5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–8

A Dual Control Variate for accelerated black-box variational inference

Xi Wang XWANG3@CS.UMASS.EDU
Tomas Geffner TGEFFNER@CS.UMASS.EDU
Justin Domke DOMKE@CS.UMASS.EDU
University of Massachusetts, Amherst

Abstract

In this paper, we aim at reducing the variance of doubly stochastic optimization, a type of stochastic optimization algorithm that contains two independent sources of randomness: The subsampling of training data and the Monte Carlo estimation of expectations. Such an optimization regime often has the issue of large gradient variance which would lead to a slow rate of convergence. Therefore we propose Dual Control Variate, a new type of control variate capable of reducing gradient variance from both sources jointly. The dual control variate is built upon approximation-based control variates and incremental gradient methods. We show that on black-box variational inference, which can be formulated as a doubly stochastic optimization problem, compared with past variance reduction approaches that take only one source of randomness into account, dual control variate leads to a gradient estimator of significantly smaller variance and demonstrates significantly faster convergence[1].

1. Introduction

Various machine learning problems can be formulated as optimizing an objective of the form

$$f(w) = \mathop{\mathbb{E}}_{\mathsf{n}} \mathop{\mathbb{E}}_{\epsilon} f(w; \mathsf{n}, \epsilon). \tag{1}$$

Here, n is a discrete random variable uniformly distributed on $\{1, \ldots, N\}$, which typically represents an index in a dataset. Meanwhile, ϵ is a continuous random variable drawn from some fixed distribution, independent of w and n. Objectives like this emerge in black box variational inference (Paisley et al., 2012; Ranganath et al., 2014; Titsias and Lázaro-Gredilla, 2014) with reparameterization gradient and variational autoencoders (Kingma and Welling, 2014; Rezende et al., 2014) (where ϵ corresponds to a sample from the latent space) and models that apply data augmentation or dropout during training (Srivastava et al., 2014) (where ϵ corresponds to the random perturbation of the data). Such objectives are typically addressed with stochastic optimization. The most obvious gradient estimator is given by drawing a random n and a random ϵ and evaluating

$$g_{\text{naive}}(w; n, \epsilon) = \nabla f(w; n, \epsilon). \tag{2}$$

This estimator is adequate for many situations but sometimes displays high variance, which slows optimization (Nemirovski et al., 2009; Bottou et al., 2018), as is shown by the blue

1. Full version at https://arxiv.org/abs/2210.07290

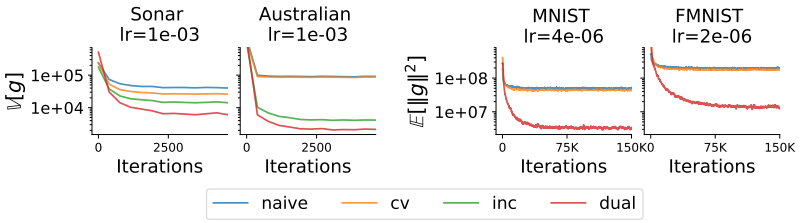


Figure 1: **On Bayesian logistic regression, simultaneously controlling subsampling and Monte Carlo noise significantly reduces gradient variance.** The naive gradient estimator (Eq. (2)) is the baseline, while the cv estimator (Eq. (4)) controls for Monte Carlo noise, the inc estimator (Eq. (7)) controls for subsampling noise, and the proposed dual estimator (Eq. (8)) controls for both. (The inc estimator is shown only for reference on smaller datasets since it is not practical.) A reduced variance allows for larger learning rates and faster optimization (Fig. 2)

lines in Fig. 1. In particular, when N is large, performing data subsampling could introduce a significant amount of gradient noise. When the variance is large, the step-size must be made very small, slowing optimization. For black-box variational inference, several recent works have thus been devoted to reducing the variance of this reparameterization estimator (Miller et al., 2017; Buchholz et al., 2018; Roeder et al., 2017; Wu et al., 2019; Geffner and Domke, 2018, 2020; Boustati et al., 2020). Many methods have been developed to reduce the variance of related objectives. Problems without subsampling correspond to an objective $f(w) = \mathbb{E}_\epsilon f(w; \epsilon)$. Many works have proposed to use *control variate* to control the gradient variance (Sec. 2.1). Other problems that *only* have subsampling can be written as $f(w) = \mathbb{E}_\mathsf{n} f(w; \mathsf{n})$. This is the *incremental gradient* setting, for which many methods have been developed, e.g. SVRG (Johnson and Zhang, 2013) and SAGA (Defazio et al., 2014a) (Sec. 2.2). Unfortunately, these methods only address variance coming from a single source of randomness, meaning there are limitations on what they can accomplish when applied to doubly-stochastic problems.

In this work we propose a novel *dual control variate* (Sec. 3) that reduces the two types of gradient variance *at the same time*. We empirically evaluate the effectiveness of the dual control variate on black box variational inference (Sec. 4). We show that the dual control variate yields a gradient estimator with variance an order of magnitude smaller than the naive estimator or an approximation-based control variate. It is also superior to a baseline control variate based on incremental gradient ideas. (This baseline is too expensive to be practical in general.) This improvement in variance enables the use of larger learning rates, and thus yields a corresponding order of magnitude increase in optimization speed.

2. Background

2.1. Approximation-based control variate

Assume we have an objective of the form $f(w) = \mathbb{E}_\epsilon f(w; \epsilon)$, where ϵ is a random variable drawn from a fixed distribution independent of w. Computing the exact gradient is often

intractable. A naive gradient estimator would be $\nabla f(w;\epsilon)$. The variance of this can often be reduced by instead using

$$g(w;\epsilon) = \nabla f(w;\epsilon) + c(w;\epsilon), \tag{3}$$

where $c(w;\epsilon)$ is a control variate, i.e. a zero-mean random variable. A general way to construct control variates involves using an approximation function $\tilde{f} \approx f$ for which the expectation $\mathbb{E}_\eta \tilde{f}(w,\eta)$ is available in closed-form (Miller et al., 2017; Geffner and Domke, 2020). Then, the control variate is defined as $c(w;\epsilon) = \mathbb{E}_\eta \nabla\tilde{f}(w;\eta) - \nabla\tilde{f}(w;\epsilon)$, which can be easily seen to have a mean zero. Approximation-based control variates can be applied to the doubly-stochastic objective using an approximation $\tilde{f}(w;n,\epsilon) \approx f(w;n,\epsilon)$ with tractable expectation with respect to ϵ. Then, using this approximation, we can define the unbiased gradient estimator

$$g_{\text{cv}}(w;n,\epsilon) = \nabla f(w;n,\epsilon) + \underbrace{\mathop{\mathbb{E}}_{\eta} \nabla\tilde{f}(w;n,\eta) - \nabla\tilde{f}(w;n,\epsilon)}_{c_{\text{cv}}(w;n,\epsilon)}, \tag{4}$$

where the second term defines a control variate. We call this the cv estimator. Notice that the cv estimator only controls the Monte Carlo noise but cannot reduce subsampling noise.

2.2. Incremental gradient methods

We now consider a stochastic optimization problem with only subsampling noise, whose objective is given by $f(w) = \mathbb{E}_{\mathsf{n}}\, f(w;\mathsf{n})$, where n is a random variable uniformly distributed on $\{1,\dots,N\}$. While one could compute f exactly, this may be costly when N is large. The naive gradient estimator $\nabla f(w;n)$ where n is randomly chosen. Incremental gradient methods (Roux et al., 2012; Shalev-Shwartz and Zhang, 2013; Johnson and Zhang, 2013; Defazio et al., 2014b; Gower et al., 2020) were developed to reduce the variance of this gradient estimator. While details vary by algorithm, the basic idea is to "recycle" previous evaluations. For example, SAGA (Defazio et al., 2014a) stores the gradients $\nabla f(w^n;n)$ where w^n is w at the most recent time $f(w;n)$ was evaluated. Then, a gradient step is taken as

$$w \leftarrow w - \lambda\left(\nabla f(w;n) + \mathop{\mathbb{E}}_{\mathsf{m}} \nabla f(w^{\mathsf{m}};\mathsf{m}) - \nabla f(w^n;n)\right), \tag{5}$$

where λ is a step size. When $w^m \approx m$, the first and last terms in Eq. (5) will approximately cancel. The final expectation over m is tracked as a running average of $\nabla f(w^m;m)$, meaning the cost per iteration is independent of N. The update rule above can also be rewritten as a gradient estimator composed of a naive gradient estimator plus a control variate

$$g_{\text{inc}}(w;n) = \nabla f(w;n) + \underbrace{\mathop{\mathbb{E}}_{\mathsf{m}} \nabla f(w^{\mathsf{m}};\mathsf{m}) - \nabla f(w^n;n)}_{c_{\text{inc}}(w;n)}. \tag{6}$$

This method, however, cannot be easily adapted to the doubly-stochastic setting. To see this, consider the following inc estimator

$$g_{\text{inc}}(w;n,\epsilon) = \nabla f_n(w;n,\epsilon) + \underbrace{\mathop{\mathbb{E}}_{\mathsf{m}} \nabla f(w^{\mathsf{m}};\mathsf{m},\epsilon) - \nabla f(w^n;n,\epsilon)}_{c_{\text{inc}}(w;n,\epsilon)}. \tag{7}$$

Algorithm 1 Stochastic gradient descent with the dual control variate.

Require: Learning rate λ.

Initialize the parameter w_0, the parameter table $W = \{w^1, \ldots, w^N\}$ and the running mean $M = \mathbb{E}_{\mathsf{m}} \mathbb{E}_{\eta} \nabla \tilde{f}(w_0; \mathsf{m}, \eta)$.

for $k = 1, 2, \cdots$ **do**

Sample n and ϵ.

Extract the value of w^n from the table W.

Compute the base gradient $g \leftarrow \nabla f(w_k; n, \epsilon)$.

Compute the control variate $c \leftarrow M - \nabla \tilde{f}(w^{\mathsf{n}}; n, \epsilon)$. (Uses that $M = \mathbb{E}_{\mathsf{m}} \mathbb{E}_{\eta} \nabla \tilde{f}(w^{\mathsf{m}}; \mathsf{m}, \eta)$.)

Update the running mean $M \leftarrow M + \frac{1}{N}\left(\mathbb{E}_{\eta} \nabla \tilde{f}(w_k; n, \eta) - \mathbb{E}_{\eta} \nabla \tilde{f}(w^{\mathsf{n}}; n, \eta)\right)$

Update the table $w^n \leftarrow w_k$ and update the parameter $w_{k+1} \leftarrow w_k - \lambda(g + c)$.

end for

The first drawback for g_{inc} is that, similar to g_{cv}, g_{inc} also possesses limitations in how much it can reduce variance: It only controls the subsampling noise from n while retaining the irreducible variance from ϵ. However, a more critical issue of this estimator is that it is not *practical* to be applied. This is because the value of $\nabla f(w^{\mathsf{n}}; n, \epsilon)$ is dependent on ϵ, which is resampled at each iteration. This means that it is not possible to efficiently maintain $\mathbb{E}_{\mathsf{m}} \nabla f(w^{\mathsf{m}}; \mathsf{m}, \epsilon)$, and so each gradient estimate, it requires a full pass over the dataset. Of course, doing this would be pointless—it would be better to simply explicitly sum out n and solve an optimization problem that only has randomness due to ϵ. Nevertheless, g_{inc} serves as an important point of reference. Surprisingly, when we introduce our dual control variate below, this computational issue disappears. Thus, we will compare to g_{inc} when possible to give more insight into which source of variance is more important. Unlike g_{cv}, g_{inc}, the variance of g_{dual} can in principle be arbitrarily small. The variance is only limited by how close $\tilde{f}$ is to f and how close the previously evaluated values w^n are to w.

3. Proposed method: Dual control varaite

We now introduce a new approach for controlling the variance of gradient estimators for doubly stochastic optimization problems, the central contribution of this paper. The idea is to introduce an approximation $\tilde{f} \approx f$ and take an expectation over ϵ as with approximation-based control variates, but to also recycle past evaluations at different values of n as in incremental gradient methods. We propose the estimator

$$g_{\text{dual}}(w; n, \epsilon) = \nabla f(w; n, \epsilon) + \underbrace{\mathop{\mathbb{E}}_{\mathsf{m}} \mathop{\mathbb{E}}_{\eta} \nabla \tilde{f}(w^{\mathsf{m}}; \mathsf{m}, \eta) - \nabla \tilde{f}(w^n; n, \epsilon)}_{c_{\text{dual}}(w; n, \epsilon)}. \tag{8}$$

Note that g_{dual} does not suffer from the same computational issue as g_{inc}. The approximation $\tilde{f}$ is designed so that $\mathbb{E}_{\eta} \nabla \tilde{f}(w^n; n, \eta)$ can be computed in closed-form. By caching these values, the expectation over m can be computed using a running average, rather than iterating though all possible values of n. The full algorithm is presented in Algorithm. 1.

4. Experiments

In this section, we empirically evaluate the proposed dual control variate on black box variational inference with mean-field Gaussian as the variational posterior, in which we approximate the true posterior distribution with a variational posterior of form $q_w(\mathsf{z}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}))$, and we find the optimal $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ by *maximizing* the evidence lower bound (ELBO), which is equivalent to minimizing the following objective

$$f(w) = -\mathop{\mathbb{E}}_{\mathsf{n}} \mathop{\mathbb{E}}_{q_w(\mathsf{z})} \Big[N \log p(x_{\mathsf{n}} \mid \mathsf{z}) + \log p(\mathsf{z}) \Big] - \mathbb{H}(w), \tag{9}$$

where we would have $w = (\boldsymbol{\mu}, \boldsymbol{\sigma})$. In order to estimate the objective's gradient with respect to w via Monte Carlo sampling, one typically has to apply the reparameterization trick (Kingma and Welling, 2014; Rezende et al., 2014; Titsias and Lázaro-Gredilla, 2014) on z, which aims to represent the sampling routine $\mathsf{z} \sim q_w(\mathsf{z})$ as a deterministic and differentiable function $\mathsf{z} = \mathcal{T}_w(\epsilon)$ of a w-independent base random variable ϵ. With the reparameterized form, we can rewrite the function inside the nested expectation as $k_n(\mathcal{T}_w(\epsilon)) = N \log p(x_n \mid \mathcal{T}_w(\epsilon)) + \log p(\mathcal{T}_w(\epsilon))$. Then, we can rewrite the objective as

$$f(w) = -\mathop{\mathbb{E}}_{\mathsf{n}} \mathop{\mathbb{E}}_{\epsilon} f(w; \mathsf{n}, \epsilon),$$

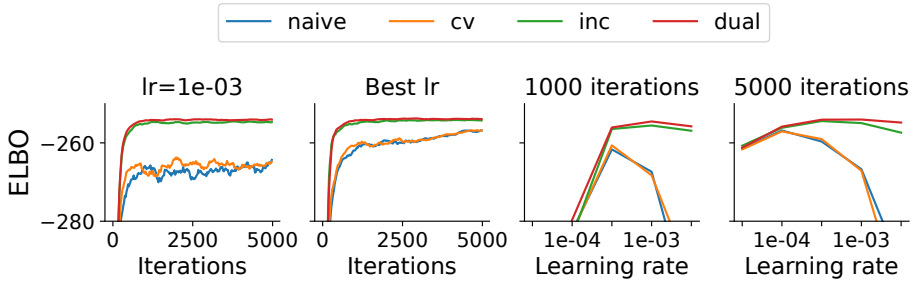
where we define $f(w; n, \epsilon) = k_n(\mathcal{T}_w(\epsilon)) + \mathbb{H}(w)$ and $\mathbb{H}(w)$ denotes the entropy of q_w. Inspired by previous work (Miller et al., 2017), we propose to get an approximation for $f(w; n, \epsilon)$ using a second order Taylor expansion for $k_n(\cdot)$ around $z_0 = \mathcal{T}_w(0)$, which yields

$$\tilde{f}(w; n, \epsilon) = k_n(z_0) + (\mathcal{T}_w(\epsilon) - z_0)^\top \nabla k_n(z_0) + \frac{1}{2}(\mathcal{T}_w(\epsilon) - z_0)^\top \nabla^2 k_n(z_0)(\mathcal{T}_w(\epsilon) - z_0)^\top + \mathbb{H}(w), \tag{10}$$

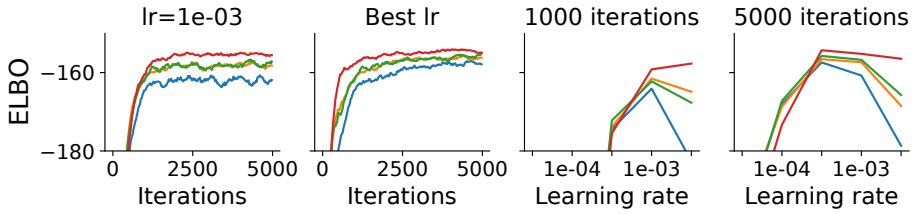
in which we assume the entropy can be computed in closed-form.

Results We evaluated our methods on Bayesian logistic regression with standard Gaussian prior. We first experiment with two small-scale dataset: the Australian dataset and the sonar dataset, where g_{inc} can be computed tractably as a baseline. We estimate gradients using a minibatch size of 5 all using a single sample of ϵ. We optimize using stochastic gradient descent (without momentum) with learning rates ranging from 10^{-5} to 5×10^{-3}. The results are presented in the left two columns in Fig. 1, Figs. 2(a) and 2(b). Note that both the inc and cv estimators have lower variance on the naive estimator, but this varies by the dataset. The excellent performance of the inc estimator on Australian shows the importance of reducing subsampling noise.

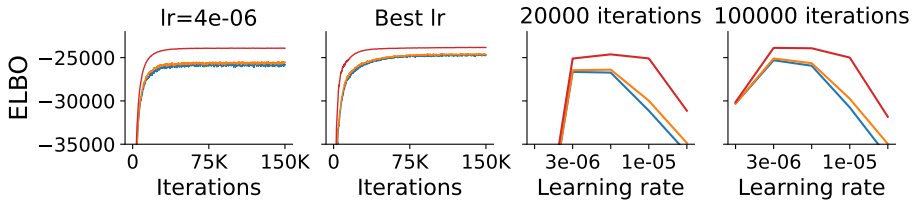
Next, we experiment with two larger-scale datasets: MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao et al., 2017) (FMNIST). We estimate the gradients using a mini-batch of 100 samples with one sample of ϵ. Because of the scale of the datasets, we are forced to use the expected gradient norm as a proxy for the gradient variance. We experiment with learning rates ranging from 4e−8 to 1.5e−5 and 2e−8 to 7.5e−6 for MNIST and FMNIST respectively. Results are shown in the right two columns in Fig. 1, Fig. 2(c) and Fig. 2(d). Here g_{dual} shows a much lower variance than g_{naive} or g_{cv}, which again emphasizes the importance of controlling subsampling noise in BBVI. (While g_{inc} is too expensive to run, we conjecture that it would perform well.)



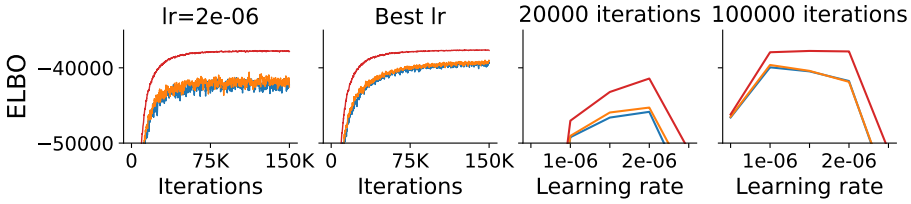
(*a*) Australian



(*b*) Sonar



(*c*) MNIST



(*d*) FMNIST

Figure 2: **For mean-field black-box variational inference, the dual estimator leads to improved convergence at higher learning rates.** This improvement is a consequence of lower variance (Fig. 1). Australian and Sonar are small datasets, included because it is possible to compute the inc estimator using brute force, which gives some insight into where the improvement in the dual estimator comes from. MNIST and FMNIST are larger-scale problems where the inc estimator is intractable. Given the small improvement of the cv estimator over the naive estimator on these problems, we suspect that most of the improvement in the dual estimator comes from reduced subsampling variance. All the figures are based on 10 trials of different random seeds.

References

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Ayman Boustati, Sattar Vakili, James Hensman, and ST John. Amortized variance reduction for doubly stochastic objective. In *Conference on Uncertainty in Artificial Intelligence*, pages 61–70. PMLR, 2020.

Alexander Buchholz, Florian Wenzel, and Stephan Mandt. Quasi-Monte Carlo variational inference. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 668–677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/buchholz18a.html.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014a.

Aaron Defazio, Justin Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133. PMLR, 2014b.

Tomas Geffner and Justin Domke. Using large ensembles of control variates for variational inference. In *Advances in Neural Information Processing Systems*, pages 9982–9992, 2018.

Tomas Geffner and Justin Domke. Approximation based variance reduction for reparameterization gradients. *Advances in Neural Information Processing Systems*, 33, 2020.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

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

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Andrew C Miller, Nicholas J Foti, Alexander D'Amour, and Ryan P Adams. Reducing reparameterization gradient variance. *Advances in Neural Information Processing Systems*, 2017:3709–3719, 2017.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

John Paisley, David M Blei, and Michael I Jordan. Variational bayesian inference with stochastic search. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1363–1370, 2012.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

Geoffrey Roeder, Yuhuai Wu, and David K Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/e91068fff3d7fa1594dfdf3b4308433a-Paper.pdf.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(2), 2013.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International conference on machine learning*, pages 1971–1979. PMLR, 2014.

Mike Wu, Noah Goodman, and Stefano Ermon. Differentiable antithetic sampling for variance reduction in stochastic variational inference. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2877–2886. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/wu19c.html.

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