TODO

Kenny Chiu

March 17, 2022

1 Critical analysis

1.1 Introduction

TODOTitsias and Ruiz [14] introduce Unbiased Implicit Variational Inference (UIVI) as a variational inference method that allows for a flexible variational family and that addresses the issues of the methods that it is built on. In this analysis, we summarize the work of Titsias and Ruiz [14] in the context of the literature and critically examine the strengths and limitations of UIVI. This analysis is organized as follows: Section 1.2 introduces the problem context and previous work; Section 1.3 describes how UIVI works and how it addresses the limitations of previous methods, and discusses its own limitations. TODOnew section?

1.2 Context and previous work

Variational inference (VI) [3] is a Bayesian inference method that formulates the problem of finding the posterior distribution $p(\mathbf{z}|\mathbf{x})$ of latent variables $\mathbf{z}$ given data $\mathbf{x}$ as an optimization problem. VI posits a variational family $\mathcal{Q} = \{q_\theta\}$ of distributions indexed by variational parameters θ and aims to approximate the posterior distribution by some simpler variational distribution $q_\theta(\mathbf{z}) \in \mathcal{Q}$. In standard VI, the selected distribution q_θ is the one that minimizes the Kullback-Leibler (KL) divergence of q_θ and $p(\mathbf{z}|\mathbf{x})$ or equivalently, the one that maximizes the evidence lower bound (ELBO) denoted as

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

To maximize the ELBO, standard VI places strong restrictions on the choice of the model and the variational family in order to allow the use of a coordinate ascent algorithm. These restrictions include (1) a mean-field assumption where the latent variables are marginally independent and the variational distribution factorizes as $q_\theta(\mathbf{z}) = \prod_{i=1}^{d} q_{\theta_i}(\mathbf{z}_i)$ and (2) the model has conjugate conditionals where $p(\mathbf{z}_i)$ and $p(\mathbf{z}_i|\mathbf{x}, \mathbf{z}_{\neg i})$ are from the same distribution family.

A major development from standard VI was black box VI (BBVI) [9] which relaxed the restrictive assumptions by optimizing the ELBO using a different approach. By rewriting the ELBO gradient in terms of an expectation, the gradient could be estimated via Monte Carlo approaches. Exchanging the above assumptions for the different assumption that one can sample from the variational distribution $q_\theta(\mathbf{z})$ expanded the possibilities for the choice of the variational family. In particular, hierarchical variational families [10] of the form $q_\theta(\mathbf{z}) = \int q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}) q_\theta(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon}$ were proposed with their main strength being the ability to model marginal dependencies between latent variables (which the mean-field family could not) through the mixing distribution $q_\theta(\boldsymbol{\varepsilon})$.

Further pushing the assumption that one only needs to be able to sample from the variational distribution, another development in hierarchical variational models was the incorporation of implicit distributions [6] in various forms. Without needing to evaluate the density of the implicit distribution, flexible models such as normalizing flows [12] and deep neural networks could be leveraged to expand the modeling capacity of the variational family. Using implicit distributions came at a cost of making the log density ratio in the ELBO intractable. Density ratio estimation is one approach for tackling this problem [e.g., 2, 6], but it is known to struggle in high-dimensional regimes [13].

The method that predates UIVI and that was proposed to address the challenges of using implicit distributions in hierarchical variational models is semi-implicit VI (SIVI) [15]. SIVI requires the variational conditional $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}) = q(\mathbf{z}|\boldsymbol{\varepsilon})$ to be reparameterizable [4] and explicit and the mixing distribution $q_\theta(\boldsymbol{\varepsilon})$ also to be reparameterizable but possibly implicit. SIVI then avoids the density ratio estimation problem by instead optimizing a lower bound for the ELBO that is only exact as the number of samples in each iteration goes to infinity [7, 15].

1.3 Current work

Titsias and Ruiz [14] propose UIVI as an alternative to SIVI that directly maximizes the ELBO as an objective

rather than a surrogate lower bound. The idea is that doing so leads to a tighter ELBO bound and therefore ideally faster convergence to the solution.

1.3.1 Unbiased implicit variational inference

Like SIVI, UIVI starts with a hierarchical variational model setup where the variational distribution is

$$q_\theta(\mathbf{z}) = \int q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})q(\boldsymbol{\varepsilon})d\boldsymbol{\varepsilon} \ .$$

UIVI requires the variational conditional $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})$ to be reparameterizable, i.e., that any sample $\mathbf{z} \sim q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})$ can be rewritten as

$$\mathbf{z} = h_\theta(\mathbf{u};\boldsymbol{\varepsilon}) := h_{\boldsymbol{\psi}=g_\theta(\boldsymbol{\varepsilon})}(\mathbf{u})$$

where $h_{\boldsymbol{\psi}}$ is some function with parameters $\boldsymbol{\psi}$ that are the output of some function g_θ with variational parameters θ and input $\boldsymbol{\varepsilon}$. Noise variables $\mathbf{u} \sim q(\mathbf{u})$ and $\boldsymbol{\varepsilon} \sim q(\boldsymbol{\varepsilon})$ are sampled from auxiliary distributions that are fixed. UIVI also requires that $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})$ and its log-gradient $\nabla_{\mathbf{z}} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})$ can be evaluated, which holds for many reparameterizable distributions.

Under these assumptions, the ELBO can be rewritten as an expectation with respect to the noise distributions $q(\mathbf{u})$ and $q(\boldsymbol{\varepsilon})$, and its gradient can be decomposed into two terms given by

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

The first expectation can be estimated using samples from $q(\varepsilon)$ and $q(\mathbf{u})$ while the second expectation is more difficult as $\nabla_z \log q(\mathbf{z})$ may not be evaluated if $q(\mathbf{z})$ is implicit. The first key trick in UIVI is to rewrite the gradient as an expectation given by

$$\nabla_z \log q(\mathbf{z}) = \mathbb{E}_{q_\theta(\boldsymbol{\varepsilon}|\mathbf{z})}\left[\nabla_{\mathbf{z}} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})\right]$$

which then allows for Monte Carlo estimation using samples from $q_\theta(\boldsymbol{\varepsilon}|\mathbf{z}) \propto q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})q(\boldsymbol{\varepsilon})$. The second key trick in UIVI is to reuse a sample $(\mathbf{z}_i, \boldsymbol{\varepsilon}_i)$ from estimating the outer expectation as an initial point in a Markov chain Monte Carlo (MCMC) sampler. As the initial point is a sample from the same joint distribution $q_\theta(\mathbf{z},\boldsymbol{\varepsilon})$, no burn-in is necessary and the only purpose of the MCMC is to break the dependence between samples used to estimate the inner and outer expectations. Thus, the gradient of the ELBO is estimated by

$$\widehat{\nabla}_\theta \mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\mathbf{z}} \log p(\mathbf{x},\mathbf{z})\big|_{\mathbf{z}=h_\theta(\mathbf{u}_i;\boldsymbol{\varepsilon}_i)} - \frac{1}{m}\sum_{j=1}^{m}\nabla_{\mathbf{z}} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}'_j)\big|_{\mathbf{z}=h_\theta(\mathbf{u}_i;\boldsymbol{\varepsilon}_i)}\right)\nabla_\theta h_\theta(\mathbf{u}_i;\boldsymbol{\varepsilon}_i)$$

where $\boldsymbol{\varepsilon}_i \sim q(\boldsymbol{\varepsilon})$, $\mathbf{u}_i \sim q(\mathbf{u})$, $\boldsymbol{\varepsilon}'_j \sim q_\theta(\boldsymbol{\varepsilon}|\mathbf{z})$ and with $n = 1$ said to be used in practice.

1.3.2 Other contributions

Aside from the UIVI algorithm, other contributions of the paper by Titsias and Ruiz [14] include the empirical evaluations of UIVI on synthetic and benchmark datasets. Using a Gaussian conditional with a neural network for the mean parameter, Hamiltonian Monte Carlo (HMC) for the MCMC estimation of the ELBO gradient, and otherwise a fairly standard setup, UIVI is visually shown to be able to approximate various synthetic 2D distributions. Under a similar setup, UIVI is shown to be able to achieve better predictive performance than SIVI on the MNIST and HAPT [11] datasets while being comparable in terms of time per iteration. Finally, Titsias and Ruiz [14] show that using a semi-implicit variational distribution in a variational autoencoder (VAE) [4], UIVI achieves a greater marginal log-likelihood on the test set compared to standard VAE and SIVI on the MNIST and Fashion-MNIST datasets.

1.3.3 Limitations

The paper by Titsias and Ruiz [14] has few limitations. The main limitation is the lack of theoretical guarantees for the performance and convergence of UIVI. However, this is a common problem across the VI literature and generally stems from the challenge of analyzing general purpose methods that may include intractable and non-analytic components. TODOother limitations?

In terms of UIVI itself, related work published after UIVI reported limitations in scalability to the number of latent parameters [5, 8]. This is likely a consequence of optimizing the ELBO via stochastic gradients, for which iterations are cheap but at a cost of requiring more iterations particularly with increased dimensions. Using MCMC may also lead to higher variance in the ELBO gradient estimates [1]. TODOother issues with MCMC?

TODOlabel switching issues with mixtures?

1.4 Other related work

TODOwork that came after?

2 Project report

TODOtitle

Abstract

TODO

2.1 Introduction

2.2 Quality of approximation

2.3 Variance of gradient

2.4 Discussion

References

[1] Michael Betancourt. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *International Conference on Machine Learning*, pages 533–540. PMLR, 2015.

[2] Ferenc Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

[3] Michael I. Jordan, Zoubin Ghahramani, and et al. An introduction to variational methods for graphical models. In *Machine Learning*, pages 183–233. MIT Press, 1999.

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[5] Vincent Moens, Hang Ren, Alexandre Maraval, Rasul Tutunov, Jun Wang, and Haitham Ammar. Efficient semi-implicit variational inference. *arXiv preprint arXiv:2101.06070*, 2021.

[6] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[7] Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, and Dmitry Vetrov. Doubly semi-implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2593–2602. PMLR, 2019.

[8] Iuliia Molchanova, Dmitry Molchanov, Novi Quadrianto, and Dmitry Vetrov. Structured semi-implicit variational inference. 2019.

[9] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

[10] Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333. PMLR, 2016.

[11] Jorge-Luis Reyes-Ortiz, Luca Oneto, Alessandro Ghio, Albert Samá, Davide Anguita, and Xavier Parra. Human activity recognition on smartphones with awareness of basic activities and postural transitions. In *International Conference on Artificial Neural Networks*, pages 177–184. Springer, 2014.

[12] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

[13] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[14] Michalis K Titsias and Francisco Ruiz. Unbiased implicit variational inference. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 167–176. PMLR, 2019.

[15] Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *International Conference on Machine Learning*, pages 5660–5669. PMLR, 2018.