j>i} |\mathbb{E}[\Delta_i \Delta_j]| &\le \sqrt{c} \cdot 2^{-\frac{1}{2}(1+\alpha)i} \cdot \sqrt{c} \cdot 2^{-\frac{1}{2}(1+\alpha)(i+1)} \cdot \frac{1}{1 - 2^{-\frac{1}{2}(1+\alpha)}} \\ &= O(2^{-(1+\alpha)i}).\end{aligned}$$

Now $\mathbb{P}(K \ge k) = (1-r)^k$, where $1 - r \in \left(2^{-(1+\alpha)}, \frac{1}{2}\right)$. Therefore $\mathbb{E}\left[\tilde{\mathrm{RR}}^2\right] < \infty$, and RR has finite variance. □

6 Using Thermodynamic Integration

6.1 Proof of Proposition 5

Proposition. *The following identity holds*

$$\hat{\ell}^{(k)}(\boldsymbol{\theta}) = \int_0^1 \hat{\pi}_\beta^{(k)}[\log w] \mathrm{d}\beta. \tag{8}$$

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 $\mathrm{Var}(\hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})) \ge \mathrm{Var}(\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta}))$.

Proof. We have that

$$\hat{\pi}_\beta^{(k)}[\log w] = \sum_{i=1}^{k} \frac{w_i^\beta}{\sum_{j=1}^k w_j^\beta} \log w_i. \tag{9}$$

By directly differentiating $\beta \mapsto F(\beta) = \log\left(\frac{1}{k}\sum_{i=1}^k w_i^\beta\right)$ with respect to β, we have that $F'(\beta) = \sum_{i=1}^k \frac{w_i^\beta}{\sum_{j=1}^k w_j^\beta} \log w_i$. Hence $\int_0^1 \hat{\pi}_\beta^{(k)}[\log w]\mathrm{d}\beta = F(1) - F(0) = \log\left(\frac{1}{k}\sum_{i=1}^k w_i\right) = \ell^{(k)}(\boldsymbol{\theta})$.

For $\hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})$,

$$\begin{aligned}\Delta_k = &\sum_{i=1}^{2^k} \frac{\left(w_i^O\right)^\beta}{\sum_{j=1}^{2^k}\left(w_j^O\right)^\beta + \sum_{j=1}^{2^k}\left(w_j^E\right)^\beta} \log w_i^O + \sum_{i=1}^{2^k} \frac{\left(w_i^E\right)^\beta}{\sum_{j=1}^{2^k}\left(w_j^O\right)^\beta + \sum_{j=1}^{2^k}\left(w_j^E\right)^\beta} \log w_i^E \\ &- \frac{1}{2}\left(\sum_{i=1}^{2^k} \frac{\left(w_i^O\right)^\beta}{\sum_{j=1}^{2^k}\left(w_j^O\right)^\beta} \log w_i^O + \sum_{i=1}^{2^k} \frac{\left(w_i^E\right)^\beta}{\sum_{j=1}^{2^k}\left(w_j^E\right)^\beta} \log w_i^E\right).\end{aligned}$$

Directly integrating gives that $\int_0^1 \Delta_k(\beta)\mathrm{d}\beta = \log\left(\frac{1}{2}(O_K + E_K)\right) - \frac{1}{2}(\log O_K + \log E_K)$, which corresponds to Δ_k for $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$. Hence $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta}) = \int_0^1 \hat{\ell}_{\mathrm{TI}}^{\mathrm{ML}}(\boldsymbol{\theta})\mathrm{d}\beta$. □

6.2 Proof of Proposition 6

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

Proof. $\hat{\ell}^{(k)}_{\text{TVO}}(\boldsymbol{\theta}) = \sum_{t=0}^{T-1}(\beta_{t+1}-\beta_t)\hat{\pi}^{(k)}_{\beta_t}[\log w]$ can be viewed as a left Riemann sum approximation of the integral $\int_0^1 \hat{\pi}^{(k)}_{\beta}[\log w]\mathrm{d}\beta = \hat{\ell}^{(k)}(\boldsymbol{\theta})$, where the equality is due to Proposition 5. However,

$$\frac{\partial}{\partial\beta}\hat{\pi}^{(k)}_{\beta}[\log w] = \frac{\left(\sum_{i=1}^k w_i^\beta (\log w_i)^2\right)\left(\sum_{j=1}^k w_j^\beta\right) - \left(\sum_{i=1}^k w_i^\beta \log w_i\right)\left(\sum_{j=1}^k w_j^\beta \log w_j\right)}{\left(\sum_{j=1}^k w_j^\beta\right)^2} \geq 0$$

by the Cauchy–Schwarz inequality. Therefore, the integrand $\hat{\pi}^{(k)}_{\beta}[\log w]$ is non-decreasing in β, and the left Riemann sum approximation $\hat{\ell}^{(k)}_{\text{TVO}}(\boldsymbol{\theta}) \leq \hat{\ell}^{(k)}(\boldsymbol{\theta})$, hence $\mathbb{E}[\hat{\ell}^{(k)}_{\text{TVO}}(\boldsymbol{\theta})] \leq \mathbb{E}[\hat{\ell}^{(k)}(\boldsymbol{\theta})] = \ell^{(k)}_{\text{IWAE}}(\boldsymbol{\theta})$. □

6.3 Proof of Theorem 7

By the propositions, it is thus not favorable to unbiasedly estimate the log-likelihood via the thermodynamic identity, or use SNIS to biasedly estimate $\ell_{\text{TVO}}(\boldsymbol{\theta})$. However, it is possible to unbiasedly estimate $\ell_{\text{TVO}}(\boldsymbol{\theta})$, and an advantage is that the moment conditions in Theorem 1 can be relaxed for multilevel Monte Carlo to have finite variance:

Theorem. *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* $\hat{\ell}^{\text{ML}}_{\text{TVO}}(\boldsymbol{\theta})$ *satisfies Theorem 1 for* $r \in \left(\frac{1}{2}, 1-\frac{1}{2^{1+\alpha}}\right)$, *where* $\alpha = \min\left(\frac{\epsilon}{2}, 1\right)$.

Proof. This is a direct application of Corollary 7. Unbiasedness and finite expected computation follows directly from the unbiasedness and finite expected computation of $\hat{\pi}^{(k)}_{\beta_t}[\log w]$ for each β_t.

Finite variance follows directly from Corollary 7 as well, since in order for the tempered importance weights $w(\boldsymbol{z})^{\beta_t}$ in $\hat{\pi}^{(k)}_{\beta_t}[\log w]$ to have finite $2+\epsilon$ moments, equivalently $w(\boldsymbol{z})$ only needs to have finite $\beta_t(2+\epsilon)$ moments. Since we require this for all β_t in $0 = \beta_0 < \beta_1 < \cdots < \beta_{T-1} < 1$, we require that $\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[w(\boldsymbol{z})^{\beta_{T-1}(2+\epsilon)}\right] < \infty$. □

6.4 SNIS Bias in the Covariance Gradient Estimator for $\boldsymbol{\theta}$

We note that Masrani et al. (2019) directly consider the gradient via the covariance gradient estimator

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

for $\boldsymbol{\lambda} = \{\boldsymbol{\theta}, \boldsymbol{\phi}\}$, where

$$\mathrm{Cov}_{\pi_\beta}[\nabla_{\boldsymbol{\lambda}}\log\tilde{\pi}_\beta, \log w] = \mathbb{E}_{\pi_\beta}\left[\left(\nabla_{\boldsymbol{\lambda}}\log\tilde{\pi}_\beta - \mathbb{E}_{\pi_\beta}\left[\nabla_{\boldsymbol{\lambda}}\log\tilde{\pi}_\beta\right]\right)\left(\log w - \mathbb{E}_{\pi_\beta}\left[\log w\right]\right)\right].$$

However, Masrani et al. (2019) use SNIS for the RHS, and in particular for the $\boldsymbol{\theta}$ gradients the SNIS estimator for the covariance gradient estimator is precisely

$$\begin{aligned}
&\sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log w_i + \sum_{i=1}^k \overline{w_i^\beta}\left(\nabla_{\boldsymbol{\theta}}\log\tilde{\pi}_\beta(\boldsymbol{z}_i) - \sum_{j=1}^k \overline{w_j^\beta}\nabla_{\boldsymbol{\theta}}\log\tilde{\pi}_\beta(\boldsymbol{z}_j)\right)\left(\log w_i - \sum_{j=1}^k \overline{w_j^\beta}\log w_j\right)\\
&= \sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log w_i + \sum_{i=1}^k \overline{w_i^\beta}\log w_i\nabla_{\boldsymbol{\theta}}\log\tilde{\pi}_\beta(\boldsymbol{z}_i) - \left(\sum_{i=1}^k \overline{w_i^\beta}\log w_i\right)\left(\sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log\tilde{\pi}_\beta(\boldsymbol{z}_i)\right)\\
&= \sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log w_i + \beta\sum_{i=1}^k \overline{w_i^\beta}\log w_i\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}_i) - \left(\sum_{i=1}^k \overline{w_i^\beta}\log w_i\right)\left(\beta\sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}_i)\right)
\end{aligned}$$

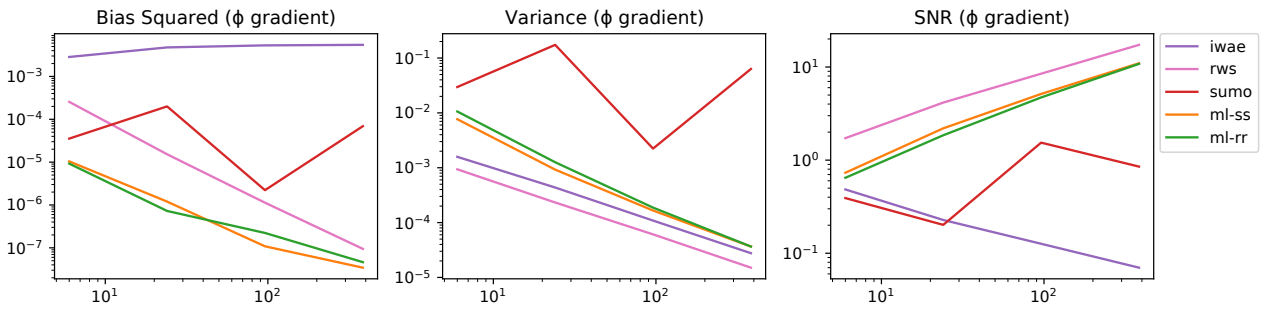


Figure 2: Empirical bias squared, variance and signal-to-noise ratio (SNR) of estimators of $\nabla_{\boldsymbol{\phi}}\mathrm{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$, plotted against expected computational cost.

This is the identical to differentiating the SNIS estimator $\hat{\pi}_\beta^{(k)}[\log w]$ directly, since

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\hat{\pi}_\beta^{(k)}[\log w] &= \nabla_{\boldsymbol{\theta}}\sum_{i=1}^k \frac{w_i^\beta}{\sum_{j=1}^k w_j^\beta}\log w_i \\
&= \sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log w_i + \beta\sum_{i=1}^k \overline{w_i^\beta}\log w_i \nabla_{\boldsymbol{\theta}}\log w_i - \frac{\left(\sum_{i=1}^k w_i^\beta \log w_i\right)\left(\beta\sum_{j=1}^k w_j^\beta \nabla_{\boldsymbol{\theta}}\log w_j\right)}{\left(\sum_{j=1}^k w_j^\beta\right)^2} \\
&= \sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log w_i + \beta\sum_{i=1}^k \overline{w_i^\beta}\log w_i \nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}_i) - \left(\sum_{i=1}^k \overline{w_i^\beta}\log w_i\right)\left(\beta\sum_{i=1}^k \overline{w_i^\beta}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z}_i)\right).
\end{aligned}$$

Therefore, using the covariance gradient estimator for $\boldsymbol{\theta}$ is the same as $\nabla_{\boldsymbol{\theta}}\hat{\pi}_\beta^{(k)}[\log w]$, so the SNIS bias in $\hat{\ell}_{\mathrm{TVO}}^{(k)}(\boldsymbol{\theta})$ carries over to the covariance estimator.

We note that this analysis applies to the $\boldsymbol{\theta}$ objective and its gradients, when viewing TVO as an estimator of $\ell(\boldsymbol{\theta})$, but not the use of TVO for the $\boldsymbol{\phi}$ gradients.

7 Further Experiment Details

We provide some further experiment details and results in this section.

7.1 Linear Gaussian Experiment

With the same linear Gaussian example, we further analyze different estimators of the gradient of the forward KL $\nabla_{\boldsymbol{\phi}}\mathrm{KL}(p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})||q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}))$. We compare the ML-SS, ML-RR and SUMO estimators of $\hat{\pi}[\psi], \psi(\boldsymbol{z}) := -\nabla_{\boldsymbol{\phi}}\log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ against the $\boldsymbol{\phi}$ gradient in Reweighted Wake-Sleep (RWS), which uses a biased SNIS estimator.

From Figure 2, we observe that SS and RR achieve the lowest bias squared again compared to RWS and SUMO at the same computational cost, while obtaining higher variance than RWS but lower variance than SUMO. As IWAE suffers from the vanishing gradient problem, SS and RR also achieve higher SNR than the IWAE estimator, even at a low computational cost.

7.2 2D Density Modeling Experiment

For the second example, we compare VAE training using different estimators with the Figure-8 dataset proposed by Yacoby et al. (2020). We use the same flexible VAE architecture, with the encoder and decoder both parameterized by a neural network with 3 hidden layers and 50 leaky ReLU units in each layer, so that the model is flexible enough to learn the ground truth model. We use the Adam optimizer with a learning rate of $2.5 \cdot 10^{-4}$ and train for a maximum of 2000 epochs with early stopping applied based on the validation set.

Table 1: Test negative log-likelihood of the trained models for the Figure-8 dataset when trained using different estimators and $q_{\boldsymbol{\phi}}$ objectives. All estimators have an expected sampling cost of 5 terms (except for SUMO which has a expected sampling cost of 6 terms).

(a)

Baselines	
ELBO	0.9659±0.0046
IWAE	0.9325±0.0054
RWS-STL	0.9387±0.0074

(b)

	$q_{\boldsymbol{\phi}}$ objectives			
Estimators	ELBO	IWAE	Unbiased RWS-STL	Var
SUMO	0.9836±0.0107	**0.9278±0.0058**	**0.9222±0.0036**	**0.9226±0.0091**
ML-SS	0.9843±0.0029	**0.9263±0.0019**	**0.9216±0.0041**	**0.9228±0.0076**
ML-RR	0.9727±0.0057	**0.9258±0.0027**	**0.9202±0.0048**	**0.9222±0.0071**

Table 2: Other test metrics of the trained models for the Figure-8 dataset when trained with the ML-SS log-likelihood estimator and different ϕ objectives.

	$q_{\boldsymbol{\phi}}$ objectives			
	ELBO	IWAE	Unbiased RWS-STL	Var
$\mathrm{KL}(q_{\boldsymbol{\phi}} \parallel p_{\boldsymbol{\theta}})$	**0.1570±0.0114**	0.7390±0.2292	1.0066±0.2469	1.3684±0.0663
$\mathbb{E}[w^4]^{1/4}$	3.6667±0.4088	2.1612±0.7074	1.7557±0.1921	**1.7534±0.2612**
$\hat{k}$	1.0122±0.2337	0.2475±0.1783	0.3259±0.1668	**-0.7768±0.1823**

We provide a further study of suitable training objectives for $q_{\boldsymbol{\phi}}$ here. Although the SUMO, ML-SS, and ML-RR log-likelihood estimators are unbiased estimators under reasonable conditions, we observe from Table 1 that the choice of the $q_{\boldsymbol{\phi}}$ objective can still have a large impact on training. Maximizing the ELBO (or equivalently minimizing the reverse KL) resulted in the worst test log-likelihood among all tested $q_{\boldsymbol{\phi}}$ training objectives due to the mode-seaking property of the reverse KL. In particular, an important quantity we considered is $\mathbb{E}[w^4]^{1/4}$, which is an essential quantity in our proofs of finite variance. We also compute the Pareto-smoothed importance sampling diagnostic $\hat{k}$ (Vehtari et al., 2016), whose inverse gives an estimate of the number of existing moments of w when $\hat{k} > 0$. Finite variance of $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$ can be achieved approximately when $\hat{k} < 0.5$. From Table 2, we see that maximizing the ELBO for $q_{\boldsymbol{\phi}}$ resulted in the lowest reverse KL but the highest sample 4th moment and $\hat{k} > 1$, which confirms that $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ does not cover the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ well and can lead to infinite variance even biased estimators. On the contrary, using the variance or the unbiased RWS-STL objective (both reuse samples and do not require generating new samples) achieved the best test log-likelihood, and in particular the variance objective achieves the lowest value of $\hat{k} < 0$, which justifies our approach and indicates finite variance of $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$ for all $r \in (0.5, 0.75)$.

7.3 Image Modeling Experiment

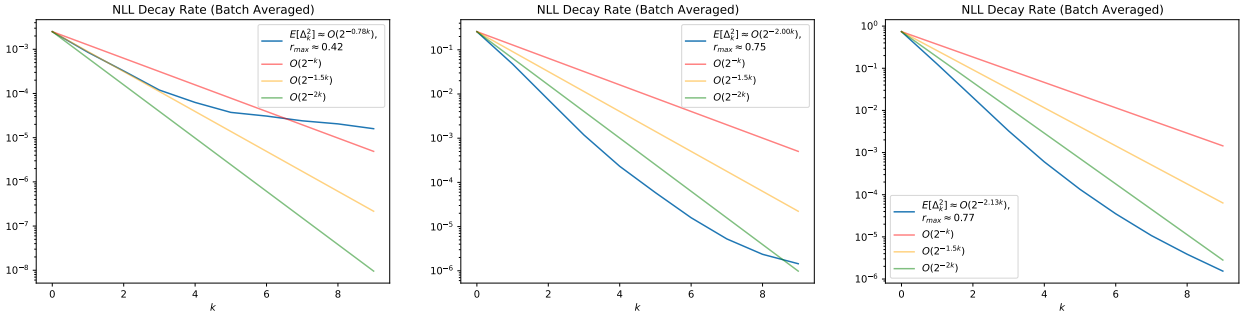
For the image modeling experiment, we follow the training scheme proposed by Luo et al. (2020, Appendix A.8) with batch size 100, the Amsgrad optimizer (Reddi et al., 2018) and the same gradient norm clipping scheme for $p_{\boldsymbol{\theta}}$. However, in order to accurately compare different estimators under same budget, we fix the number of training epochs to 3280 for all estimators and use a linear learning rate decay from 10^{-3} to 10^{-4}. We also only modify the $p_{\boldsymbol{\theta}}$ objective for this experiment, with different estimators in place of IWAE for $p_{\boldsymbol{\theta}}$, which provides a fair comparison to the IWAE baseline. IWAE is used to train $q_{\boldsymbol{\phi}}$ and new samples are drawn for all estimators.

For this example, we find that the convergence rate of $\mathbb{E}[\Delta_k^2]$ at initialization is $O\left(2^{-(1+\alpha)k}\right)$ with $\alpha \approx 0.5$, which justifies the use of $r \in (0.5, 0.6464)$, but the value of α decays to less than 0 as training progresses. To solve this, similar to Luo et al. (2020), we propose to modify the tail of the sampling distribution of K as described in Section 5.4. We set $k_{\max}$ so that ML-RR is an unbiased estimator of IWAE with $k = 128$, while having a much smaller expected computational cost (5 or 15 terms in our experiments)[1]. This can also lower the

[1]Luo et al. (2020) propose to softly truncates the tail of K after $k = 80$ terms, so our modification is in line with

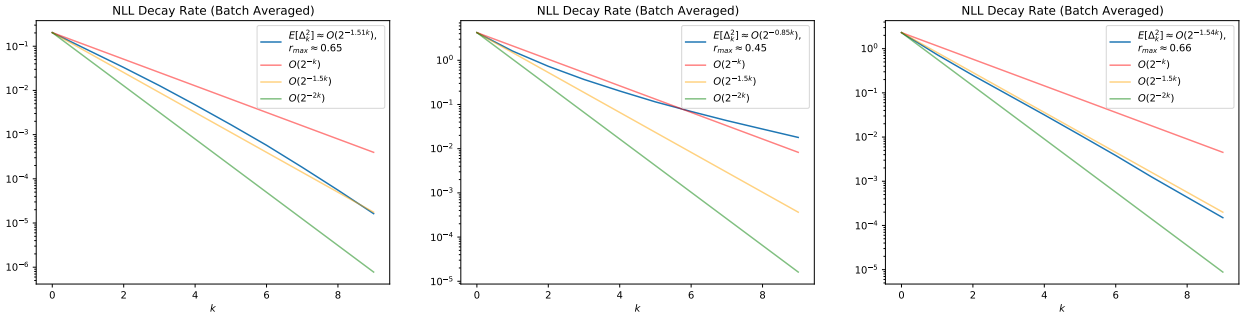
computational cost and limit the computation/memory usage to be less than $2^{k_{\max}+1}$ samples. For ML-TVO-RR , we use 5 intermediate β_t values and follow Masrani et al. (2019) for their placement. By applying thermodynamic integration, faster convergence rate of $\mathbb{E}[\Delta_k^2]$ can be established. This shows that the relaxation of the moment condition from $\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[w(\boldsymbol{z})^{2+\epsilon}\right] < \infty$ in Theorem 2 of main text to $\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[w(\boldsymbol{z})^{\beta_{T-1}(2+\epsilon)}\right] < \infty$ in Theorem 7 is crucial to the convergence speed of $\mathbb{E}[\Delta_k^2]$, at the cost of an approximation error in the numerical integration of the 1D integral.

existing methods.



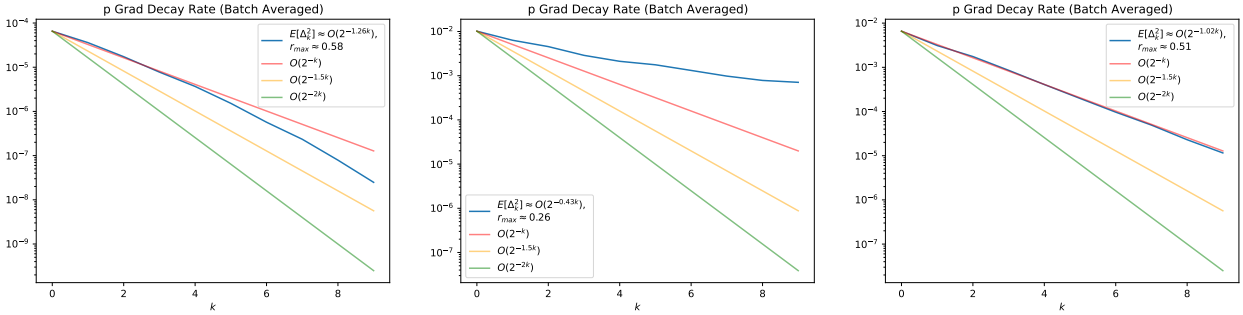
(a) (b) (c)

Figure 3: Empirical confirmation of the convergence rate of $\mathbb{E}[\Delta_k^2]$ for $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$, evaluated on the models trained with $p_{\boldsymbol{\theta}}$ objective $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta})$ and $q_{\boldsymbol{\phi}}$ objective (a) ELBO; and (b) unbiased RWS-STL; (c) Var $\hat{\ell}^{\text{ML-SS}}(\boldsymbol{\theta})$ on the Figure-8 dataset. Clearly training with ELBO as the $q_{\boldsymbol{\phi}}$ objective results in much slower convergence speed of $\mathbb{E}[\Delta_k^2]$.



(a) (b) (c)

Figure 4: Empirical confirmation of the convergence rate of $\mathbb{E}[\Delta_k^2]$ for $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$, evaluated on (a) model at initialization; (b) model trained with $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$ on the MNIST dataset. (c) Convergence rate of $\mathbb{E}[\Delta_k^2]$ for $\hat{\ell}^{\mathrm{ML}}_{\mathrm{TVO}}(\boldsymbol{\theta})$, evaluated on model trained with $\hat{\ell}^{\mathrm{ML}}_{\mathrm{TVO}}(\boldsymbol{\theta})$. $T = 5$ and $\beta_1 = 0.01, \beta_{T-1} \approx 0.3162$ is taken using the log-uniform spacing as in Masrani et al. (2019).



(a) (b) (c)

Figure 5: Empirical confirmation of the convergence rate of $\mathbb{E}[\Delta_k^2]$ for $\nabla_{\boldsymbol{\theta}}\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$, evaluated on (a) model at initialization; (b) model trained with $\hat{\ell}^{\mathrm{ML}}(\boldsymbol{\theta})$ on the MNIST dataset. (c) Convergence rate of $\mathbb{E}[\Delta_k^2]$ for $\nabla_{\boldsymbol{\theta}}\hat{\ell}^{\mathrm{ML}}_{\mathrm{TVO}}(\boldsymbol{\theta})$, evaluated on model trained with $\hat{\ell}^{\mathrm{ML}}_{\mathrm{TVO}}(\boldsymbol{\theta})$.

References

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.

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

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

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.

Sashank Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, August 2016.

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.