Supplementary Material: Evaluating the Variance of Likelihood-Ratio Gradient Estimators

Seiya Tokui [1 2] **Issei Sato** [3 2]

A. Derivations of Estimators for Binary Variables

In this section, we give the derivations of estimators for binary variables given in Sec. 6. Let

$$q_{\phi_i}(z_i|\mathrm{pa}_i) = \mu_i^{z_i}(1-\mu_i)^{1-z_i}$$

be a Bernoulli distribution of a mean parameter $\mu_i = \mu_i(\mathrm{pa}_i, \phi_i)$. Here we only focus on the derivative w.r.t. μ_i instead of ϕ_i. The derivative w.r.t. ϕ_i can be derived by simply multiplying $\nabla_{\phi_i}\mu_i$ to the derivative given in this section.

The log probability of z_i is given by

$$\log q_{\phi_i}(z_i|\mathrm{pa}_i) = z_i \log \mu_i + (1-z_i)\log(1-\mu_i),$$

and its derivative is

$$\frac{\partial}{\partial \mu_i}\log q_{\phi_i}(z_i|\mathrm{pa}_i) = \frac{z_i}{\mu_i} - \frac{1-z_i}{1-\mu_i} = \begin{cases} \frac{1}{\mu_i} & \text{if } z_i = 1 \\ -\frac{1}{1-\mu_i} & \text{if } z_i = 0. \end{cases} \tag{1}$$

The derivative of the probability is simply given as follows.

$$\frac{\partial}{\partial \mu_i} q_{\phi_i}(z_i|\mathrm{pa}_i) = \begin{cases} 1 & \text{if } z_i = 1 \\ -1 & \text{if } z_i = 0. \end{cases} \tag{2}$$

Let $f_k = f(x, z_i = k, z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i}))$ be the simulated objective function f evaluated at fixed $z_i = k \in \{0, 1\}$ and the noise $\epsilon_{\setminus i}$.

Likelihood-Ratio Estimator: Recall that the likelihood-ratio estimator for a general class of distribution is formulated by the following equation.

$$\begin{aligned} &\frac{\partial}{\partial \mu_i}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x, \epsilon)) \\ &= \mathbb{E}_{\epsilon_i}(f(x,z) - b_i(x, \epsilon_{\setminus i}))\frac{\partial}{\partial \mu_i}\log q_{\phi_i}(z_i|\mathrm{pa}_i). \end{aligned} \tag{3}$$

Here we consider an *independent* baseline, i.e., $b_i = b_i(x, \epsilon_{\setminus i})$ is constant against ϵ_i. When q_{ϕ_i} is a Bernoulli distribution, the Monte Carlo estimate of Eq. (3) is written as follows using Eq. (1).

$$\begin{aligned} &(f(x,z) - b_i(x, \epsilon_{\setminus i}))\frac{\partial}{\partial \mu_i}\log q_{\phi_i}(z_i|\mathrm{pa}_i) \\ &= (f_{z_i} - b_i)\left(\frac{z_i}{\mu_i} - \frac{1-z_i}{1-\mu_i}\right) \\ &= \begin{cases} (f_1 - b_i)/\mu_i & \text{if } z_i = 1, \\ -(f_0 - b_i)/(1-\mu_i) & \text{if } z_i = 0. \end{cases} \end{aligned}$$

Since $z_i = 1$ holds with probability μ_i, the estimator is written as follows.

$$\Delta_i^{\mathrm{LR}} = \begin{cases} (f_1 - b_i)/\mu_i & \text{w.p. } \mu_i, \\ -(f_0 - b_i)/(1-\mu_i) & \text{w.p. } 1-\mu_i. \end{cases} \tag{4}$$

Optimal Estimator: The optimal estimator is given by the following formula.

$$\begin{aligned} &\frac{\partial}{\partial \mu_i}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x, \epsilon)) \\ &= \sum_{z_i} f(x,z)\frac{\partial}{\partial \mu_i} q_{\phi_i}(z_i|\mathrm{pa}_i)\Big|_{z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})}. \end{aligned} \tag{5}$$

By substituting Eq. (2), we obtain the following estimator.

$$\begin{aligned} \Delta_i^\star &= \sum_{z_i} f(x,z)\frac{\partial}{\partial \mu_i} q_{\phi_i}(z_i|\mathrm{pa}_i)\Big|_{z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})} \\ &= f_1 \times 1 + f_0 \times (-1) \\ &= f_1 - f_0. \end{aligned}$$

It can also be derived by simply calculating the mean of Eq. (4).

Local Expectation Gradient: The local expectation gradient estimator is given by the following expectation.

$$\begin{aligned} &\frac{\partial}{\partial \mu_i} F(\phi; x) \\ &= \mathbb{E}_{q_\phi(z_{\setminus i}|x)} \sum_{z_i} \frac{q_\phi(z_i|\mathrm{mb}_i)}{q_{\phi_i}(z_i|\mathrm{pa}_i)} f(x,z)\frac{\partial}{\partial \mu_i} q_{\phi_i}(z_i|\mathrm{pa}_i). \end{aligned} \tag{6}$$

Let $\pi_i = q_\phi(z_i = 1|\mathrm{mb}_i)$ and $f'_k = f(x, z_i = k, z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i = 1-k, \epsilon_{\setminus i}))$. Here f'_k is the value of f evaluated at $z_i = k \in \{0, 1\}$ and other variables $z_{\setminus i}$ computed

from $\epsilon_{\setminus i}$ and $z_i = 1 - k$. Note that the evaluation of the local expectation gradient estimator proceeds as follows:

1. Sample ϵ and compute $z = g_\phi(x, \epsilon)$.
2. Discard z_i so that we obtain $z_{\setminus i} \sim q_\phi(z_{\setminus i}|x)$.
3. Compute the summation in Eq. (6).

Therefore, if $z_i = 1$ is sampled at the first step, the value of f used in the estimation is f_1 and f'_0 instead of f_0, i.e., $z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})$ is not re-evaluated at $z_i = 0$. Similarly, if $z_i = 0$ is sampled at the first step, the value of f used in the estimation is f'_1 and f_0. Based on these observations, we derive the estimator using Eq. (2). If $z_i = 1$, then

$$\sum_k \frac{q_\phi(z_i = k|\mathrm{mb}_i)}{q_{\phi_i}(z_i = k|\mathrm{pa}_i)} f(x, z_i = k, z_{\setminus i}) \frac{\partial}{\partial \mu_i} q_{\phi_i}(z_i = k|\mathrm{pa}_i)$$
$$= \frac{\pi_i}{\mu_i} f_1 - \frac{1 - \pi_i}{1 - \mu_i} f'_0. \tag{7}$$

It is further transformed as follows.

$$\begin{aligned}
&\frac{\pi_i}{\mu_i} f_1 - \frac{1 - \pi_i}{1 - \mu_i} f'_0 \\
&= \frac{1}{\mu_i} \left(\pi_i f_1 - \frac{1 - \pi_i}{1 - \mu_i} \mu_i f'_0 \right) \\
&= \frac{1}{\mu_i} \left(f_1 - (1 - \pi_i) f_1 - \frac{1 - \pi_i}{1 - \mu_i} \mu_i f'_0 \right) \\
&= \frac{1}{\mu_i} \left(f_1 - \frac{1 - \pi_i}{1 - \mu_i} ((1 - \mu_i) f_1 + \mu_i f'_0) \right).
\end{aligned} \tag{8}$$

If $z_i = 0$, then

$$\sum_k \frac{q_\phi(z_i = k|\mathrm{mb}_i)}{q_{\phi_i}(z_i = k|\mathrm{pa}_i)} f(x, z_i = k, z_{\setminus i}) \frac{\partial}{\partial \mu_i} q_{\phi_i}(z_i = k|\mathrm{pa}_i)$$
$$= \frac{\pi_i}{\mu_i} f'_1 - \frac{1 - \pi_i}{1 - \mu_i} f_0. \tag{9}$$

It is further transformed similarly to Eq. (8) as follows.

$$\begin{aligned}
&\frac{\pi_i}{\mu_i} f'_1 - \frac{1 - \pi_i}{1 - \mu_i} f_0 \\
&= -\frac{1}{1 - \mu_i} \left((1 - \pi_i) f_0 - \frac{\pi_i}{\mu_i} (1 - \mu_i) f'_1 \right) \\
&= -\frac{1}{1 - \mu_i} \left(f_0 - \frac{\pi_i}{\mu_i} (\mu_i f_0 + (1 - \mu_i) f'_1) \right).
\end{aligned} \tag{10}$$

By combining Eq. (8) and Eq. (10), we obtain the following estimator.

$$\Delta_i^{\mathrm{LEG}} = \begin{cases} \frac{f_1 - \frac{1-\pi_i}{1-\mu_i}((1-\mu_i) f_1 + \mu_i f'_0)}{\mu_i} & \text{w.p. } \mu_i, \\ -\frac{f_0 - \frac{\pi_i}{\mu_i}((1-\mu_i) f'_1 + \mu_i f_0)}{1-\mu_i} & \text{w.p. } 1 - \mu_i. \end{cases}$$

It can be seen as a likelihood-ratio estimator with baseline given by

$$b_i^{\mathrm{LEG}} = \begin{cases} \frac{1-\pi}{1-\mu_i}((1-\mu_i) f_1 + \mu_i f'_0) & \text{if } z_i = 1, \\ \frac{\pi}{\mu_i}((1-\mu_i) f'_1 + \mu_i f_0) & \text{if } z_i = 0. \end{cases}$$

Note that this baseline might depend on z_i. Since the local expectation gradient is an unbiased estimator of the true gradient, the residual term of the likelihood-ratio gradient estimation C_i is kept 0.

B. Additional Results from the Experiments

The performance of each method on the training datasets is shown in Fig. 1 and Fig. 2. The trends are almost same as those of the validation scores.

The results on the test dataset are shown in Table 1. The log likelihood is approximated by the Monte Calo lower bound of Burda et al. (2015) with a sample size of 50,000.

C. Relationship between the RAM estimator and the Use of Monte Carlo Objectives

It has been shown that the use of Monte Carlo objectives (Burda et al., 2015; Mnih & Rezende, 2016) can improve the learning of generative models including variational autoencoders and sigmoid belief networks. The importance-weighted autoencoders (IWAE) (Burda et al., 2015) and VIMCO estimator (Mnih & Rezende, 2016) are gradient estimators for Monte Carlo objectives. Both of these evaluate the function f at multiple points, while the RAM estimator also evaluates multiple points. The use of multi-point evaluation is orthogonal between these methods and the RAM estimator. In the methods for Monte Carlo objectives, the multi-point evaluation comes from the improved lower bound of the log likelihood, i.e., they alter the objective function. On the other hand, in the RAM estimator, the multi-point evaluation is introduced purely for variance reduction. In particular, the RAM estimator is applicable to any kind of stochastic computational graphs. The experiments in this study is conducted to compare the variance of each gradient estimator for the same objective function, and therefore VIMCO estimator is omitted from the comparison.

References

Burda, Yuri, Grosse, Roger, and Salakhutdinov, Ruslan. Importance weighted autoencoders. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Mnih, Andriy and Rezende, Danilo Jimenez. Variational inference for monte carlo objectives. In *Proceedings of*

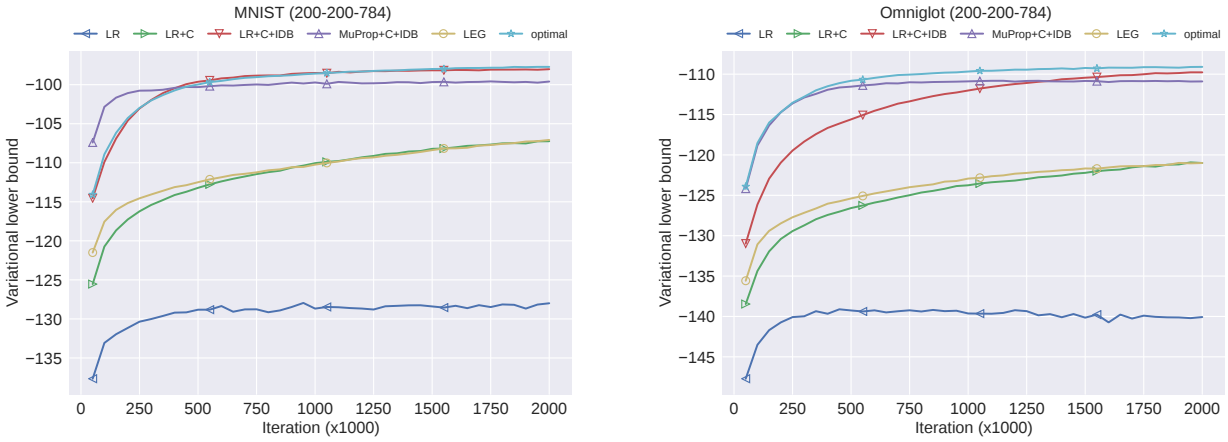

Figure 1. Training curves of two-layer SBN. Left: results using MNIST dataset. Right: results using Omniglot dataset.

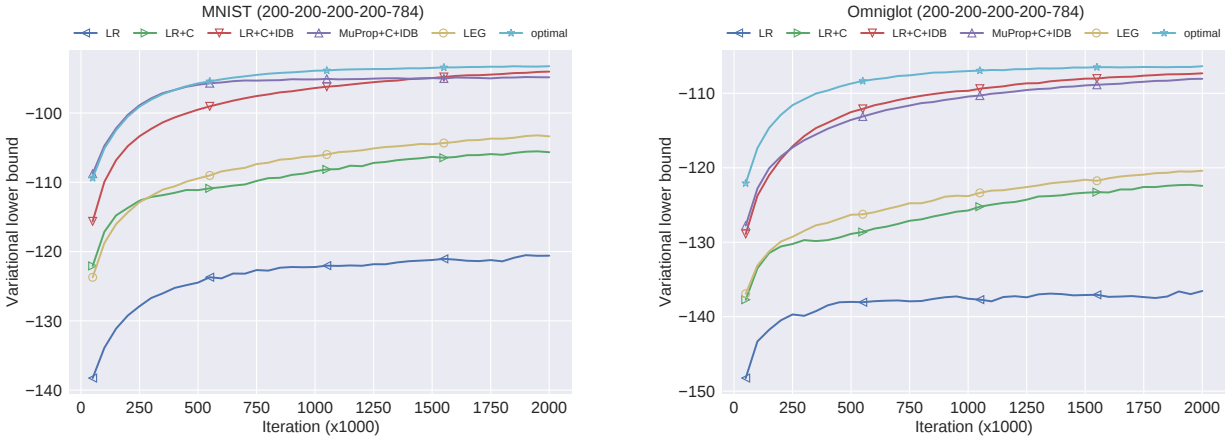

Figure 2. Training curves of four-layer SBN. Left: results using MNIST dataset. Right: results using Omniglot dataset.

the 33rd International Conference on Machine Learning, 2016.

Table 1. Test results of gradient estimators. The learning rate with the best validation performance is used in each method. VB stands for variational lower bound of the log likelihood, and LL stands for the log likelihood estimation.

	MNIST (shallow)		MNIST (deep)		Omniglot (shallow)		Omniglot (deep)	
	VB	LL	VB	LL	VB	LL	VB	LL
LR	-127.33	-108.53	-119.93	-103.53	-139.17	-124.08	-137.54	-122.85
LR+C	-107.21	-97.90	-105.38	-95.30	-122.10	-113.87	-123.27	-114.27
LR+C+IDB	-98.04	-92.68	-94.10	-89.02	-111.10	-107.14	-108.72	-105.00
MuProp+C+IDB	-99.96	-94.23	-95.03	-89.83	-112.97	-108.28	-109.55	-105.52
LEG	-106.75	-98.22	-103.26	-93.26	-121.68	-113.56	-121.27	-112.80
optimal	-97.64	-92.55	-93.31	-88.97	-110.60	-106.90	-108.17	-104.85