Evaluating the Variance of Likelihood-Ratio Gradient Estimators

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

Abstract

The likelihood-ratio method is often used to estimate gradients of stochastic computations, for which baselines are required to reduce the estimation variance. Many types of baselines have been proposed, although their degree of optimality is not well understood. In this study, we establish a novel framework of gradient estimation that includes most of the common gradient estimators as special cases. The framework gives a natural derivation of the optimal estimator that can be interpreted as a special case of the likelihood-ratio method so that we can evaluate the optimal degree of practical techniques with it. It bridges the likelihood-ratio method and the reparameterization trick while still supporting discrete variables. It is derived from the exchange property of the differentiation and integration. To be more specific, it is derived by the reparameterization trick and local marginalization analogous to the local expectation gradient. We evaluate various baselines and the optimal estimator for variational learning and show that the performance of the modern estimators is close to the optimal estimator.

1. Introduction

The success of deep learning owes much to efficient gradient computation using backpropagation (Rumelhart et al., 1986). When the model of interest includes internal stochasticities, the objective function is often written as a stochastic computational graph (Schulman et al., 2015). In this case, the exact gradient computation is intractable in general, and an approximate estimation is required. The variance introduced by the approximation often degrades the optimization performance for deep models, and therefore variance reduction is crucial for practical learning. However, few things are known about its theoretical aspects, and we often struggle with model-specific heuristics whose degree of optimality is difficult to know.

The likelihood-ratio method (Glynn, 1990; Williams, 1992) and the reparameterization trick (Williams, 1992; Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lzaro-gredilla, 2014) are widely used for the gradient estimation. The likelihood-ratio method only requires the computation of density functions and their derivatives, and therefore it is applicable to a wide range of models including those with discrete variables. It requires variance reduction techniques in practice. The most common technique is the use of a baseline value (Paisley et al., 2012; Bengio et al., 2013; Ranganath et al., 2014; Mnih & Gregor, 2014; Gu et al., 2016a) which is subtracted from a sampled objective value. The optimal baseline is difficult to compute in general, and we often use alternatives that are efficiently computed, some of which are based on model-specific heuristics. The reparameterization trick, on the other hand, has a small estimation variance in practice and is only applicable to models with certain continuous variables. Various models with continuous variables have been proposed using it, whereas less progress on the research of deep discrete variable models has been made because of the inapplicability of this method.

In this paper, we give a novel framework to formulate gradient estimators. It is derived by the reparameterization and the local marginalization analogous to the local expectation gradient (Titsias & Lázaro-Gredilla, 2015). The likelihood-ratio method and the reparameterization trick can be formalized under this framework, and therefore it bridges these two families of estimators. We can derive the optimal estimator, which gives a lower bound of the variance of all estimators covered by the framework. Since the estimator is derived by applying local marginalization to the reparameterized gradient, we named it the *reparameterization and marginalization (RAM) estimator*. This estimator is an instance of the likelihood-ratio estimator with the optimal baseline, and therefore it can be used to evaluate the variance of existing baseline techniques.

When the variable of interest follows a Bernoulli distribution, we can derive a tighter connection of a wider range of

[1]Preferred Networks, Tokyo, Japan [2]The University of Tokyo, Tokyo, Japan [3]RIKEN, Tokyo, Japan. Correspondence to: Seiya Tokui <tokui@preferred.jp>, Issei Sato <sato@k.u-tokyo.ac.jp>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

estimators to the framework. For example, the local expectation gradient (Titsias & Lázaro-Gredilla, 2015) becomes covered by our framework, and the straight-through estimator (Hinton, 2012; Bengio et al., 2013; Raiko et al., 2015) approximates the optimal estimator where the finite difference of the objective function is replaced by the infinitesimal first-order approximation. Furthermore, the optimal estimator is reduced to a likelihood-ratio estimator with an input-dependent baseline, which implies that a practical baseline technique might achieve a near-optimal variance.

The rest of this paper is organized as follows. We overview the related work in Sec.2 and formulate the gradient estimation problem in Sec.3. We introduce our framework in Sec.4 and derive important estimators with it in Sec.5. We also introduce a wider range of estimators for Bernoulli variables in Sec.6. We then show experimental results in Sec.7 and give a conclusion in Sec.8.

2. Related Work

The gradient estimation problem was being studied in the field of simulation around 1990, which is well summarized in L'Ecuyer (1991). The likelihood-ratio method (Glynn, 1989) is a general approach for solving the problem, in which the parametric density $q_\phi(z)$ is replaced by $\frac{q_\phi(z)}{q_0(z)} q_0(z)$ where $q_0 := q_\phi$ is fixed against ϕ on differentiation. The ratio $\frac{q_\phi(z)}{q_0(z)}$ is called the likelihood ratio, hence the name of this method. It can be seen as an importance sampling method that uses a proposal q_0 (Jie & Abbeel, 2010), with which there is a study on reducing the variance by using a proposal better than q_0 (Ruiz et al., 2016a). Another approach is the finite-difference method, in which the use of common random numbers, i.e., using the same random numbers to run two perturbed simulations, is effective in reducing the variance. The common random numbers naturally appear in the formulation of the optimal estimator of our framework.

The likelihood-ratio method has been combined with baselines and was introduced to the policy gradient methods for reinforcement learning, which is called the REINFORCE algorithm (Williams, 1992). The baseline technique is used for reducing the variance. A simple estimation of the average reward is commonly used as the baseline, and the optimal constant baseline that minimizes the variance is also derived (Weaver & Tao, 2001). The likelihood-ratio estimator has also been used for black-box variational inference (Ranganath et al., 2014). The likelihood-ratio estimator is used to derive the gradient estimation without depending on the specific form of the distributions. From a statistical point of view, the baseline can be seen as a special form of control variates, for which the optimal one can be derived again. The baseline technique has been further made sophisticated by involving the variable-wise baselines and those depending on the varaible of interest (Mnih & Gregor, 2014; Gu et al., 2016a). Some of them are also exported to policy-gradient methods (Gu et al., 2016b). Taking the local expectation of the likelihood-ratio estimator (Titsias & Lázaro-Gredilla, 2015) is another approach of variance reduction.

For variational inference of models with continuous variables, the reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lzaro-gredilla, 2014) is widely used. It is easy to implement with modern frameworks of automatic differentiation. It also has low variance in practice, although the superiority to the likelihood-ratio estimator is not guaranteed in general (Gal, 2017). This method is also applied to the continuous relaxation of discrete variables (Jang et al., 2016; Maddison et al., 2016).

On the one hand, the connection between the likelihood-ratio method and the reparameterization trick is studied in some literature, especially on continuous variables for which a tractable reparameterization is not available (Ruiz et al., 2016b). On the other hand, there are fewer studies for discrete variables. This paper provides a bridge between these estimators for discrete variables.

3. Problem Formulation

Our task is to optimize an expectation over a parameterized distribution. The objective function is given as $F(\phi; x) = \mathbb{E}_{q_\phi(z|x)} f(x, z)$, where f is a feasible function, $q_\phi(z|x) = \prod_{i=1}^{M} q_{\phi_i}(z_i|\mathrm{pa}_i)$ is a directed graphical model of M variables $z = (z_1, \ldots, z_M)$ conditioned on an input to the system x, pa_i are the parent nodes of z_i, and ϕ are the model parameters. Each conditional $q_{\phi_i}(z_i|\mathrm{pa}_i)$ is continuously differentiable w.r.t. ϕ_i and is typically a simple distribution such as a Bernoulli or Gaussian whose parameters are computed by a neural network with weights ϕ_i. For simplicity, we will assume that ϕ_i and $\phi_{i'}$ for $i \neq i'$ do not share any parameters; however, this assumption can be easily removed. We want to optimize F by stochastic gradient methods, which require an unbiased estimation of its gradient $\nabla_\phi F$.

A motivating example is variational learning of a generative model $p_\theta(x, z)$ with an approximate posterior $q_\phi(z|x)$. In this case, the objective function is the expectation of $f(x, z) = \log p_\theta(x, z) - \log q_\phi(z|x)$, which gives a lower bound of the log likelihood $\log p_\theta(x)$. On the one hand, the gradient w.r.t. the generative parameter θ is easily estimated by a Monte Carlo simulation. On the other hand, estimating the gradient w.r.t. ϕ is not trivial, which falls into the above general setting. Note that we omit the gradient incurred from the dependency of the second term $-\log q_\phi(z|x)$ on ϕ from our discussions since this gradi-

ent term is easy to estimate with low variance.

4. Proposed Framework

Here we give a general formulation of our framework of gradient estimation. The framework is based on the reparameterization of variables, which we also review.

Suppose that each sample drawn from a conditional is reparameterized as follows.

$$z_i \sim q_{\phi_i}(z_i|\mathrm{pa}_i) \Leftrightarrow z_i = g_{\phi_i}(\mathrm{pa}_i, \epsilon_i),\ \epsilon_i \sim p(\epsilon_i).$$

Here ϵ_i is a noise variable. We will give concrete examples of reparameterization later, and here we only emphasize that g_{ϕ_i} might be a non-continuous function. We write the whole reparameterization as $z = g_\phi(x, \epsilon)$.

Using this reparameterization, we derive the general form of gradient estimation. Let $\epsilon_{\setminus i} = \{\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_{i+1}, \ldots, \epsilon_M\}$. We partially exchange the differentiation and integration as follows.

$$\begin{aligned}\nabla_{\phi_i} F(\phi; x) &= \nabla_{\phi_i} \mathbb{E}_\epsilon f(x, g_\phi(x, \epsilon)) \\ &= \mathbb{E}_{\epsilon_{\setminus i}} \nabla_{\phi_i} \mathbb{E}_{\epsilon_i} f(x, g_\phi(x, \epsilon)). \end{aligned} \quad (1)$$

Unlike the reparameterization trick, this equation holds even if the function g_ϕ is not continuous because the local expectation $\mathbb{E}_{\epsilon_i} f(x, g_\phi(x, \epsilon))$ is differentiable. The technique of separating variables in leave-one-out manner is similar to Eq. (8) of Titsias & Lázaro-Gredilla (2015), whereas it is applied to reparameterized, mutually-independent noise variables in our case.

Equation (1) gives our framework of gradient estimation. Given a way to estimate the local gradient $\nabla_{\phi_i} \mathbb{E}_{\epsilon_i} f(x, g_\phi(x, \epsilon))$, we can estimate $\nabla_{\phi_i} F(\phi; x)$ by sampling $\epsilon_{\setminus i}$ and estimating the local gradient. Many existing estimators are derived by specifying a method of local gradient estimation, which we review in the next section.

Examples of Reparameterization: We introduce typical ways of reparameterizing popular distributions. When $q_{\phi_i}(z_i|\mathrm{pa}_i) = \mathcal{N}(z_i|\mu_i, \sigma_i^2)$ is a Gaussian distribution, z_i can be reparameterized as $z_i = \mu_i + \epsilon_i \sigma_i, \epsilon_i \sim \mathcal{N}(\epsilon_i; 0, 1)$ (Kingma & Welling, 2014). In this case, the change-of-variable formula $g_{\phi_i}(\mathrm{pa}_i, \epsilon_i) = \mu_i(\mathrm{pa}_i; \phi_i) + \epsilon_i \sigma_i(\mathrm{pa}_i; \phi_i)$ is differentiable w.r.t. ϕ_i. We can derive a reparameterization for Bernoulli variables as well. Suppose $z_i \in \{0, 1\}$ is a binary variable following a Bernoulli distribution $q_{\phi_i}(z_i|\mathrm{pa}_i) = \mu_i^{z_i}(1-\mu_i)^{1-z_i}$. It can be reparameterized using a uniform noise $\epsilon_i \sim U(0, 1)$ as $z_i = H(\mu_i - \epsilon_i)$, where $H(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \leq 0) \end{cases}$ is the Heaviside step function. In this case, the change-of-variable formula $z_i = H(\mu_i(\mathrm{pa}_i; \phi_i) - \epsilon_i)$ is not continuous in general. For categorical variables, we can use the Gumbel-Max trick (Gumbel, 1954; Jang et al., 2016; Maddison et al., 2016) for the reparameterization in a similar way.

5. Derivation of Gradient Estimators

We derive existing estimators on the basis of our general framework (1). We also derive the estimator that is optimal in terms of the estimation variance.

5.1. Likelihood-Ratio Estimator

The likelihood-ratio estimator is derived by using the log-derivative trick for the local gradient estimation. Let $b_i(x, \epsilon)$ be a baseline for z_i, and $\epsilon_{\setminus i} \sim p(\epsilon_{\setminus i})$. We use $z = g_\phi(x, \epsilon)$ and omit the dependency of z on ϵ. The likelihood-ratio estimator with baseline b_i is a Monte Carlo estimate of the following expectation.

$$\begin{aligned}&\nabla_{\phi_i} \mathbb{E}_{\epsilon_i} f(x, g_\phi(x, \epsilon)) \\ &= \mathbb{E}_{\epsilon_i}(f(x, z) - b_i(x, \epsilon)) \nabla_{\phi_i} \log q_{\phi_i}(z_i|\mathrm{pa}_i) + C_i(x, \epsilon_{\setminus i}).\end{aligned} \quad (2)$$

Here $C_i(x, \epsilon_{\setminus i}) = \mathbb{E}_{\epsilon_i} b_i(x, \epsilon) \nabla_{\phi_i} \log q_{\phi_i}(z_i|\mathrm{pa}_i)$, which has to be analytically computed.

There are many baseline techniques for variance reduction. We classify them into four categories as follows.

- *Constant baseline* is a constant of all variables $\{x, \epsilon\}$. It is a common choice for the baseline. In this case, it holds that $C_i = 0$. An exponential moving average of the simulated function f is often used.
- *Independent baseline* is a baseline that is constant against ϵ_i. It can depend on other variables, $\{x, \epsilon_{\setminus i}\}$. In this case, it again holds that $C_i = 0$. Two techniques proposed by Mnih & Gregor (2014) can be seen as examples of baselines in this class. One is the input-dependent baseline, which is a neural network that predicts the sampled objective value $f(x, z)$ from x and pa_i. The other one is the use of local signals, where the terms of f that are not descendants of z_i in the stochastic computational graphs are omitted. It can be seen as a baseline that includes all these terms.
- *Linear baseline* is a baseline that is a linear function of z_i, i.e., $b_i = z_i^\top u_i(x, \epsilon_{\setminus i}) + v_i(x, \epsilon_{\setminus i})$ [1], where u_i and v_i are arbitrary functions. In this case, we can write $C_i = (\nabla_{\phi_i} \mu_i)^\top u_i(x, \epsilon_{\setminus i})$, where $\mu_i = \mathbb{E}_{q_{\phi_i}(z_i|\mathrm{pa}_i)} z_i$ is the mean of z_i. The MuProp estimator (Gu et al.,

[1] When z_i is a binary or continuous scalar, the transposition is not needed. When z_i is a categorical variable, we represent it by a one-hot vector, for which $z_i^\top u_i$ is the innerproduct of two vectors.

2016a) is an example of estimators with linear baselines, where the baseline is given as the first-order approximation of the mean-field network of f at μ_i.

- *Fully-informed baseline* is a baseline that depends on all of x and ϵ, possibly in a nonlinear way. This is the most general class of baselines.

It is easily expected that the fully-informed baseline can achieve the lowest variance. We will show that the optimal estimator under the framework (1) falls into this category.

5.2. Reparameterization Trick Estimator

The reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014; Titsias & Lzaro-gredilla, 2014) is a common way to estimate the gradient for models with continuous variables. It is derived by exchanging the differentiation and integration of the local gradient as follows.

$$\nabla_{\phi_i}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x,\epsilon)) = \mathbb{E}_{\epsilon_i}\nabla_{\phi_i} f(x, g_\phi(x,\epsilon)). \quad (3)$$

Note that this equation holds only if the function $g_\phi(x,\epsilon)$ is differentiable, and therefore the reparameterization trick is only applicable to continuous variables.

The reparameterization trick often gives a better estimation of the gradient compared with the likelihood-ratio estimator, although there is no theoretical guarantee. Indeed, we can construct an example for which the likelihood-ratio estimator gives a better estimation (Gal, 2017).

5.3. Optimal Estimator

The optimal estimator is obtained by analytically computing the local gradient $\nabla_{\phi_i}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x,\epsilon))$. Let $z_{\setminus i} := \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_M\}$. When we fix $\epsilon_{\setminus i}$ and modify the value of z_i, the descendant variables of z_i might be changed because they are functions of z_i and noise variables. We denote the resulting values of $z_{\setminus i}$ by $z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})$. The function $h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})$ represents the ancestral sampling procedure of $z_{\setminus i}$ with given $\epsilon_{\setminus i}$ and clamped z_i. Using the reparameterization again, we obtain $\mathbb{E}_{\epsilon_i} f(x, g_\phi(x,\epsilon)) = \mathbb{E}_{q_{\phi_i}(z_i|\mathrm{pa}_i)} f(x,z)$. The local gradient is then computed as follows.

$$\nabla_{\phi_i}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x,\epsilon)) = \sum_{z_i} f(x,z)\nabla_{\phi_i} q_{\phi_i}(z_i|\mathrm{pa}_i)\Big|_{z_{\setminus i}=h_{\phi_{\setminus i}}(x,z_i,\epsilon_{\setminus i})}. \quad (4)$$

If z_i is continuous, the summation is replaced by an integral, which is approximated numerically. The resulting algorithm is shown in Alg. 1, which we name the *reparameterization and marginalization (RAM) estimator*. It requires M times evaluations of h, and therefore it scales worse than other estimators [2]. However, these evaluations are easily parallelized, and it runs fast enough for models of moderate size.

Algorithm 1 Algorithm for RAM estimator (4) for discrete z_i's. If z_i is continuous, the loop over all the configurations of z_i is replaced by a loop over integration points.

Require: a set of parameters ϕ and an input variable x.
1: Sample $\epsilon \sim p(\epsilon)$.
2: **for** $i = 1, \dots, M$ **do**
3: **for all** configurations of z_i **do**
4: $z_{\setminus i} := h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})$.
5: $f_{z_i} := f(x,z)\nabla_{\phi_i} q_{\phi_i}(z_i|\mathrm{pa}_i)$.
6: **end for**
7: $\Delta_i := \sum_{z_i} f_{z_i}$.
8: **end for**
9: **return** $(\Delta_1, \dots, \Delta_M)$ as an estimation of $\nabla F(\phi; x)$.

The optimality of this estimator is stated in the following theorem. Let ϕ_{ij} be the j-th element of the vector of parameters ϕ_i, and let $\partial_{ij} = \partial/\partial\phi_{ij}$ for notational simplicity.

Theorem 1. *Suppose an unbiased estimator δ_{ij} of the local derivative $\partial_{ij}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x,\epsilon))$, i.e., δ_{ij} is a random variable whose expectation matches the local derivative. Let V_{ij} be the variance of the estimator δ_{ij} and $V_{ij}^\star$ be the variance of the RAM estimator. Then, it holds that $V_{ij}^\star \le V_{ij}$.*

Proof. This follows from the standard Rao-Blackwellization argument. □

We briefly review the relationships between the RAM estimator and existing ones.

Relationship to the Likelihood-Ratio Estimator: The RAM estimator can be seen as an example of the likelihood-ratio estimators with fully-informed baselines. Let $b_i(x,\epsilon) = f(x, g_\phi(x,\epsilon))$. Then, the log-derivative term of Eq. (2) is canceled, and only the residual $C_i(x, \epsilon_{\setminus i}) = \nabla_{\phi_i}\mathbb{E}_{\epsilon_i} f(x, g_\phi(x,\epsilon))$ remains. The analytically-computed residual is equivalent to the RAM estimator. Since this estimator gives the minimum variance, our likelihood-ratio formulation (2) contains the optimal estimator.

While the fully-informed baseline is too general in practice to be efficiently computed, much more restrictive independent baselines can achieve the optimal estimator when z_i follows a Bernoulli distribution. Let $V_{ij}^{\mathrm{LR}}(b_i)$ be the variance of the likelihood-ratio estimator with baseline b_i.

Theorem 2. *Suppose that $q_{\phi_i}(z_i|\mathrm{pa}_i)$ is a Bernoulli distribution. Then, there is one and only one baseline $b_i^\star$ such*

[2] The difference in computational cost against the local expectation gradient (Titsias & Lázaro-Gredilla, 2015) comes from the inapplicability of pivot samples.

that $b_i^\star$ is constant against ϵ_i and $V_{ij}^\star = V_{ij}^{\mathrm{LR}}(b_i^\star)$.

The proof is given in Sec.6. This result implies that, for Bernoulli variables, the optimal variance might be obtained by a practical class of baseline techniques. Note that the optimal baseline $b_i^\star$ might depend on the noise variables corresponding to the descendants of z_i, which are not used by existing baseline techniques.

Relationship to the Reparameterization Trick: Theorem 1 also states that the RAM estimator gives a variance not larger than that of the reparameterization trick. Indeed, the RAM estimator is based on an analytical computation of the integral (3), and therefore gives the lower or equal variance. In practice, it is infeasible to compute the local gradient analytically, and numerical approximation is required. We can approximate the integral with high precision in practice, because z_i is usually a scalar variable and therefore we only need to evaluate the function $f(x, g_\phi(x, \epsilon))$ at a few integration points of z_i.

Relationship to the Local Expectation Gradient: The local expectation gradient (Titsias & Lázaro-Gredilla, 2015) is an application of local marginalization to the likelihood-ratio estimator. Let mb_i be the Markov blanket of z_i. This estimator is then derived as follows.

$$\begin{aligned}
&\nabla_{\phi_i} F(\phi; x) \\
&= \mathbb{E}_{q_\phi(z|x)} f(x,z) \nabla_{\phi_i} \log q_{\phi_i}(z_i|\mathrm{pa}_i) \\
&= \mathbb{E}_{q_\phi(z_{\setminus i}|x)} \mathbb{E}_{q_\phi(z_i|\mathrm{mb}_i)} f(x,z) \nabla_{\phi_i} \log q_{\phi_i}(z_i|\mathrm{pa}_i) \\
&= \mathbb{E}_{q_\phi(z_{\setminus i}|x)} \sum_{z_i} \frac{q_\phi(z_i|\mathrm{mb}_i)}{q_{\phi_i}(z_i|\mathrm{pa}_i)} f(x,z) \nabla_{\phi_i} q_{\phi_i}(z_i|\mathrm{pa}_i). \quad (5)
\end{aligned}$$

For the Monte Carlo simulation of $z_{\setminus i}$, z is first sampled from $q_\phi(z|x)$, and then z_i is discarded. It corresponds to sampling ϵ and computing $z_{\setminus i}$ by using it in the reparameterized notation. If the latent variables $z_1, \ldots, z_M$ are mutually independent given x, the density ratio factor $q_\phi(z_i|\mathrm{mb}_i)/q_{\phi_i}(z_i|\mathrm{pa}_i)$ equals 1, and therefore this estimator is equivalent to the RAM estimator (4). Otherwise, the density ratio factor remains, and these estimators do not match in general. The inference distribution $q_\phi(z_i|\mathrm{mb}_i)$ can be computed as

$$q_\phi(z_i|\mathrm{mb}_i) = \frac{q_\phi(z_i, \mathrm{mb}_i \setminus \mathrm{pa}_i|\mathrm{pa}_i)}{\sum_{z_i} q_\phi(z_i, \mathrm{mb}_i \setminus \mathrm{pa}_i|\mathrm{pa}_i)}.$$

Therefore, the density ratio is proportional to $q_\phi(z_i, \mathrm{mb}_i \setminus \mathrm{pa}_i|\mathrm{pa}_i)$. It tends to concentrate on z_i used in the sampling of $\mathrm{mb}_i \setminus \mathrm{pa}_i$, in which case the estimator degenerates to the plain likelihood-ratio estimator. Therefore, it cannot be guaranteed to have a lower variance than the likelihood-ratio estimator with baselines in general.

The RAM estimator can be seen as an application of the same technique to the reparameterized expectation (3). Thanks to the reparameterization, there is no need to solve the inference problem $q_\phi(z_i|\mathrm{mb}_i)$, and therefore the problematic density ratio factor does not appear. The evaluation of $f(x, z)$ with fixed $z_{\setminus i}$ does not reflect the full influence of the choice of z_i, whereas the reparameterized counterpart $f(x, z)|_{z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i})}$ does reflect it.

6. Analyzing Estimators for Binary Variables

A Bernoulli variable is the most fundamental example of a discrete variable, and some estimators are dedicated for it. It is beneficial to study the applications of any estimators to Bernoulli variables because they facilitate understanding and still contain most of the essential characteristics of discrete distributions. In some cases, an estimator has a connection to other estimators only when applied to Bernoulli variables. Here we focus on Bernoulli variables and introduce how each estimator can be formalized and related to others. The derivations are given in the supplementary material [3]

Suppose that $q_{\phi_i}(z_i|\mathrm{pa}_i)$ is a Bernoulli distribution of the mean parameter $\mu_i = \mu_i(\mathrm{pa}_i, \phi_i)$. Let $f_k = f(x, z_i = k, z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i, \epsilon_{\setminus i}))$ for $k \in \{0, 1\}$, i.e., f_k is the reparameterized objective value for $z_i = k$. All estimators we introduce here can be written as an estimation of the gradient w.r.t. μ_i multiplied by $\nabla_{\phi_i} \mu_i$, and therefore we only focus on the gradient w.r.t. μ_i denoted by Δ_i.

6.1. Likelihood-Ratio Estimator

The likelihood-ratio estimator for a Bernoulli variable with an independent baseline b_i is written as follows.

$$\Delta_i^{\mathrm{LR}} = \begin{cases} (f_1 - b_i)/\mu_i & \text{w.p. } \mu_i, \\ -(f_0 - b_i)/(1 - \mu_i) & \text{w.p. } 1 - \mu_i. \end{cases} \quad (6)$$

It can be interpreted as an importance sampling estimation of the sum of $f_1 - b_i$ and $-(f_0 - b_i)$. Indeed, the likelihood-ratio estimator for the general class of distributions can be seen as an importance sampling estimation of the expectation. When the distribution has low entropy (i.e., μ_i is close to 0 or 1), the variance of Δ_i^{LR} becomes large. However, it does not always mean that the variance of the gradient w.r.t. ϕ_i becomes huge, because in this case the sigmoid activation that outputs μ_i is in a flat regime so that its small derivative somehow alleviates the large variance. Neither does it mean that the variance is always small enough to optimize complex models.

[3] The supplementary material is attached to the arXiv version of this paper.

6.2. Optimal Estimator

The RAM estimator of the gradient w.r.t. μ_i is simply written as the difference of the f value at $z_i = 1$ and $z_i = 0$.

$$\Delta_i^\star = f_1 - f_0. \tag{7}$$

Using this formulation, we can prove Theorem 2.

Proof of Theorem 2. Let $b_i = (1-\mu_i)f_1 + \mu_i f_0$. Then, the likelihood-ratio estimator (6) with baseline b_i is equivalent to the RAM estimator (7). In this case, both cases of (6) are equal to $\Delta_i^\star$. This baseline does not depend on ϵ_i, and therefore we conclude the proof by letting $b_i^\star = b_i$. □

Interestingly, the optimal baseline is an expectation of f_k with the mean of k being $1-\mu$ instead of μ. This is different from the mean objective value $\bar{f} = \mu_i f_1 + (1-\mu_i)f_0$, which a constant mean baseline approximates. The difference becomes large when μ_i is close to 0 or 1.

The optimal estimator still has a positive variance since the noise variables $\epsilon_{\setminus i}$ are not integrated out. In Eq. (7), f_1 and f_0 are evaluated with the same configurations of these noise variables. The likelihood-ratio estimator can also be seen as an estimator that separately samples f_1 and f_0 at different iterations in which they use the separate samples of $\epsilon_{\setminus i}$. When the number of variables is large, the influence of one variable z_i on the objective value $f(x,z)$ is small, and we can expect that f_1 and f_0 have a positive covariance. In general, the estimation variance of the difference of two random variables X, Y is reduced by estimating them with a positive covariance since $\mathbb{V}[X-Y] = \mathbb{V}X + \mathbb{V}Y - 2\mathrm{Cov}(X,Y)$, and therefore our estimator effectively reduces the variance by using the same configuration of the noise variables. This technique is known as *common random numbers*, which is also used to reduce the variance of the gradient estimations with the finite difference method for stochastic systems (L'Ecuyer, 1991).

6.3. Local Expectation Gradient

The local expectation gradient (5) has a special view as a likelihood-ratio estimator when z_i is a Bernoulli variable. Let $\pi_i = q_\phi(z_i = 1|\mathrm{mb}_i)$. Let $f'_k = f(x, z_i = k, z_{\setminus i} = h_{\phi_{\setminus i}}(x, z_i = 1-k, \epsilon_{\setminus i}))$, i.e., f'_{1-k} is the objective value with fixed $z_{\setminus i}$ and flipped z_i. Note that f'_k can be different from f_k when some variables in $z_{\setminus i}$ depend on z_i. Then, the local expectation gradient is written as follows.

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

It is further rewritten as follows.

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

Thus, it can be seen as a likelihood-ratio estimator with baseline

$$b_i^{\mathrm{LEG}} = \frac{1 - q_\phi(z_i|\mathrm{mb}_i)}{1 - q_{\phi_i}(z_i|\mathrm{pa}_i)} b'_{ik}$$

where $b'_{ik} = (1 - q_{\phi_i}(z_i|\mathrm{pa}_i))f_k + q_{\phi_i}(z_i|\mathrm{pa}_i)f'_{1-k}$. The unweighted value b'_{ik} has a similar form as the optimal baseline $b_i^\star$, where f_{1-k} is replaced by f'_{1-k}. The final baseline b_i^{LEG} is given by multiplying b'_{ik} by the density ratio. We have seen that it tends to be close to 0, in which case the baseline is also close to 0 so that the estimator degenerates to the plain likelihood-ratio estimator. Even if the weight does not vanish, Theorem 2 shows that the local expectation gradient estimator for Bernoulli variables has higher variance than the optimal one unless all latent variables are mutually independent given x.

6.4. Straight-Through Estimator

We give one example of an estimator dedicated for Bernoulli variables, the straight-through estimator (Hinton, 2012; Bengio et al., 2013; Raiko et al., 2015). It is a biased estimator that leverages the gradient of f so that we can obtain the high-dimensional information of the direction towards which the objective would be decreasing. The estimator is written as follows.

$$\Delta_i^{\mathrm{ST}} = \begin{cases} \left.\frac{\partial f}{\partial z_i}\right|_{z_i=1} & \text{w.p. } \mu_i, \\ \left.\frac{\partial f}{\partial z_i}\right|_{z_i=0} & \text{w.p. } 1-\mu_i. \end{cases} \tag{8}$$

Observing that Eq. (7) gives the finite difference of f between $z_i = 1$ and $z_i = 0$, we can see that the straight-through estimator is its infinitesimal counterpart at the sampled z_i. Thus, this estimator is equivalent to our estimator (and is therefore unbiased) when f is a linear function of z_i. The difference between these estimators becomes large when the nonlinearity of f increases. If we consider a general class of the evaluation function f, we can construct an adversarial function f such that the derivative at $z_i \in \{0,1\}$ has the opposite sign against the finite difference (7). For example, when $f(x,z) = \sum_i z_i$, this estimator is equivalent to the optimal one. However, if we modify it to $\tilde{f}(x,z) = \sum_i z_i - \sin(2\pi \sum_i z_i)$, the straight-through estimator always gives a gradient opposite to the steepest direction of the expectation, which does not change as a result of the modification, i.e., $\mathbb{E}\tilde{f}(x,z) = \mathbb{E}f(x,z)$.

7. Experiments

We conduct experiments to empirically verify Theorem 1 and to demonstrate a procedure to analyze the optimal de-

gree of a given estimator covered by our framework. All the methods are implemented with Chainer (Tokui et al., 2015).

7.1. Experimental Settings

The task is variational learning of sigmoid belief networks (SBN) (Neal, 1992), which is a directed graphical model with layers of Bernoulli variables. Let $x \in \{0,1\}^d$ be a binary vector of input data and $Z_\ell = (z_{\ell,1}, \ldots, z_{\ell,N_\ell}) \in \{0,1\}^{N_\ell}$ be a binary vector of the latent variables at the ℓ-th layer. Denote the input layer as $Z_0 = x$ for notational simplicity. Let L be the number of latent layers. The generative model is specified by a conditional of each layer $p_\theta(Z_\ell|Z_{\ell+1})$ and the prior of the deepest layer $p_\theta(Z_L)$. In our experiments, the prior of each variable $z_{L,i} \in Z_L$ is independently parameterized by its logit. The conditional $p_\theta(Z_\ell|Z_{\ell+1})$ is modeled by an affine transformation of $Z_{\ell+1}$ that outputs the logit of Z_ℓ. The parameters of the affine transformation are optimized through the learning. We use two models with $L = 2$ and $L = 4$, respectively. Each layer consists of $N_\ell = 200$ Bernoulli variables.

We model the approximate posterior $q_\phi(z|x)$ by a reverse-directional SBN. In this case, the prior $q(x)$ is not modeled, and each conditional $q_\phi(Z_{\ell+1}|Z_\ell)$ is specified by its logit as an affine transformation of Z_ℓ.

Both the generative parameter θ and the variational parameter ϕ are optimized simultaneously to maximize the following variational lower bound.

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(z|x)} \frac{p_\theta(x,z)}{q_\phi(z|x)} \geq \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)} = \mathcal{L}.$$

The second line follows Jensen's inequality with the concavity of $\log$. The gradient w.r.t. θ is estimated by a Monte Carlo simulation of $\nabla_\theta \mathcal{L} = \mathbb{E}_{q_\phi(z|x)} \nabla_\theta \log p_\theta(x,z)$. We use gradient estimators for approximating the gradient w.r.t. ϕ.

The plain likelihood-ratio estimator is denoted by LR, whereas the constant baseline using the moving average of $f(x,z) = \log p_\theta(x,z) - \log q_\phi(z|x)$ and the input-dependent baseline of Mnih & Gregor (2014) are expressed by the postfixes +C and +IDB, respectively. We also run experiments for MuProp and the Local Expectation Gradient (LEG). Algorithm 1 is used to obtain the results for the optimal estimator.

We use MNIST (Lecun et al., 1998) and Omniglot (Lake et al., 2015) for our experiments. These are sets of 28x28 pixel gray-scale images of hand-written digits and hand-written characters from various languages. We binarize each pixel by sampling from a Bernoulli distribution with the mean equal to the pixel intensity (Salakhutdinov & Murray, 2008). The binarization is done in an online manner, i.e., we sample binarized vectors at each iteration. For the MNIST dataset, we use the standard split of 60,000 training images and 10,000 test images. The training images are further split into 50,000 images and 10,000 images, the latter of which are used for validation. For the Omniglot dataset, we use the standard split of 24,345 training images and 8,070 test images used in the official implementation of Burda et al. (2015) [4]. The training images are further split into 20,288 images and 4,057 images, the latter of which are used for validation.

We used RMSprop (Tieleman & Hinton, 2012) with a mini-batch size of 100 to optimize the variational lower bound. We apply a weight decay of the coefficient 0.001 for all parameters. All the weights are initialized with the method of Glorot & Bengio (2010). The learning rate is chosen from $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$. We evaluate the model on the validation set during training, and choose the learning rate with which the best validation performance with early-stopping beats the others. After each evaluation, we also measure the variance of the gradient estimations of variational parameters for the training set with the same mini-batch size.

Each experiment is done on an Intel(R) Xeon(R) CPU E5-2623 v3 at 3.00 GHz and an NVIDIA GeForce Titan X. Thanks to the parallel computation using the GPU, the computational time of the RAM estimator is only two times larger than the plain likelihood-ratio estimator.

7.2. Results

The results for the two-layer SBN and four-layer SBN are shown in Fig. 1 and Fig. 2, respectively. The results for these models have almost the same trends. As is predicted by Theorem 1, the optimal estimator gives the lower bound of the estimation variance. The plots imply that the modern baseline techniques effectively reduce the estimation variance, which is approaching the optimal value. However, the gap between these practical methods and the optimal one is not negligible, and there is still room for improvements. The local expectation gradient actually does not degenerate to the plain likelihood-ratio estimator, whereas the variance reduction effect is limited so that its variance stays at a similar level to that of the likelihood-ratio estimator with a constant baseline. The validation score almost agrees with the variance level, although there are some exceptions caused by the differences in selected learning rates.

Note that we do not align the computational cost by sampling multiple values in the experiments because the purpose of these experiments is evaluating the optimal degree

[4] https://github.com/yburda/iwae

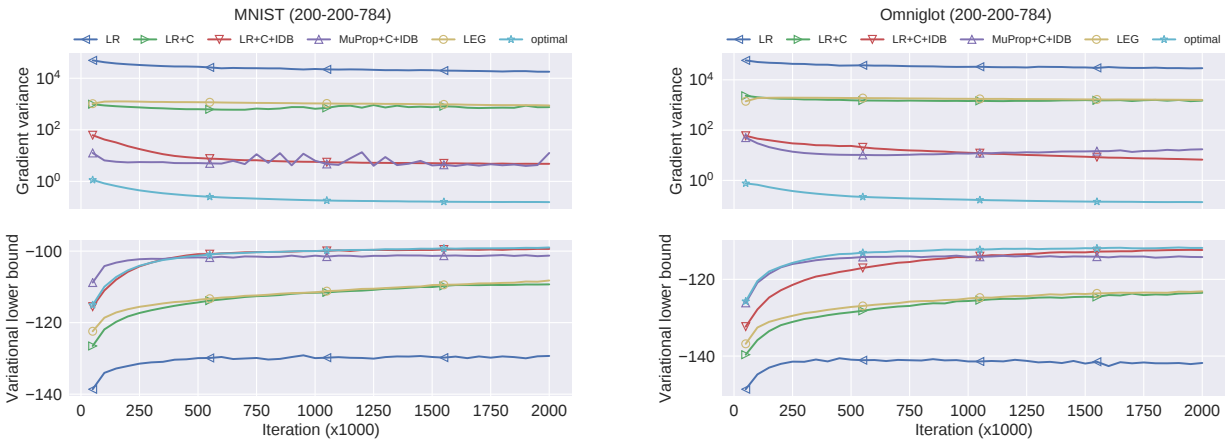

Figure 1. Results of two-layer SBN. Left: results using MNIST dataset. Right: results using Omniglot dataset. The mean of the gradient variances of the variational parameters are plotted in the top figures. The validation performance is plotted in the bottom figures.

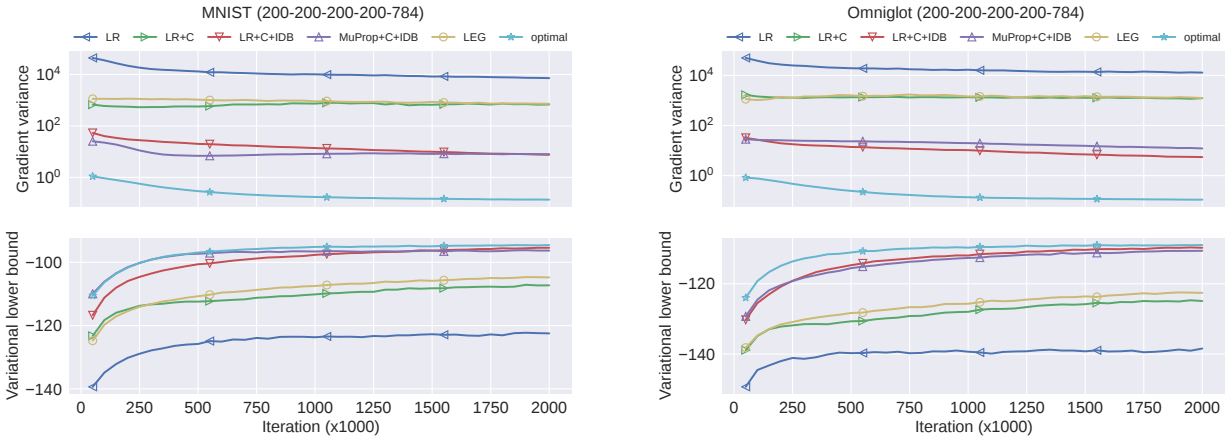

Figure 2. Results of four-layer SBN. Left: results using MNIST dataset. Right: results using Omniglot dataset.

of each method. We can infer the performance with aligned computational budget by comparing the variance and the computational cost.

8. Conclusion

We introduced a novel framework of gradient estimation for stochastic computations using reparameterization. The framework serves as a bridge between the likelihood-ratio method and the reparameterization trick. The optimal estimator is naturally derived under the framework. It provides the minimum variance attainable by the likelihood-ratio estimators with the general class of baselines, and therefore can be used to evaluate the optimal degree of each practical baseline technique. We actually evaluated the common baseline techniques against the optimal estimator for variational learning of sigmoid belief networks and showed that the modern techniques achieve a variance level close to the lower bound.

Comparison between continuous variable models and discrete variable models is needed for the further development of deep probabilistic modeling, which should consider the adequacy of the use of these variables in each task and the efficiency of gradient estimators available for these models. While this study does not provide a way to compare such models in general, it bridges the gradient estimators of them through the optimal case, and therefore provides some insights on their relationships. Observing the experimental results, the modern estimators for Bernoulli variables achieve variance close to the optimal one, and therefore we can expect that the modern estimators for Bernoulli variables are maturing and could be applied to much larger models capturing discrete phenomena.

Acknowledgments

We thank members of Preferred Networks, especially Daisuke Okanohara, for the helpful discussions.

References

Bengio, Yoshua, Léonard, Nicholas, and Courville, Aaron C. Estimating or propagating gradients through stochastic neurons for conditional computation. *ArXiv*, 1308.3432, 2013.

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

Gal, Yarin. *Uncertainty in Deep Learning*. PhD thesis, Department of Engineering, University of Cambridge, 2017.

Glorot, Xavier and Bengio, Yoshua. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10)*, 2010.

Glynn, P. W. Optimization of stochastic systems via simulation. In *Proceedings of the 21st Conference on Winter Simulation*, pp. 90–105, 1989. doi: 10.1145/76738.76750.

Glynn, Peter W. Likelihood ratio gradient estimation for stochastic systems. *Communication of the ACM*, 33(10): 75–84, 1990. doi: 10.1145/84537.84552.

Gu, Shixiang, Levine, Sergey, Sutskever, Ilya, and Mnih, Andriy. Muprop: Unbiased backpropagation for stochastic neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016a.

Gu, Shixiang, Lillicrap, Timothy, Ghahramani, Zoubin, Turner, Richard E., and Levine, Sergey. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *NIPS 2016 Deep Reinforcement Learning Workshop*, 2016b.

Gumbel, Emil Julius. *Statistical theory of extreme values and some practical applications*. U. S. Govt. Print. Office, 1954.

Hinton, Geoffrey. Neural networks for machine learning. Coursera, video lectures, 2012.

Jang, E., Gu, S., and Poole, B. Categorical Reparameterization with Gumbel-Softmax. *ArXiv e-prints*, 2016. To appear in ICLR 2017.

Jie, Tang and Abbeel, Pieter. On a connection between importance sampling and the likelihood ratio policy gradient. In *Advances in Neural Information Processing Systems 23*, pp. 1000–1008. 2010.

Kingma, Diederik P. and Welling, Max. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Lake, Brenden M., Salakhutdinov, Ruslan, and Tenenbaum, Joshua B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266): 1332–1338, 2015. doi: 10.1126/science.aab3050.

Lecun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

L'Ecuyer, Pierre. An overview of derivative estimation. In *Proceedings of the 23rd Conference on Winter Simulation*, pp. 207–217, 1991. ISBN 0-7803-0181-1.

Maddison, Chris J., Mnih, Andriy, and Teh, Yee Whye. The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712, 2016. To appear in ICLR 2017.

Mnih, Andriy and Gregor, Karol. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 1791–1799, 2014.

Neal, Radford M. Connectionist learning of belief networks. *Artificial Intelligence*, 56(1):71–113, 1992.

Paisley, John, Blei, David M., and Jordan, Michael I. Variational bayesian inference with stochastic search. In *Proceedings of the 29 th International Conference on Machine Learning (ICML)*, 2012.

Raiko, Tapani, Berglund, Mathias, Alain, Guillaume, and Dinh, Laurent. Techniques for learning binary stochastic feedforward neural networks. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David M. Black box variational inference. In *Artificial Intelligence and Statistics (AISTATS)*, pp. 814–822, 2014.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pp. 1278–1286, 2014.

Ruiz, F. J. R., Titsias, M. K., and Blei, D. M. Overdispersed black-box variational inference. In *Uncertainty in Artificial Intelligence (UAI)*, 2016a.

Ruiz, F. J. R., Titsias, M. K., and Blei, D. M. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, 2016b.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pp. 696–699. MIT Press, 1986.

Salakhutdinov, Ruslan and Murray, Iain. On the quantitative analysis of Deep Belief Networks. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML)*, pp. 872–879, 2008.

Schulman, John, Heess, Nicolas, Weber, Theophane, and Abbeel, Pieter. Gradient estimation using stochastic computation graphs. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 3528–3536, 2015.

Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. In CORSERA: Neural Networks for Machine Learning, 2012.

Titsias, Michalis and Lázaro-Gredilla, Miguel. Local expectation gradients for black box variational inference. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 2638–2646. 2015.

Titsias, Michalis and Lzaro-gredilla, Miguel. Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1971–1979, 2014.

Tokui, Seiya, Oono, Kenta, Hido, Shohei, and Clayton, Justin. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015. URL http://learningsys.org/papers/LearningSys_2015_paper_33.pdf.

Weaver, Lex and Tao, Nigel. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 538–545, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1. URL http://dl.acm.org/citation.cfm?id=2074022.2074088.

Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.