Unbiased Implicit Variational Inference

Michalis K. Titsias *Athens University of Economics and Business* MTITSIAS@AUEB.GR
Francisco J. R. Ruiz *Univ. of Cambridge & Columbia University* F.RUIZ@COLUMBIA.EDU

Abstract

We develop *unbiased implicit variational inference* (UIVI), a method that expands the applicability of variational inference by defining an expressive variational family. UIVI considers an implicit variational distribution obtained in a hierarchical manner using a simple reparameterizable distribution whose variational parameters are defined by arbitrarily flexible deep neural networks. Unlike previous works, UIVI directly optimizes the evidence lower bound (ELBO) rather than an approximation to the ELBO. We demonstrate UIVI on several models, including Bayesian multinomial logistic regression and variational autoencoders, and show that UIVI achieves both tighter ELBO and better predictive performance than existing approaches at a similar computational cost.

Keywords: variational inference, implicit distributions, stochastic gradient

1. Introduction

We consider the problem of approximating the posterior distribution $p(z \mid x)$ of a probabilistic model $p(x, z)$ using an *implicit* variational distribution $q_\theta(z)$. A distribution $q_\theta(z)$ is implicit when it is not possible to evaluate its density but it is possible to draw samples from it. One typical way to draw from an implicit distribution in variational inference (VI) is to first sample a noise vector and then push it through a deep neural network (Mohamed and Lakshminarayanan, 2016; Huszár, 2017; Tran et al., 2017; Li and Turner, 2018; Mescheder et al., 2017; Shi et al., 2018).

VI maximizes the evidence lower bound (ELBO), given by

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)} \left[\log p(x, z) - \log q_\theta(z) \right]. \tag{1}$$

Implicit VI expands the variational family making $q_\theta(z)$ more expressive, but computing the entropy term in the ELBO—or its gradient—becomes intractable. To address that, implicit VI typically relies on density ratio estimation (Goodfellow et al., 2014). However, density ratio estimation is challenging in high-dimensional settings (Sugiyama et al., 2012).

We develop an unbiased estimator of the gradient of the ELBO that avoids density ratio estimation. Our approach builds on semi-implicit variational inference (SIVI) (Yin and Zhou, 2018) in that we also define the variational distribution by mixing the variational parameter with an implicit distribution. In contrast to SIVI, we propose an unbiased optimization method that directly maximizes the ELBO rather than a bound. We call our method *unbiased implicit variational inference* (UIVI). We show experimentally that UIVI can achieve better ELBO and predictive log-likelihood than SIVI at a similar computational cost.

2. Method Description

Semi-Implicit Variational Distribution. unbiased implicit variational inference (UIVI) uses a semi-implicit variational distribution $q_\theta(z)$ (Yin and Zhou, 2018); that is, $q_\theta(z)$ is defined in a hierarchical manner with a mixing parameter,

$$\varepsilon \sim q(\varepsilon), \quad z \sim q_\theta(z \,|\, \varepsilon), \qquad \text{or equivalently,} \quad q_\theta(z) = \int q_\theta(z \,|\, \varepsilon) q(\varepsilon) d\varepsilon. \tag{2}$$

Eq. 2 reveals why the resulting variational distribution $q_\theta(z)$ is implicit, as we can obtain samples from it but cannot evaluate its density, as the integral is intractable.

The dependence of the conditional $q_\theta(z \,|\, \varepsilon)$ on the random variable ε can be arbitrarily complex. In UIVI, its parameters are the output of a deep neural network (parameterized by the variational parameters θ) that takes ε as input.

Assumptions. In UIVI, the conditional $q_\theta(z \,|\, \varepsilon)$ must satisfy two assumptions. First, it must be reparameterizable. That is, to sample from $q_\theta(z \,|\, \varepsilon)$, we can first draw an auxiliary variable u and then set z as a deterministic function $h_\theta(\cdot)$ of the sampled u. That is, the process $u \sim q(u)$, $z = h_\theta(u\,;\,\varepsilon)$ generates a sample $z \sim q_\theta(z \,|\, \varepsilon)$. The transformation $h_\theta(u\,;\,\varepsilon)$ is parameterized by the random variable ε and the variational parameters θ, but the auxiliary distribution $q(u)$ has no parameters.

The second assumption on the conditional $q_\theta(z \,|\, \varepsilon)$ is that it is possible to evaluate the log-density $\log q_\theta(z \,|\, \varepsilon)$ and its gradient with respect to z, $\nabla_z \log q_\theta(z \,|\, \varepsilon)$. This is not a strong assumption; indeed it holds for most reparameterizable distributions.

Unbiased Gradient Estimator. Now we derive the unbiased gradient estimator of the ELBO. First, UIVI uses the reparameterization $z = h_\theta(u\,;\,\varepsilon)$ to rewrite Eq. 1 as an expectation with respect to $q(\varepsilon)$ and $q(u)$,

$$\mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)q(u)} \left[\log p(x,z) - \log q_\theta(z) \Big|_{z=h_\theta(u\,;\,\varepsilon)} \right].$$

To obtain the gradient of the ELBO with respect to θ, the gradient operator can now be pushed inside the expectation, as in the standard reparameterization method (Kingma and Welling, 2014; Titsias and Lázaro-Gredilla, 2014; Rezende et al., 2014). This gives two terms: one corresponding to the model and one corresponding to the entropy, $\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)q(u)} \left[g_\theta^{\text{mod}}(\varepsilon, u) + g_\theta^{\text{ent}}(\varepsilon, u) \right]$. These two terms are, respectively,

$$g_\theta^{\text{mod}}(\varepsilon, u) \triangleq \nabla_z \log p(x,z) \Big|_{z=h_\theta(u\,;\,\varepsilon)} \nabla_\theta h_\theta(u\,;\,\varepsilon), \tag{3}$$

$$g_\theta^{\text{ent}}(\varepsilon, u) \triangleq -\nabla_z \log q_\theta(z) \Big|_{z=h_\theta(u\,;\,\varepsilon)} \nabla_\theta h_\theta(u\,;\,\varepsilon). \tag{4}$$

To obtain this decomposition, we have applied the identity that the expected value of the score function is zero, $\mathbb{E}_{q_\theta(z)} \left[\nabla_\theta \log q_\theta(z) \right] = 0$, which reduces the variance of the estimator (Roeder et al., 2017). UIVI estimates the model component in Eq. 3 using samples from $q(\varepsilon)$ and $q(u)$. However, estimating the entropy component in Eq. 4 is harder because the term $\nabla_z \log q_\theta(z)$ cannot be evaluated—the variational distribution $q_\theta(z)$ is an implicit distribution. UIVI addresses this issue rewriting Eq. 4 as an expectation, which enables Monte Carlo estimates of $g_\theta^{\text{ent}}(\varepsilon, u)$. In particular, UIVI rewrites as an expectation the intractable log-density gradient in Eq. 4,

$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon \,|\, z)} \left[\nabla_z \log q_\theta(z \,|\, \varepsilon) \right]. \tag{5}$$

We prove Eq. 5 in Appendix A. This equation shows that the problematic gradient $\nabla_z \log q_\theta(z)$ can be expressed in terms of an expression that can be evaluated—the gradient $\nabla_z \log q_\theta(z \,|\, \varepsilon)$ can be evaluated by assumption. UIVI rewrites the entropy term in Eq. 4 using Eq. 5,

$$g_\theta^{\text{ent}}(\varepsilon, u) = -\mathbb{E}_{q_\theta(\varepsilon' \,|\, z)} \left[\nabla_z \log q_\theta(z \,|\, \varepsilon') \right] \Big|_{z = h_\theta(u \,;\, \varepsilon)} \times \nabla_\theta h_\theta(u \,;\, \varepsilon). \tag{6}$$

The expectation in Eqs. 5 and 6 is taken with respect to the distribution $q_\theta(\varepsilon \,|\, z) \propto q_\theta(z \,|\, \varepsilon) q(\varepsilon)$. We call this distribution the *reverse conditional*. Although the conditional $q_\theta(z \,|\, \varepsilon)$ has a simple form (by assumption, it is a reparameterizable distribution for which we can evaluate the density and its gradient), the reverse conditional is complex because the conditional $q_\theta(z \,|\, \varepsilon)$ is parameterized by deep neural networks that take ε as input. We show below how to efficiently draw samples from the reverse conditional to obtain an estimator of the entropy component in Eq. 6.

Full Algorithm. UIVI builds the stochastic gradient of the ELBO by estimating Eqs. 3 and 6 with a single sample $\varepsilon_s \sim q(\varepsilon)$ and $u_s \sim q(u)$. Estimating the entropy component (Eq. 6) is challenging because it contains an intractable expectation with respect to the reverse conditional $q_\theta(\varepsilon \,|\, z)$. Thus, UIVI forms a Monte Carlo estimator using samples ε'_s from the reverse conditional.

The reverse conditional is a complex distribution due to the complex dependency of the (direct) conditional $q_\theta(z \,|\, \varepsilon)$ on the random variable ε. Consequently, sampling from the reverse conditional may be challenging. UIVI exploits the fact that the samples ε_s that generated z_s are also samples from the reverse conditional. This is because the sampling procedure in Eq. 2 implies that each pair of samples (z_s, ε_s) comes from the joint $q_\theta(z, \varepsilon)$, and thus ε_s can be seen as a draw from the reverse conditional $q_\theta(\varepsilon \,|\, z_s)$. However, although ε_s is a valid sample from the reverse conditional, setting $\varepsilon'_s = \varepsilon_s$ in the estimation of the entropy component (Eq. 6) would break the assumption that ε'_s and ε_s are independent. Instead, UIVI runs a Markov chain Monte Carlo (MCMC) method, such as Hamiltonian Monte Carlo (HMC) (Neal, 2011), to draw samples from the reverse conditional. Crucially, UIVI initializes the MCMC chain at ε_s. In this way, there is no burn-in period in the MCMC procedure, in the sense that the sampler starts from stationarity so that any subsequent MCMC draw gives a sample from the reverse conditional (Robert and Casella, 2005). To reduce the correlation between the sample ε'_s and the initialization value ε_s, UIVI runs more than one MCMC iterations and allows for a short burn-in period. (In the experiments of Section 3, we use 10 MCMC iterations where only the final 5 samples are used to form the Monte Carlo estimate.)

3. Experiments: Bayesian Multinomial Logistic Regression

We now apply UIVI to assess the goodness of the resulting variational approximation and the computational complexity. As a baseline, we compare against SIVI (Yin and Zhou, 2018), which has been shown to outperform other approaches like mean-field VI and be on par with MCMC methods.

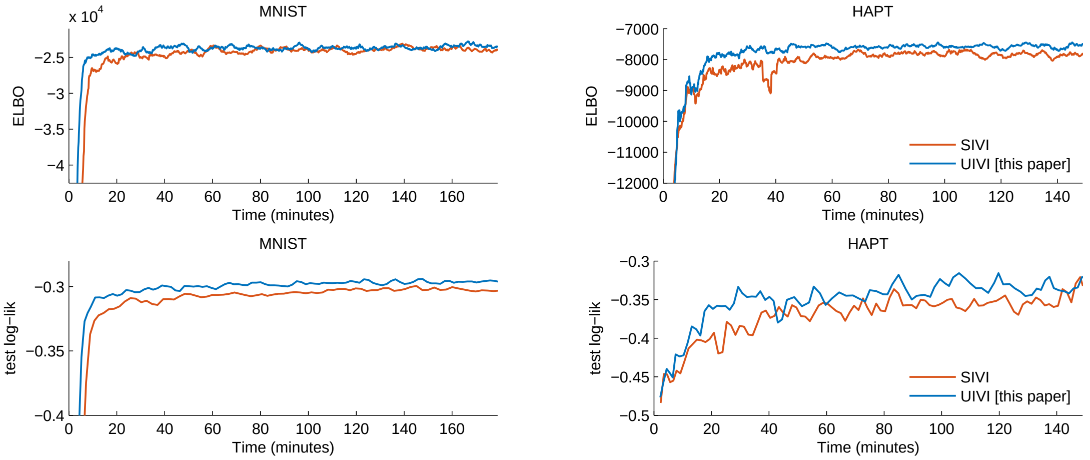

Figure 1. Estimates of the ELBO and the test log-likelihood as a function of wall-clock time. Compared to SIVI (red), UIVI (blue) achieves a better bound on the marginal likelihood and has better predictive performance.

We consider Bayesian multinomial logistic regression. For a dataset of N features x_n and labels $y_n \in \{1, \ldots, K\}$, the model is $p(z) \prod_n p(y_n \,|\, x_n, z)$, where z denotes the latent weights and biases. We set the prior $p(z)$ to be standard Gaussian; the categorical likelihood is $p(y_n = k \,|\, x_n, z) \propto \exp(x_n^\top z_k + z_{0k})$.

We use two datasets, MNIST and HAPT (Reyes-Ortiz et al., 2016). To define the variational distribution, we choose a standard 100-dimensional Gaussian prior for $q(\varepsilon)$. We use a Gaussian conditional $q_\theta(z \,|\, \varepsilon) = \mathcal{N}(z \,|\, \mu_\theta(\varepsilon), \text{diag}\,(\sigma))$, whose mean is parameterized by a neural network with two hidden layers of 200 ReLu units each. We set a diagonal covariance that we also optimize (for simplicity, it does not depend on ε). We run $100{,}000$ iterations of UIVI and SIVI, subsampling minibatches of data at each iteration (Hoffman et al., 2013). (We use a minibatch size of $2{,}000$ for MNIST and 863 for HAPT.) For SIVI, we set the parameter $K = 200$ (Yin and Zhou, 2018).

Results. We obtain a Monte Carlo estimate of the ELBO every 100 iterations. Figure 1 (top) shows the ELBO estimates; the plot has been smoothed using a rolling window of size 20 for easier visualization. UIVI provides a similar bound on the marginal likelihood than SIVI on MNIST and a slightly tighter bound on HAPT. In addition, we also estimate the predictive log-likelihood on the test set every $1{,}000$ iterations. Figure 1 (bottom) shows the test log-likelihood as a function of the wall-clock time for both methods and datasets; the plot has been smoothed with a rolling window of size 2. UIVI achieves better predictions on both datasets.

Finally, we found that the time per iteration was comparable for both methods: SIVI took 0.14 seconds per iteration on MNIST and 0.09 seconds on HAPT, while UIVI took 0.11 and 0.10 seconds.

Acknowledgments

Francisco J. R. Ruiz is supported by the EU Horizon 2020 programme (Marie Skłodowska-Curie Individual Fellowship, grant agreement 706760).

References

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, May 2013.

F. Huszár. Variational inference using implicit distributions. In *arXiv:1702.08235*, 2017.

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

Y. Li and R. E. Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, 2017.

S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. In *arXiv:1610.03483*, 2016.

R. M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 2011.

J. L. Reyes-Ortiz, L. Oneto, A. Samà, X. Parra, and D. Anguita. Transition-aware human activity recognition using smartphones. *Neurocomputing*, 171(C):754–767, jan 2016.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

G. Roeder, Y. Wu, and D. Duvenaud. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, 2017.

J. Shi, S. Sun, and J. Zhu. Kernel implicit variational inference. In *International Conference on Learning Representations*, 2018.

M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

M. K. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.

D. Tran, R. Ranganath, and D. M. Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, 2017.

M. Yin and M. Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, 2018.

Appendix A. Proof of Eq. 5

Here we show how to express the gradient $\nabla_z \log q_\theta(z)$ as an expectation. We start with the log-derivative identity,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \nabla_z q_\theta(z).$$

Next we use the definition of the semi-implicit distribution $q_\theta(z)$ through a mixing distribution (Eq. 2) and we push the gradient into the integral,

$$\begin{aligned} \nabla_z \log q_\theta(z) &= \frac{1}{q_\theta(z)} \nabla_z \int q_\theta(z \,|\, \varepsilon) q(\varepsilon) d\varepsilon \\ &= \frac{1}{q_\theta(z)} \int \nabla_z q_\theta(z \,|\, \varepsilon) q(\varepsilon) d\varepsilon. \end{aligned}$$

We now apply the log-derivative identity on the conditional $q_\theta(z \,|\, \varepsilon)$,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int q_\theta(z \,|\, \varepsilon) q(\varepsilon) \nabla_z \log q_\theta(z \,|\, \varepsilon) d\varepsilon.$$

Finally, we apply Bayes' theorem to obtain Eq. 5. □