Variance Reduction in Black-box Variational Inference by Adaptive Importance Sampling

Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang*
College of Computer Science and Technology, Jilin University, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
liximing86@gmail.com

Abstract

Overdispersed black-box variational inference employs importance sampling to reduce the variance of the Monte Carlo gradient in black-box variational inference. A simple overdispersed proposal distribution is used. This paper aims to investigate how to adaptively obtain better proposal distribution for lower variance. To this end, we directly approximate the optimal proposal in theory using a Monte Carlo moment matching step at each variational iteration. We call this adaptive proposal moment matching proposal (MMP). Experimental results on two Bayesian models show that the MMP can effectively reduce variance in black-box learning, and perform better than baseline inference algorithms.

1 Introduction

Variational inference (VI) [Jordan, 1999; Wainwright and Jordan, 2008] is an effective approximating inference algorithm for intractable posterior distributions of probabilistic models, an alternative to Markov chain Monte Carlo (MCMC). In contrast to MCMC, VI tends to be faster, and it has been applied to many popular Bayesian models, such as factorial models [Ghahramani and Jordan, 1997] and topic models [Blei *et al.*, 2003; Blei and Lafferty, 2007].

The basic idea of VI is that it first defines a family of variational distributions, and then transforms the posterior inference into an optimization with respect to the variational parameters. The commonly used objective is the Kullback-Leibler divergence between the variational distribution and the true posterior.

In some settings (e.g., conditionally conjugate models), VI can be easily solved by coordinate ascent, where one can derive the updating equations of variational parameters analytically. However, for many complex models the variational optimization becomes difficult, and we have to design model-specific updating strategies, e.g., numerical quadrature [Honkela and Valpola, 2004], alternative bounds [Blei and Lafferty, 2007], and Laplace approximation [Wang and Blei, 2013]. Such model-specific derivations make variational computations expensive, especially for the users that have no acquaintance with variational methods.

To make VI more practical, the researchers developed a variety of black-box extensions of VI, which can be directly applied to a wider range of models [Paisley *et al.*, 2012; Ranganath *et al.*, 2014; Kingma and Welling, 2014; Titsias and Lazaro-Gredilla, 2015; Knowles, 2015; Ruiz *et al.*, 2016b; Mnih and Rezende, 2016]. These extensions optimize the variational parameters following the spirit of stochastic optimization. They form noisy gradients by using Monte Carlo samples, leading to model-independent derivations. Such optimization will converge to a local optimum, since the noisy Monte Carlo gradients are unbiased estimators for the true gradients.

To the best of our knowledge, there exist two primary ways of forming Monte Carlo gradients, the log-derivative trick [Ranganath *et al.*, 2014] and the reparameterization trick [Kingma and Welling, 2014]. In this work, we are interested in the former one since it is more generic. Its basic idea is that first derives an expectation form of the gradient of the variational objective using the log-derivative trick, and then approximates the gradient using Monte Carlo samples from the variational distribution. A prior art proposed by [Paisley *et al.*, 2012] focuses on approximating the gradient of the intractable terms in the variational objective. Further, black-box variational inference (BBVI) [Ranganath *et al.*, 2014] directly forms the Monte Carlo estimator for the full gradient. This BBVI is generic and feasible for many models, however, the variance of the Monte Carlo gradient may be too large to be useful. This often leads to slower convergence and even worse approximations. A recent algorithm, namely overdispersed black-box variational inference (O-BBVI) [Ruiz *et al.*, 2016b], uses importance sampling for variance reduction, suggesting a simple overdispersed proposal distribution that has heavier tails than the variational distribution. It yields lower variance than the standard BBVI and therefore leads to more precise optimization.

The variance reduction effects of O-BBVI mainly depend on the type of the proposal distribution. In this paper, we aim to further investigate how to obtain a better proposal than the overdispersed proposal used in O-BBVI. Actually, previous importance sampling literatures [Owen, 2013; Rubinstein and Kroese, 2016] show us that the unnormalized

*corresponding author

density of the optimal proposal is known, where "optimal" means least variance estimator. Following this, at each iteration of the variational optimization we directly approximate the current optimal proposal for the parameters of interest. This is achieved by using a Monte Carlo moment matching step, so that we call the proposed black-box proposal moment matching proposal (MMP). For more precise MMPs, we use a moving average across variational iterations to approximate the MMPs.

We studied the proposed MMPs on two Bayesian models. Both synthetic and real world data sets are used in the empirical evaluations. Experimental results show that the MMPs can reduce the variance than the standard O-BBVI with the overdispersed proposal, and perform better.than baseline algorithms.

2 Background

We first review variational inference [Jordan, 1999; Wainwright and Jordan, 2008],and then introduce black-box variational inference [Ranganath *et al.*, 2014] and overdispersed black-box variational inference [Ruiz *et al.*, 2016b].

2.1 Variational Inference

Consider a probabilistic model that involves observations x, hidden variables θ and a prior distribution $p(\theta|\lambda_0)$. The joint distribution of this model is:

$$p(x, \theta|\lambda_0) = p(\theta|\lambda_0) \prod_{n=1}^{N} p(x_n|\theta)$$

What we commonly concern is the posterior distribution of hidden variables, i.e., $p\left(\theta|x, \lambda_0\right)$. However, for many models it is intractable to compute. Variational inference (VI) is a popular approximate inference algorithm.

In VI, we first posit a simple variational family of distributions over hidden variables $q\left(\theta|\widehat{\lambda}\right)$ with variational parameters $\widehat{\lambda}$, and then find the optimal member of the variational family that minimizes the Kullback-Leibler (KL) divergence to the true posterior [Jordan, 1999; Wainwright and Jordan, 2008]:

$$\min_{q} \ \mathrm{KL}\left(q\left(\theta|\widehat{\lambda}\right)||p\left(\theta|x, \lambda_0\right)\right)$$

This minimization is equivalent to maximizing a lower bound of the log marginal likelihood of the observations, i.e., evidence lower bound (ELBO), leading to a variational objective with respect to $\widehat{\lambda}$:

$$\mathcal{L}(\widehat{\lambda}) \triangleq \mathrm{E}_q\left[\log p\left(x, \theta|\lambda_0\right) - \log q\left(\theta|\widehat{\lambda}\right)\right] \tag{1}$$

where $\mathrm{E}_q\left[\cdot\right]$ is the expectation with respect to $q(\theta|\widehat{\lambda})$.

2.2 Black-box Variational Inference

For conditionally conjugate models, we can optimize the variational objective of Eq.1 using coordinate ascent in closed form [Ghahramani and Beal, 2001]. However, for many complex models, e.g., non-conjugate models, there is no analytic solution.

Black-box variational inference (BBVI) [Ranganath *et al.*, 2014] is a generic approximate inference algorithm that can be directly applied to a wider range of models. BBVI is built on stochastic optimization [Robbins and Monro, 1951], where it optimizes the variational objective by forming Monte Carlo gradients. This is achieved by transforming the gradient of $\mathcal{L}(\widehat{\lambda})$ into an expectation form using the log-derivative trick [Williams, 1992]:

$$\nabla_{\widehat{\lambda}}\mathcal{L} = \mathrm{E}_q\left[f\left(\theta\right)\right] \tag{2}$$

where

$$f\left(\theta\right) = \nabla_{\widehat{\lambda}} \log q\left(\theta|\widehat{\lambda}\right)\left(\log p\left(x, \theta|\lambda_0\right) - \log q\left(\theta|\widehat{\lambda}\right)\right)$$

Then approximate this expectation using Monte Carlo samples θ^s from the variational distribution $q(\theta|\widehat{\lambda})$:

$$\nabla_{\widehat{\lambda}}\mathcal{L} \approx \frac{1}{S}\sum_{s=1}^{S} f\left(\theta^s\right) \qquad \theta^s \sim q\left(\theta|\widehat{\lambda}\right) \tag{3}$$

where S is the number of Monte Carlo samples.

We easily identify that the expectation of the noisy gradient is equivalent to the true gradient of $\mathcal{L}(\widehat{\lambda})$, giving the following update of $\widehat{\lambda}$ with a step size ρ:

$$\widehat{\lambda}_{t+1} \leftarrow \widehat{\lambda}_t + \rho_t \nabla_{\widehat{\lambda}_t}\mathcal{L} \tag{4}$$

This update process guarantees to converge to a local maximum of $\mathcal{L}(\widehat{\lambda})$, if the step size ρ satisfies the Robbins-Monro conditions [Robbins and Monro, 1951].

$$\sum_{t=1}^{\infty}\rho_t = \infty, \qquad \sum_{t=1}^{\infty}\rho_t^2 < \infty$$

2.3 Overdispersed Black-box Variational Inference

A potential danger of BBVI is that the Monte Carlo gradient may suffer from high variance, resulting in slower convergence and even worse approximations.

Overdispersed black-box variational inference (O-BBVI) [Ruiz *et al.*, 2016b] aims to reduce the variance of Monte Carlo gradients in BBVI. The main idea is to employ importance sampling. In O-BBVI, the gradient is rewritten by the following expectation form:

$$\nabla_{\widehat{\lambda}}\mathcal{L} = \mathrm{E}_r\left[\frac{q\left(\theta|\widehat{\lambda}\right)}{r\left(\theta\right)} f\left(\theta\right)\right] \tag{5}$$

where $r(\theta)$ is the proposal distribution. Then it can form noisy gradients with Monte Carlo samples from the proposal distribution:

$$\nabla_{\widehat{\lambda}}\mathcal{L} \approx \frac{1}{S}\sum_{s=1}^{S}\frac{q\left(\theta^s|\widehat{\lambda}\right)}{r\left(\theta^s\right)} f\left(\theta^s\right) \qquad \theta^s \sim r(\theta) \tag{6}$$

The importance sampling trick works well when an appropriate proposal distribution is used. O-BBVI constrains that

the variational distribution $q(\theta|\widehat{\lambda})$ is in the exponential family:

$$q\left(\theta|\widehat{\lambda}\right) = exp\left(T\left(\theta\right)^T \widehat{\lambda} - A\left(\widehat{\lambda}\right)\right) \tag{7}$$

where $T(\theta)$ and $A(\widehat{\lambda})$ are the vectors of sufficient statistics and the log normalization constant, respectively; and here $\widehat{\lambda}$ denotes the natural parameter. It then defines the proposal distribution that is in the same exponential family with a dispersion parameter τ [Jorgensen, 1987]:

$$r\left(\theta\right) = exp\left(\frac{T\left(\theta\right)^T \widehat{\lambda}}{\tau} - A\left(\widehat{\lambda}, \tau\right)\right) \tag{8}$$

The dispersion parameter τ is set to be greater than 1 (i.e., $\tau \geq 1$), leading to an overdispersed proposal distribution with heavier tails that is closer to the optimal proposal distribution discussed in [Owen, 2013].

3 Adaptive Optimal Proposal

In terms of importance sampling, the type of the proposal distribution that is used highly influences variance reduction effects. We aim to further investigate how to set better proposals beyond the overdispersed proposal in O-BBVI.

Reviewing the expectation form with importance sampling in Eq.5, the optimal proposal arrives at [Owen, 2013; Rubinstein and Kroese, 2016]:

$$r^*\left(\theta\right) \propto q\left(\theta|\widehat{\lambda}\right)\left|f\left(\theta\right)\right| \tag{9}$$

Unfortunately, this optimal proposal $r^*(\theta)$ is not easy to use in black-box learning. That is because (1) its normalization constant is often intractable to compute; and (2) more importantly, the model-specific term $f(\theta)$ makes the proposal "black-box" by no means. That is why O-BBVI alienates such form.

In this work, we attempt to directly approximate the optimal proposal of Eq.9 under black-box learning, and then improve O-BBVI for Monte Carlo gradients with lower variance. An adaptive proposal method, namely moment matching proposal (MMP), is proposed.

3.1 Moment Matching Proposal

The spirit of our MMP is to use a black-box moment matching step that directly approximates the optimal importance proposal $r^*(\theta)$ of Eq.9. In more detail, we posit a simple family of distributions $\ddot{r}(\theta|\ddot{\eta})$, and then find the optimal member of this family as an approximation to $r^*(\theta)$, using the moment matching computations:

$$\mathrm{E}_{r^*}\left[T\left(\theta\right)\right] = \mathrm{E}_{\ddot{r}}\left[T\left(\theta\right)\right] \tag{10}$$

We call the family $\ddot{r}(\theta|\ddot{\eta})$ *moment matching approximation* with parameter $\ddot{\eta}$.

For example, we can define the moment matching approximation by a multivariate Gaussian with mean $\ddot{\mu}$ and covariance matrix $\ddot{\Sigma}$, i.e., $\ddot{\eta} = \left\{\ddot{\mu}, \ddot{\Sigma}\right\}$. By setting $T(\theta) = \left\{\theta, \theta\theta^T\right\}$, we can find an optimal Gaussian approximation $\ddot{r}(\theta|\ddot{\mu}, \ddot{\Sigma})$ to $r^*(\theta)$ using the following moment matching computations:

$$\ddot{\mu} = \mathrm{E}_{r^*}\left[\theta\right]$$

$$\ddot{\Sigma} = \mathrm{E}_{r^*}\left[\theta\theta^T\right] - \ddot{\mu}\ddot{\mu}^T$$

The central issue of this approximating process is the computation of the moments with respect to $r^*(\theta)$. Because the target proposal $r^*(\theta)$ involves model-specific terms, we resort to a black-box step to compute its moments.

Reviewing Eq.9, we can rearrange the moment of $r^*(\theta)$ by an expectation form with respect to the variational distribution:

$$\begin{aligned}\mathrm{E}_{r^*}\left[T\left(\theta\right)\right] &= \int \frac{1}{Z} q\left(\theta|\widehat{\lambda}\right)\left|f\left(\theta\right)\right| T\left(\theta\right) d\theta \\ &= \mathrm{E}_q\left[\frac{1}{Z}\left|f\left(\theta\right)\right| T\left(\theta\right)\right]\end{aligned} \tag{11}$$

where Z denotes the normalization constant of $r^*(\theta)$. This normalization Z can also be represented by an expectation form with respect to the variational distribution:

$$Z = \int q\left(\theta|\widehat{\lambda}\right)\left|f\left(\theta\right)\right| d\theta = \mathrm{E}_q\left[\left|f\left(\theta\right)\right|\right] \tag{12}$$

Combining Eq.11 and Eq.12 leads to a new transformation of the concerned moment:

$$\mathrm{E}_{r^*}\left[T\left(\theta\right)\right] = \frac{\mathrm{E}_q\left[\left|f\left(\theta\right)\right| T\left(\theta\right)\right]}{\mathrm{E}_q\left[\left|f\left(\theta\right)\right|\right]} \tag{13}$$

Using this transformation, we can compute a noisy Monte Carlo moment with samples from the variational distribution:

$$\mathrm{E}_{r^*}\left[T\left(\theta\right)\right] \approx \frac{\sum_{m=1}^{M}\left|f\left(\theta^m\right)\right| T\left(\theta^m\right)}{\sum_{m=1}^{M}\left|f\left(\theta^m\right)\right|} \qquad \theta^m \sim q\left(\theta|\widehat{\lambda}\right) \tag{14}$$

where *M* is the number of Monte Carlo samples.

We call the above noisy Monte Carlo moment (MC-moment). Using such noisy moment, we can easily reach the MMP $\ddot{r}(\theta|\ddot{\eta})$, which is a black-box approximation to the true optimal proposal described in Eq.9. Additionally, we declare that the setting of the moment matching approximation type is very feasible. One can use the prior distribution of the hidden variable θ, the Gaussian and any other popular distribution.

3.2 Moving Average of Monte Carlo Moments

Since the MMP estimation is an additional step for computing the importance sampling gradient at each iteration of O-BBVI, we would be inclined to use a small number (i.e., *M*) of samples to estimate the MC-moment in Eq.14. Unfortunately, this may make the variance of the MC-moment very large, resulting in inaccurate MMPs in practice.

We address this problem by approximating the Monte Carlo moments with moving averages across iterations. For each iteration *t*, we update the Monte Carlo moment by

$$\mathrm{E}_{r^*}\left[T\left(\theta\right)\right] \approx \frac{\widehat{g}_t}{\widehat{h}_t} = \frac{\left(1-\frac{1}{P}\right)\widehat{g}_{t-1} + \frac{1}{P} g_t}{\left(1-\frac{1}{P}\right)\widehat{h}_{t-1} + \frac{1}{P} h_t} \tag{15}$$

where P is the iteration window size; and g_t and h_t are the Monte Carlo approximations of the current iteration as follows:

$$g_t = \mathrm{E}_{q_t}\left[\left|f_t\left(\theta\right)\right| T\left(\theta\right)\right] \approx \frac{1}{M}\sum_{m=1}^{M}\left|f_t\left(\theta^m\right)\right| T\left(\theta^m\right)$$

$$h_t = \mathrm{E}_{q_t}\left[T\left(\theta\right)\right] \approx \frac{1}{M}\sum_{m=1}^{M} T\left(\theta^m\right)$$

$$\theta^m \sim q_t\left(\theta|\widehat{\lambda}_t\right) \quad (16)$$

We call this new noisy moving average Monte Carlo moment (AMC-moment).

3.3 Algorithm Outline

We briefly outline the full algorithm of O-BBVI with MMPs in *Algorithm 1*. Actually, we have employed the Rao-Blackwellization and control variates as described in [Ranganath *et al.*, 2014; Ruiz *et al.*, 2016b]. Due to the space limit, we omit details in this paper.

Algorithm 1 O-BBVI with MMPs

1: **Initialize** the variational parameter $\widehat{\lambda}$ randomly
2: **While** *algorithm has not converged* **do**
3: Compute the AMC-moments using Eq.15
4: Compute the MMPs following Eq.10
5: Compute the Monte Carlo gradient $\nabla_{\widehat{\lambda}}\mathcal{L}$ using Eq.6
6: Update the variational parameter $\widehat{\lambda}$ using Eq.4
7: **End**

3.4 Discussion

Reviewing Eq.10, the MMP is the member of the moment matching approximation $\ddot{r}(\theta|\ddot{\eta})$, whose moments are equivalent to the corresponding moments of the optimal proposal $r^*(\theta)$ in Eq.9. If the moment matching approximation $\ddot{r}(\theta|\ddot{\eta})$ is in the exponential family, the MMP becomes the optimal distribution that minimizes the KL-divergence between $r^*(\theta)$ and $\ddot{r}(\theta|\ddot{\eta})$ [Bishop, 2006]:

$$\min_{\ddot{r}} \ \mathrm{KL}\left(r^*\left(\theta\right)||\ddot{r}\left(\theta|\ddot{\eta}\right)\right)$$

Actually, this minimization follows the spirit of expectation propagation [Minka, 2001; Minka and Lafferty, 2002]. We know that the approximation computed by expectation propagation often tends to represent the target distribution's mode with large mass [Minka, 2005], That is, our MMP will tend to have heavier tails to some extent. In the framework of O-BBVI, this is often helpful for the importance sampling step [Ruiz *et al.*, 2016b].

4 Related Work

To the best of our knowledge, the stochastic search VI (SSVI) [Paisley *et al.*, 2012] is an early black-box inference algorithm for non-conjugate models. In SSVI, the variational ELBO is divided into a tractable term and an intractable term. By using the log-derivative trick, one can approximate the derivative of the intractable term by Monte Carlo integration, and then obtain a full gradient that is noisy but unbiased. In this sense, BBVI is a generalization of SSVI, where it directly forms Monte Carlo gradients of the ELBO, instead of the intractable term. To reduce the variance of Monte Carlo gradients, O-BBVI [Ruiz *et al.*, 2016b] uses the importance sampling with an overdispersed proposal distribution that is heavy tailed. In contrast, the proposed MMP can adaptively approach the optimal proposal distribution, leading to lower variance.

Another line of black-box variational extensions is based on the reparameterization trick, including auto-encoding variational Bayes (AEVB) [Kingma and Welling, 2014] and some recent algorithms [Burda *et al.*, 2016; Mnih and Rezende, 2016] using Monte Carlo variational objectives. The key idea is to use a transformation of a simple random variable by a differentiable mapping function. Empirically, the reparameterization-based inference algorithms often yield lower variance, but they are not as generally applicable. That is because for many distributions of interest (e.g., Gamma and Dirichlet distributions), there are no available mapping functions. Auxiliary methods [Ruiz *et al.*, 2016a; Maddison *et al.*, 2017; Jang *et al.*, 2017; Naesseth *et al.*, 2017] are often needed to leverage the reparameterization trick.

5 Empirical Study

We have described an adaptive proposal, namely MMP, for the importance sampling step in O-BBVI. In this section, we evaluate the performance O-BBVI with MMPs (abbr. O-BBVI-MMP) on two Bayesian models, including Mixture of Gaussians and Bayesian logistic regression [Jaakkola and Jordan, 1997].

We choose three black-box inference algorithms as baselines, including BBVI [Ranganath *et al.*, 2014], O-BBVI [Ruiz *et al.*, 2016b] and AEVB [Kingma and Welling, 2014].

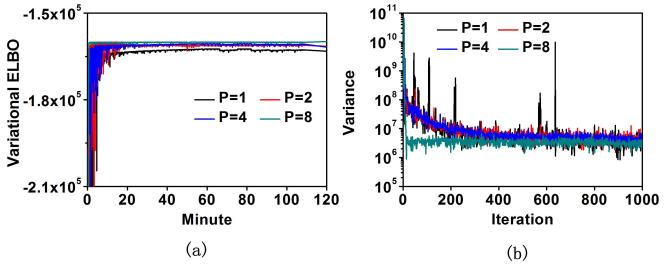
5.1 Mixture of Gaussians

Mixture of Gaussians is a linear combination of K Gaussian bases. It generates each observation by first drawing a Gaussian base z_n from the mixture proportion π, and then drawing the observation $x_n \in \mathcal{R}^D$ from the corresponding Gaussian bases $\mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$. We place a Gaussian prior with mean μ_0 and covariance Σ_0 on all K Gaussian means $\mu_{1:K}$. For simplicity, all K variances $\Sigma_{1:K}$ are fixed as the identity matrix I_D. The generative process of the model is as follows:

1. For $k = 1, \cdots, K$
   (a) Draw a mean parameter $\mu_k \sim \mathcal{N}(\mu_0, \Sigma_0)$
2. For $n = 1, \cdots, N$
   (a) Draw a Gaussian base $z_n \sim Multinomial(\pi)$
   (b) Draw an observation $x_n \sim \mathcal{N}(\mu_{z_n}, I_D)$

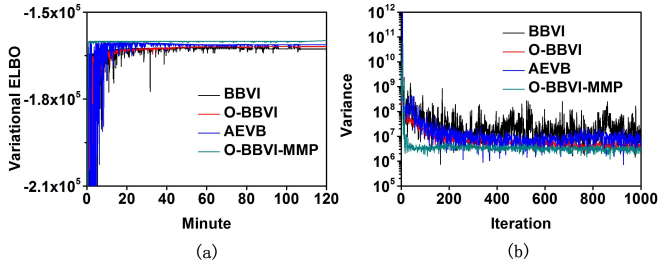
The joint probability of this mixture of Gaussians is::

$$p\left(x, \mu, z|\mu_0, \Sigma_0, \pi\right) = \prod_{k=1}^{K} p\left(\mu_k|\mu_0, \Sigma_0\right) \prod_{n=1}^{N} p\left(z_n|\pi\right) p\left(x_n|\mu_{z_n}, I_D\right)$$



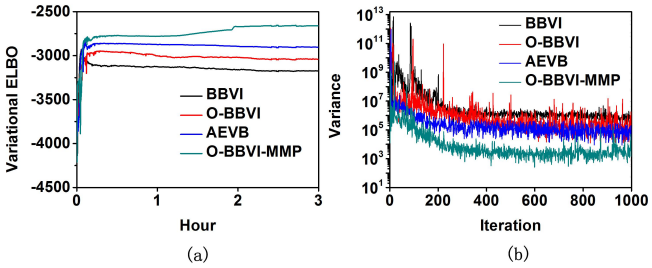
(a) (b)

Figure 1: Results of different P values: (a) the ELBO (Higher is better) and (b) averaged variance of Monte Carlo gradients (Lower is better).



(a) (b)

Figure 2: Results of mixture of Gaussians on synthetic data: (a) the ELBO and (b) averaged variance of Monte Carlo gradients.



(a) (b)

Figure 3: Results of mixture of Gaussians on the COIL-20 dataset: (a) the ELBO and (b) averaged variance of Monte Carlo gradients.

We wish to estimate the (intractable) posterior distribution $p(\mu, z|x, \mu_0, \Sigma_0, \pi)$.

Results of Synthetic Data

We will show empirical results of synthetic data of $N = 10,000$ observations. We set that $D = 8$ and $K = 5$. The number of samples S are set to 32 and 16 for baseline algorithms and our algorithm, respectively. Specially, the sample number M used in MMP estimation is set to 8. In this evaluation, we set the moment matching approximation to be the prior distribution of μ, i.e., Gaussian prior.

Evaluation of the AMC-moment with different P values. We first evaluate the iteration window size P of AMC-moments, which are used to estimate the MMPs.

The results of the ELBO and the averaged variance of Monte Carlo gradients are shown in Figure 1. We can observe that as the value of P increases, the ELBO tends to be higher and it converges faster. When P is set to 8, the averaged variance is lower than other P values, so that its ELBO curve seems much more stable. In early experiments, we have examined larger values of P, e.g., 16 and 32, and found that the performance almost unchanged when P is greater than 8. We have not shown the results of larger P values here.

We argue that this evaluation is helpful, since the results indicate that our method can work well using relatively smaller P values (e.g., $P = 8$ here).

Comparison against baseline algorithms. We then compare O-BBVI-MMP against BBVI, the standard O-BBVI and AEVB. For the standard O-BBVI, the dispersion parameters of the overdispersed proposal are adaptively computed following the method in [Ruiz *et al.*, 2016b]. For our O-BBVI-MMP, the iteration window size P is set to 8.

The experimental results are shown in Figure 2. We can observe that the ELBO of O-BBVI-MMP is higher than those of baseline algorithms. Additionally, the variance of O-BBVI-MMP is lower than baseline algorithms. O-BBVI-MMP converges faster. The possible reason is that O-BBVI-MMP can reach a low variance at the beginning of iterations, making the convergence faster.

Results of Real-world Data

We then evaluate our method across two real-world data sets, including a UCI data set *Vote*[1] and an object data set *COIL-20*. The *Vote* data set contains 435 votes for the U.S House of Representatives Congressmen on 16 key votes (i.e., 16-dimensional features), divided into 2 classes. The *COIL-20* data set contains 1,440 images, each with 1,024 pixels after being resized. The images are divided into 20 classes.

Results of the ELBO and variance. Since the ELBO and variance results of the two data sets are very similar, we only show the results across the *COIL-20* data set.

The experimental results are shown in Figure 3. We observe that O-BBVI-MMP outperforms all three baseline algorithms. More importantly, the improvement over O-BBVI implies that the MMP can better approach the optimal proposal distribution. Besides, the variance of O-BBVI-MMP is significantly lower than those of baseline algorithms. This is consistent with the ELBO performance in Figure 3(a).

Results of clustering. We also evaluate the clustering results of mixture of Gaussians learnt by different inference algorithms. Two popular evaluation metrics are used, including clustering accuracy (ACC) and normalized mutual information (NMI). The ACC is computed by:

$$ACC = \frac{\sum_{n=1}^{N} \delta\left(y_n, map\left(c_n\right)\right)}{N}$$

where y_n and c_n are the true class label and estimated cluster label of x_n, respectively; $\delta(\cdot)$ is the indicator function; and $map(c_d)$ is the mapping function using the Hungarian algorithm [Papadimitriou and Steiglitz, 1998]. The NMI is computed by:

$$NMI(Y, C) = \frac{MI(Y, C)}{\sqrt{H(Y) H(C)}}$$

where Y and C are the true label set and the estimated cluster label set, respectively; $MI(Y, C)$ is the mutual information

[1] https://archive.ics.uci.edu/ml/datasets.html

Metric	BBVI	O-BBVI	AEVB	O-BBVI-MMP
ACC	0.826±0.05	0.831±0.04	0.833±0.05	**0.837±0.03**
NMI	0.363±0.02	0.364±0.02	**0.371±0.04**	**0.371±0.01**
ACC	0.683±0.03	0.692±0.02	0.690±0.04	**0.695±0.02**
NMI	0.769±0.04	0.776±0.02	**0.784±0.03**	0.783±0.01

Table 1: The clustering results across *Vote* (top section) and *COIL-20* (bottom section)

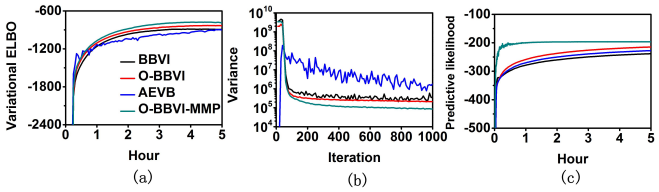


Figure 4: Results of Bayesian Logistic Regression: (a) the ELBO, (b) averaged variance of Monte Carlo gradients and (c) predictive likelihood

of Y and C; and $H(\cdot)$ is the entropy. For both metrics, a higher value indicates better performance.

The experimental results are shown in Table 1. Roughly, we can see that O-BBVI-MMP achieves the best scores on 3/4 of settings. The clustering results indirectly indicate that O-BBVI-MMP is more effective than baseline inference algorithms.

5.2 Bayesian Logistic Regression

Bayesian logistic regression is a popular non-conjugate model for binary classification. For each data point is a tuple $\{x_n, y_n\}$, where $x_n \in \mathcal{R}^D$ is the feature vector and $y_n \in \{-1, +1\}$ the class label. In Bayesian logistic regression, the class label y_n is drawn from a Bernoulli distribution parameterized by $\sigma(\theta^T x_n)$, where $\theta \in \mathcal{R}^D$ is the weight vector and $\sigma(z) \triangleq (1 + \exp(-z))$ the logistic function. It often places a Gaussian prior on the weight vector θ. Bayesian logistic regression assumes the following conditional process:

1. Draw the weight vector $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$
2. For each of the N data points with the feature vector x_n
   (a) Draw the class label $y_n \sim Bernoulli\left(\sigma\left(-\theta^T x_n\right), \sigma\left(\theta^T x_n\right)\right)$

The joint probability of observed labels and the weight vector is as follows::

$$p\left(y, \theta | x, \mu_0, \Sigma_0\right) = p\left(\theta | \mu_0, \Sigma_0\right) \prod_{n=1}^{N} \sigma\left(y_n \theta^T x_n\right)$$

We wish to estimate the posterior distribution $p(\theta|y, x, \mu_0, \Sigma_0)$, and then use it to predict the class labels for future data points. Here, we use VI-based algorithms to approximate the posterior $p(\theta|y, x, \mu_0, \Sigma_0)$. The variational distribution for the weight vector θ is a D-dimensional multivariate Gaussian:

$$q\left(\theta | \widehat{\mu}, \widehat{\Sigma}\right) = \mathcal{N}\left(\theta | \widehat{\mu}, \widehat{\Sigma}\right)$$

where we restrict the variational covariance $\widehat{\Sigma}$ to be diagonal.

We use a subset of the *MNIST* data set that includes all 14,283 examples from the digit classes 2 and 7, each with 784 pixels. The standard training set contains 12,223 examples and the remaining 2,060 examples are used for testing.

BBVI	O-BBVI	AEVB	O-BBVI-MMP
0.978±0.03	0.979±0.02	0.981±0.04	**0.983±0.01**

Table 2: The results of the classification accuracy across a subset of *MNIST*

Results of the ELBO, Variance and Predictive Likelihood

In this evaluation, we plot the ELBO and variance of the Monte Carlo gradient across the training set, and the predictive likelihood across the test set.

The experimental results are shown in Figure 4. Again, it can be seen that our O-BBVI-MMP performs better than baseline inference algorithms on both ELBO and predictive likelihood. The variance of O-BBVI-MMP is significantly lower, so that it performs more stable and converges faster. Surprisingly, in this evaluation even BBVI beats AEVB.

Results of Classification

Finally, we evaluate the classification results learnt by different inference algorithms. We train the models of Bayesian logistic regression on the training images, and compute the classification accuracy on the testing set. A higher accuracy value implies better performance.

The experimental results are shown in Table 2. Our O-BBVI-MMP achieves the best score of classification accuracy. The results further indicate that O-BBVI-MMP is more effective indirectly.

6 Conclusion

In this paper, we develop a moment matching proposal (MMP) method, which can adaptively compute the proposal distribution in O-BBVI. The idea behind MMP is to approximate the optimal proposal in theory using a Monte Carlo moment matching step. We further improve the quality of MMPs by moving average across variational iterations. Experimental results on two popular Bayesian models indicate that the MMPs can effectively reduce variance in black-box learning, and perform well.

Acknowledgements

We would like to acknowledge support for this project from the National Natural Science Foundation of China (NSFC) (grant numbers 61602204, 61572226 and 61472157).

References

[Bishop, 2006] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[Blei and Lafferty, 2007] David M. Blei and John D. Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Burda *et al.*, 2016] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *International Conference on Learning Representations*, 2016.

[Ghahramani and Beal, 2001] Zoubin Ghahramani and Matthew J. Beal. Propagation algorithms for variational Bayesian learning. *Neural Information Processing Systems*, 2001.

[Ghahramani and Jordan, 1997] Zoubin Ghahramani and Michael I. Jordan. Factorial hidden markov models. *Machine Learning*, 29(2-3):245–273, 1997.

[Honkela and Valpola, 2004] Antti Honkela and Harri Valpola. Unsupervised variational bayesian learning of nonlinear models. *Neural Information Processing Systems*, 2004.

[Jaakkola and Jordan, 1997] Tommi S. Jaakkola and Michael I. Jordan. A variational approach to bayesian logistic regression models and their extensions. *Artificial Intelligence and Statistics*, 1997.

[Jang *et al.*, 2017] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization using gumbel-softmax. *International Conference on Learning Representations*, 2017.

[Jordan, 1999] Michael I. Jordan. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[Jorgensen, 1987] Bent Jorgensen. Exponential dispersion models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 49(2):127–162, 1987.

[Kingma and Welling, 2014] Durk Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2014.

[Knowles, 2015] David A. Knowles. Stochastic gradient variational bayes for Gamma approximating distributions. *arXiv:1509.01631*, 2015.

[Maddison *et al.*, 2017] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Learning Representations*, 2017.

[Minka and Lafferty, 2002] Thomas P Minka and John Lafferty. Expectation-propagation for the generative aspect model. *Conference on Uncertainty in Artificial Intelligenc*, pages 352–359, 2002.

[Minka, 2001] Thomas P Minka. Expectation propagation for approximate Bayesian inference. *Conference on Uncertainty in Artificial Intelligenc*, pages 362–369, 2001.

[Minka, 2005] Thomas P Minka. Divergence measures and message passing. (MSR-TR-2005-173), 2005.

[Mnih and Rezende, 2016] Andriy Mnih and Danilo J. Rezende. Variational inference for monte carlo objective. *International Conference on Machine Learning*, 2016.

[Naesseth *et al.*, 2017] Christian A. Naesseth, Francisco J. R. Ruiz, Scott W. Lindermanz, and David M. Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. *International Conference on Artificial Intelligence and Statistics*, 2017.

[Owen, 2013] Art B. Owen. Monte Carlo theory, methods and examples. *http://statweb.stanford.edu/∼owen/mc/*, 2013.

[Paisley *et al.*, 2012] John Paisley, David M. Blei, and Michael I. Jordan. Variational Bayesian inference with stochastic search. *International Conference on Machine Learning*, 2012.

[Papadimitriou and Steiglitz, 1998] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[Ranganath *et al.*, 2014] Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black-box variational inference. *Artificial Intelligence and Statistics*, 2014.

[Robbins and Monro, 1951] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[Rubinstein and Kroese, 2016] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method, 3rd Edition*. Wiley Series in Probability and Statistics, 2016.

[Ruiz *et al.*, 2016a] Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. The generalized reparameterization gradient. *Neural Information Processing Systems*, 2016.

[Ruiz *et al.*, 2016b] Francisco J. R. Ruiz, Michalis K. Titsias, and David M. Blei. Overdispersed black-box variational inference. *Conference on Uncertainty in Artificial Intelligence*, pages 647–656, 2016.

[Titsias and Lazaro-Gredilla, 2015] Michalis K. Titsias and Miguel Lazaro-Gredilla. local expectation gradients for black box variational inference. *Neural Information Processing Systems*, pages 2620–2628, 2015.

[Wainwright and Jordan, 2008] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[Wang and Blei, 2013] Chong Wang and David M. Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:1005–1031, 2013.

[Williams, 1992] Ronald J. Williams. Simple statistical gradient following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.