TODO

Kenny Chiu

March 20, 2022

1 Critical analysis

1.1 Introduction

TODOTitsias and Ruiz [18] 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 [18] 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, how it addresses the limitations of previous methods, and its own limitations; and Section 1.4 highlights related work in the recent literature and discusses the general direction that the literature is moving towards.

1.2 Context and previous work

Variational inference (VI) [4] 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 $\mathbf{z}$ 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) [12] 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. One such proposed family was the hierarchical variational model (HVM) [13] containing distributions of the form $q_\theta(\mathbf{z}) = \int q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}) q_\theta(\boldsymbol{\varepsilon}) d\boldsymbol{\varepsilon}$. An advantage of these hierarchical distributions over other variational distributions is the ease in being able 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, one development in the hierarchical variational model literature was the incorporation of implicit distributions [8] in various forms. Without needing to evaluate the density of the implicit distribution, flexible models such as normalizing flows [15] 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., 3, 8], but it is known to struggle in high-dimensional regimes [17].

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) [19]. SIVI requires the variational conditional $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}) = q(\mathbf{z}|\boldsymbol{\varepsilon})$ to be reparameterizable [5] and explicit and requires 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 [9, 19].

1.3 Current work

Titsias and Ruiz [18] 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 deterministic function with parameters $\boldsymbol{\psi}$ that are the output of some function g_θ that depends on variational parameters θ and input $\boldsymbol{\varepsilon}$. To sample from $q_\theta(\mathbf{z})$, noise variables $\mathbf{u} \sim q(\mathbf{u})$ and $\boldsymbol{\varepsilon} \sim q(\boldsymbol{\varepsilon})$ are first sampled from fixed auxiliary distributions and then fed through h_θ. 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 common reparameterizable distributions such as Gaussian.

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 $q(\boldsymbol{\varepsilon})$ and $q(\mathbf{u})$ while the second expectation is more difficult as $\nabla_z \log q(\mathbf{z})$ may not be computable if $q(\mathbf{z})$ is implicit. The first key trick in UIVI is to rewrite the gradient in the second term 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 $q_\theta(\boldsymbol{\varepsilon}|\mathbf{z}) \propto q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})q(\boldsymbol{\varepsilon})$. A Markov chain Monte Carlo (MCMC) sampler is used to sample from $q_\theta(\boldsymbol{\varepsilon}|\mathbf{z})$, and the second key trick in UIVI is to reuse the sample $(\mathbf{z}_i, \boldsymbol{\varepsilon}_i)$ used to estimate the outer expectation as an initial point in the 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 $n = 1$, $m = 5$ 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 [18] 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 shown to be able to visually 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 [14] datasets while being comparable in terms of time per iteration. Finally, Titsias and Ruiz [18] show that using a semi-implicit variational distribution in a variational autoencoder (VAE) [5], 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 [18] has a 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 [7, 10]. This is likely a consequence of the stochastic optimization of the ELBO as well as the use of MCMC, for both of which the number of samples needed to provide a reasonable estimate of a mean grow quickly with the number of dimensions. Using MCMC may also lead to higher variance in the ELBO gradient estimates [1]. TODOother issues with MCMC? non-parallelizable ([16])?

TODOlabel switching issues with mixtures?

1.4 Other related work

While UIVI was proposed as an improved alternative to SIVI, there does not appear to be follow-up work in the literature that directly extends UIVI. As mentioned in the previous section, the inefficiency of MCMC in high-dimensional regimes is often cited as the main problem of UIVI [7, 10]. It appears that rather than trying to address this issue in UIVI, recent work in the literature tend to start with SIVI and propose methods that either improve the quality of approximation or let it scale more efficiently to high dimensions.

Several strategies for improving the SIVI approximation have been proposed in the literature around the time of or after the work by Titsias and Ruiz [18]. Molchanov et al. [9] proposed *doubly* SIVI (DSIVI) that expands the flexibility of standard SIVI by allowing both the posterior and prior to be semi-implicit. Sobolev and Vetrov [16] introduced *importance weighted hierarchical* VI (IWHVI), which optimizes a SIVI-like lower bound that incorporates elements from the bound used in importance weighted autoencoders [2]. SIVI, DSIVI and HVM can be seen as special cases of IWHVI and so the bound in IWHVI has the capacity to result in a tighter lower bound [16].

Recent work in the literature have focused more on improving the scalability of SIVI to high dimensions. Molchanova et al. [10] proposed *structured* SIVI where the high-dimensional semi-implicit distribution is assumed to factorize into low-dimensional semi-implicit distributions. Moens et al. [7] introduced *compositional implicit* VI, which integrates various mechanisms into SIVI including an adaptive solver for addressing the bias in the SIVI objective and sketch-based approximations that keeps the method computationally practical for high-dimensional regimes.

Though the majority of developments in the related literature are methodological, there have been some recent forays on the more theoretical side that attempt to provide statistical guarantees and insights for implicit VI. In particular, Plummer et al. [11] derive posterior contraction results for simple *non-linear latent variable models* by drawing connections to Gaussian convolutions. The NL-LVM has a structure that can be seen as a particular choice of the reparameterization and mixing distributions in UIVI, and so we suspect that this work may provide a reasonable starting point for a theoretical analysis of UIVI.

2 Project report

TODOtitle

Abstract

TODO

2.1 Introduction

2.2 Notation

TODOϕ_σ density of $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$. Overload distribution and density. λ Lebesgue measure on $[0,1]$. Borel σ-algebra of $\mathbb{R}$ by $\mathcal{B}$. $C^\beta(\mathcal{Z})$ β-Hölder. $\|\bullet\|_\infty$ supremum norm

2.3 Quality of approximation

TODODue to time constraints on this project, we focus on the simpler question of, *is the true posterior in the* UIVI *variational family?* TODOalso quality. Using similar arguments from the work by Plummer et al. [11] for NL-LVM, we show that under certain assumptions and particular choices of the reparameterization and mixing distribution, we can approximate the true posterior arbitrarily closely. For convenience, we assume that $p(z|x)$ is continuous. TODO: for simplicity, assume univariate latent variable

TODO

To approximate $p(z|x)$, UIVI posits the variational family $\mathcal{Q}$ of distributions of the form

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

where the variational conditional $q_\theta(z|\boldsymbol{\varepsilon})$ is reparameterizable and explicit, but the dependency on θ can be arbitrarily complex. UIVI also requires that the log-gradient $\nabla_z \log q_\theta(z|\boldsymbol{\varepsilon})$ can be evaluated.

Let $q(\boldsymbol{\varepsilon}) = \prod_{i=1}^d q(\varepsilon_i)$ where $q(\varepsilon_i) = \mathrm{Unif}(0,1)$ for $i = 1, \dots, d$, $d \geq 1$. Let $q_\theta(z|\boldsymbol{\varepsilon})$ be univariate Gaussian with mean $\mu_\theta(\boldsymbol{\varepsilon})$ and variance σ^2 where $\mu_\theta : [0,1]^d \to \mathbb{R}$ is some arbitrarily complex function. This distribution satisfies the requirements of UIVI as it is reparameterizable through the form

$$z = h_\theta(u; \boldsymbol{\varepsilon}) = \mu_\theta(\boldsymbol{\varepsilon}) + \sigma u$$

where $u \sim \mathcal{N}(0,1)$, and the log-density and its gradient is given by

$$\begin{aligned}
\log q_\theta(z|\boldsymbol{\varepsilon}) &= -\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \left(z - \mu_\theta(\boldsymbol{\varepsilon})\right)^2 \ , \\
\nabla_z \log q_\theta(z|\boldsymbol{\varepsilon}) &= -\frac{1}{\sigma^2} \left(z - \mu_\theta(\boldsymbol{\varepsilon})\right) \ .
\end{aligned}$$

The key insight of Plummer et al. [11] is that $q_\theta(z)$ has the form of a convolution with a Gaussian kernel, that is,

$$\begin{aligned}
q_\theta(z) &= \int_0^1 q_\theta(z|\boldsymbol{\varepsilon}) q(\boldsymbol{\varepsilon}) \lambda(d\boldsymbol{\varepsilon}) \\
&= \int_0^1 \phi_\sigma\left(z - \mu_\theta(\boldsymbol{\varepsilon})\right) \lambda(d\boldsymbol{\varepsilon}) \\
&= \int_{\mathbb{R}} \phi_\sigma\left(z - t\right) \nu_{\mu_\theta}(dt)
\end{aligned}$$

where $\nu_{\mu_\theta}(B) = \lambda\left(\mu_\theta^{-1}(B)\right)$, $B \in \mathcal{B}$, is the image measure of λ under μ_θ.

Proposition 1. *Let $\mathcal{Q}_\sigma$ denote the variational family described above indexed by the standard deviation σ of $q_\theta(z|\varepsilon)$. Suppose that $\mu_\theta(t) = F^{-1}_{z|x}(t)$ for all $t \in [0,1]$. Then $p(z|x) \in \mathcal{Q}_0$.*

Proof. If $\mu_\theta(t) = F^{-1}_{z|x}(t)$ for all $t \in [0,1]$, then $q_\theta(z) = \phi_\sigma * p(z|x)$. The result immediately follows using the property of Gaussian convolutions that $\phi_\sigma * p(z|x) \to p(z|x)$ pointwise as $\sigma \to 0$. □

TODOassuming that our optimization procedure is able to identify θ

Then by using the approximation properties of Gaussian convolutions, we can characterize the quality of the approximation of $p(z|x)$ by $q(z)$.

To quantify the quality of the approximation, we must make assumptions about the smoothness of $p(z|x)$ and its support. Following Plummer et al. [11], we make the following assumptions:

Assumption 1. $\log p(z|x) \in C^\beta([0,1])$. Define $l_j(z_0) = \nabla^j_z \log p(z|x)\big|_{z=z_0}$ for $j = 1, \ldots, r$ with $r = \lfloor \beta \rfloor$. For any $\beta > 0$, there exists a constant $L > 0$ such that for all $z_1 \neq z_2$,

$$|l_r(z_1) - l_r(z_2)| \leq L|z_1 - z_2|^{\beta - r} \ .$$

Assumption 2. $p(z|x)$ has compact support on $[0,1]$. There exists some interval $[z_1, z_2] \subset [0,1]$ such that $p(z|x)$ is non-decreasing and non-zero on $[0, z_1]$ and non-increasing on $[z_2, 1]$.

TODOexplain assumptions

TODOFollowing Kruijer et al. [6] and Plummer et al. [11], we can construct a sequence of functions $\{p_j\}_{j \geq 0}$ through an iterative procedure where

$$\begin{aligned} p_{j+1}(z|x) &= p(z|x) - \Delta_\sigma p_j(z|x) \ , \\ \Delta_\sigma p_j(z|x) &= \phi_\sigma * p_j(z|x) - p_j(z|x) \end{aligned}$$

for $j \geq 0$ and $p_0(z|x) = p(z|x)$.

Proposition 2. *Suppose that $p(z|x)$ satisfies Assumptions 1 and 2 with $\beta \in (2j, 2j+2]$. Let $F_{z|x}$ be the cumulative distribution function of the posterior $p(z|x)$. TODOfix this If $\mu_\theta(t) = F^{-1}_{z|x}(t)$ for all $t \in [0,1]$, then*

$$\|\phi_\sigma * p_\beta(z|x) - p(z|x)\|_\infty = O(\sigma^\beta)$$

with

$$\phi_\sigma * p_\beta(z|x) = p(z|x)\left(1 + O(\sigma^\beta)\left(\sum_{i=1}^{r} c_i |l_j(z)|^{\frac{\beta}{i}} + c_{r+1}\right)\right)$$

for non-negative constants c_i, $i = 1, \ldots, r+1$ and $z \in [0,1]$.

Proof. Suppose that $\mu_\theta(\mathbf{t}) = F^{-1}_{\mathbf{z}|\mathbf{x}}(\mathbf{t})$ and so $q_\theta(\mathbf{z}) = \phi_\sigma * p(\mathbf{z}|\mathbf{x})$. Then as $\sigma \to 0$, $q_\theta(\mathbf{z}) \to p(\mathbf{z}|\mathbf{x})$. □

2.4 Variance of gradient

2.5 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] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

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

[4] 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.

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

[6] Willem Kruijer, Judith Rousseau, and Aad Van Der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.

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

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

[9] 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.

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

[11] Sean Plummer, Shuang Zhou, Anirban Bhattacharya, David Dunson, and Debdeep Pati. Statistical guarantees for transformation based models with applications to implicit variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 2449–2457. PMLR, 2021.

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

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

[14] 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.

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

[16] Artem Sobolev and Dmitry P Vetrov. Importance weighted hierarchical variational inference. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[18] 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.

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