TODO

Kenny Chiu

March 31, 2022

1 Critical analysis

1.1 Introduction

Titsias and Ruiz [19] introduce *unbiased implicit variational inference* (UIVI) as a variational inference method with a flexible variational family and that addresses the issues of the existing methods that it is built on. In this analysis, we summarize the work of Titsias and Ruiz [19] in the context of the literature and discuss 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) [5] 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 the goal is to identify the variational distribution $q_\theta(\mathbf{z}) \in \mathcal{Q}$ that best approximates the posterior distribution. 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

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

Standard VI maximizes the ELBO using a coordinate ascent algorithm, which requires placing strong restrictions on the choice of the model and the variational family. 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. These assumptions implied that the choice of a variational family were often limited to analytical, exponential families and that marginal dependencies could not be modeled.

An important development after standard VI was black box VI (BBVI) [13], 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 unbiasedly and cheaply using Monte Carlo samples. The optimization approach of BBVI exchanges the restrictive assumptions of standard VI for the different assumption that one can sample from the variational distribution $q_\theta(\mathbf{z})$. This expanded the possibilities for the choice of the variational family. One such proposed family was the hierarchical variational model (HVM) [14] 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 capture marginal dependencies between latent variables 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 trend following the introduction of the HVM was the incorporation of deep neural networks to expand the modeling capacity of the hierarchical variational family. These models took various forms, such as through normalizing flows [16] or through implicit distributions [9] involving deep networks in which the density cannot be evaluated. Though the implicit models are flexible, the log density ratio in the ELBO is intractable in these models. Some works proposed using density ratio estimation to tackle this problem [e.g., 4, 9], but this approach is known to struggle in high-dimensional regimes [18].

The method that precedes UIVI and that was proposed to address the challenges of using implicit distributions in hierarchical variational models is *semi-implicit* VI (SIVI) [20]. SIVI makes use of a semi-implicit variational distribution in which the variational conditional $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}) = q(\mathbf{z}|\boldsymbol{\varepsilon})$ is required to be reparameterizable [6] and explicit while the mixing distribution $q_\theta(\boldsymbol{\varepsilon})$ is required to be also 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 [10, 20].

1.3 Current work

Titsias and Ruiz [19] propose UIVI as an alternative to SIVI that directly maximizes the ELBO as an objective rather than a surrogate lower bound. The motivation for UIVI is that directly optimizing the ELBO objective should be more efficient than optimizing a surrogate and therefore should result in faster convergence to the optimal variational approximation. UIVI allows for an ELBO objective by rewriting the ELBO gradient in terms of two expectations. One expectation is easily estimated using Monte Carlo samples, while the other expectation is over an inverse conditional from which UIVI draws samples using Markov chain Monte Carlo (MCMC).

1.3.1 Unbiased implicit variational inference

Like in 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 reparameterization function with parameters $\boldsymbol{\psi}$ that are the output of some arbitrarily complex 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 $q(\boldsymbol{\varepsilon})$ through a change of variables, and its gradient can be decomposed into two expectation 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 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 [19] 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 [15] datasets while being comparable in terms of time per iteration. Finally, Titsias and Ruiz [19] show that for a variational autoencoder (VAE) [6] with a semi-implicit variational distribution, 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 [19] has a few limitations. The main limitation is the lack of theoretical guarantees for the performance and convergence of UIVI. For example, it is unclear what the modeling capabilities are for the UIVI variational family, and no guidance is given on how to construct the model such that the true posterior is in or at least well-approximated by a member of the variational family. It is also unclear how good of an approximation UIVI is able to guarantee through its optimization procedure. However, we recognize that 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.

Another notable limitation of the paper is the missing discussion of the limitations of UIVI. In particular, the showcased experiments do not stress test UIVI, and there is no mention of future directions for improving or extending UIVI. Related work published after the paper by Titsias and Ruiz [19] reported limited scalability with the number of latent parameters [8, 11]. This is likely a consequence of the stochastic optimization of the ELBO as well as the use of MCMC, both for which require an increasing number of samples to maintain estimation quality in response to an increasing number of dimensions. The MCMC sampling in the UIVI optimization procedure may also result in greater variance of the ELBO gradient estimates [1] and further contribute to non-scalability by complicating potential parallelization of the algorithm [17]. In some instances, other issues common to MCMC approaches, such as poor mixing over different modes, appear to inhibit the performance of UIVI [17].

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 [8, 11]. It appears that rather than trying to address this issue in UIVI, recent work in the literature return to SIVI and propose methods that either improve the quality of its approximation or allow it to 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 [19]. Molchanov et al. [10] 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 [17] 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 [17].

Recent work in the literature have focused more on improving the scalability of SIVI to high dimensions. Molchanova et al. [11] proposed *structured* SIVI where the high-dimensional semi-implicit distribution is assumed to factorize into low-dimensional semi-implicit distributions. Moens et al. [8] 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 keep the method computationally practical for high-dimensional regimes.

Though the developments in the related literature are mostly methodological, there have been some recent forays into the more theoretical side that attempt to provide statistical guarantees and insights for implicit VI. In particular, Plummer et al. [12] 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 Plummer et al.'s work may provide a reasonable starting point for a theoretical analysis of UIVI.

2 Project report

TODOtitle

Abstract

TODO

2.1 Introduction

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.

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 $\mathcal{B}$.

TODOdelete $C^\beta(\mathcal{Z})$ β-Hölder. $\|\bullet\|_\infty$ supremum norm

2.3 Quality of approximation

TODOoverly simple example showing true posterior not in variational family?

Titsias and Ruiz [19] empirically show that UIVI is seemingly able to match the implicit variational distribution to various synthetic datasets and is able to better approximate several models compared to SIVI. However, they do not provide theoretical guarantees nor quantify the quality of the UIVI approximation. In this section, we show that TODOsimple example where the true posterior distribution is not in variational family, and that under certain assumptions and choices of the reparameterization and mixing distribution, UIVI is able to approximate the true posterior arbitrarily closely. We do so using similar arguments that Plummer et al. [12] made for non-linear latent variable models (NL-LVM). Following their work, we assume that z is a continuous, univariate latent variable. We leave other cases for future work due to time constraints on this project.

2.3.1 Normal model example

We first illustrate through a simple example that using a hierarchical variational family does not magically expand the modeling capacity of the variational distribution, and that the choice of the reparameterization and mixing distributions are still important in realizing the potential of UIVI.

Suppose that the posterior $p(z|x)$ is univariate Gaussian with mean $\mu_{Z|X}$ and known variance $\sigma^2_{Z|X}$. This may be the case, for example, when we have a Gaussian likelihood with latent mean Z and known variance, and a conjugate Gaussian prior for Z. Suppose that we choose to approximate $p(z|x)$ by the UIVI variational family with

$$z = h_\theta(u; \boldsymbol{\varepsilon}) = \theta + \boldsymbol{\varepsilon} + \sigma_{Z|X} u$$

where $\boldsymbol{\varepsilon}$ and u are independent standard normal random variables and θ is the variational parameter to optimize. Then it is easy to see that

$$\begin{aligned} q_\theta(z) &= \int_{\mathbb{R}} q_\theta(z|\boldsymbol{\varepsilon}) q(\boldsymbol{\varepsilon}) \lambda(d\boldsymbol{\varepsilon}) \\ &= \int_{\mathbb{R}} \phi_{\sigma_{Z|X}}(\theta + \boldsymbol{\varepsilon}) \phi_1(\boldsymbol{\varepsilon}) \lambda(d\boldsymbol{\varepsilon}) \\ &= \phi_{\sqrt{\sigma_{Z|X}^2+1}}(\theta) \ . \end{aligned}$$

In other words, our variational family is the set of univariate Gaussian distributions $\left\{\mathcal{N}(\theta, \sigma_{Z|X}^2 + 1) : \theta \in \mathbb{R}\right\}$. The true posterior distribution $\mathcal{N}(\mu_{Z|X}, \sigma_{Z|X}^2)$ is not in this variational family. The problem in this example is that our reparameterized distribution is too restrictive and misspecified. If we changed the reparameterization function to be

$$z = h'_{\boldsymbol{\theta}}(u; \boldsymbol{\varepsilon}) = \theta_1 + \boldsymbol{\varepsilon} + \theta_2 u$$

where both θ_1 and θ_2 are learned parameters, then the new variational family corresponding to this reparameterization includes the true posterior distribution. While this example is very simple, it nicely illustrates that relying on the hierarchical structure alone is insufficient for setting up a flexible variational family.

2.3.2 Flexible variational family

TODOdimensions of $\mathbf{z}$ and $\boldsymbol{\varepsilon}$

Consider the UIVI variational family induced by the following choices of the reparameterization and mixing distribution. We take a potentially multivariate mixing distribution of the form $q(\boldsymbol{\varepsilon}) = \prod_{i=1}^d q(\boldsymbol{\varepsilon}_i)$ where $q(\boldsymbol{\varepsilon}_i) = \text{Unif}(0,1)$ for $i = 1, \ldots, d$, $d \geq 1$. Let the variational conditional $q_{\theta,\sigma}(z|\boldsymbol{\varepsilon})$ 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. Note that we keep the variational parameters θ and σ separate for reasons that will be clear shortly. This distribution is reparameterizable through the form

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

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

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

The above properties suggest that the variational family induced by these choices satisfy the UIVI requirements. Note that the form of $h_{\theta,\sigma}$ also resembles the NL-LVM studied by Plummer et al. [12], which allows us to apply their results with slight modifications.

To study the approximation capability of this model, the key insight of Plummer et al. [12] is that $q_{\theta,\sigma}(z)$ has the form of a convolution with a Gaussian kernel, that is,

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

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. Using the approximation property of Gaussian convolutions, we characterize the relationship between the true posterior and the variational family through the following proposition.

Proposition 1. *Let $\mathcal{Q}_\sigma$ denote the variational family described above indexed by the standard deviation σ of $q_{\theta,\sigma}(z|\boldsymbol{\varepsilon})$. 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 $q_{\theta,\sigma}(z) = \phi_\sigma * p(z|x)$. The result immediately follows using the property of Gaussian convolutions that $\phi_\sigma * p(z|x) \to p(z|x)$ pointwise as $\sigma \to 0$. □

Proposition 1 says that if the inverse CDF or quantile function $F^{-1}_{z|x}$ is in the set of functions $\{\mu_\theta : \theta \in \Theta\}$ that can be modeled by μ_θ, then the true posterior $p(z|x)$ is a limiting member of the sequence of best approximations by this variational family as the bandwidth σ of the Gaussian kernel shrinks to zero. While this result suggests that the true posterior is not in the variation family for $\sigma > 0$, it does imply that for any measure of error, we can choose σ such that the best approximation will be close to the true posterior within a desired tolerance level. What this result does not address is whether we are able to achieve the best approximation for a given σ in practice. This depends on whether our functional form μ_θ is flexible enough such that $F^{-1}_{z|x} \in \{\mu_\theta : \theta \in \Theta\}$ and whether our optimization procedure is able to identify the correct θ such that $\mu_\theta = F^{-1}_{z|x}$.

TODO: a bit about why rest of Plummer et al. [12] does not apply (posterior contraction)

2.3.3 Directions of future work for quantifying quality

TODOkde

TODOuniversal approximation property of neural nets?

TODO: problem boils down to learning the quantile function.

2.3.4 TODOto delete

TODOthe results below seem to be leading towards posterior contraction; proposition 2 isn't a result about $q_{\theta,\sigma}(z)$. Theorem 3.1 may still be relevant?

We can further quantify the quality of the approximation by $q_{\theta,\sigma}(z)$ if we make assumptions about the smoothness of $p(z|x)$ and its support. Following Plummer et al. [12], 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 on $[0, z_1]$, non-zero on $[z_1, z_2]$, and non-increasing on $[z_2, 1]$.

Assumption 1 says that the derivatives of $\log p(z|x)$ up to order r are β-Hölder continuous, implying that $\log p(z|x)$ is smooth to an extent. The proofs of Kruijer et al. [7] and Plummer et al. [12] rely heavily on the assumed smoothness in order to ensure that the error between the target distribution and an approximating convolution can be bounded. Assumption 2 says that the mass of $p(z|x)$ is concentrated in some compact interval of z and that the tails of $p(z|x)$ outside this interval can be bounded above. This allows the approximation error in the tails to be bounded even as the convolution bandwidth shrinks, and so an analysis of the error only needs to focus on the closed interval in which the mass is concentrated. Plummer et al. [12] appear to specify an interval of $[0,1]$ for analytical convenience, whereas the intervals in similar assumptions made by Ghosal et al. [3] feature arbitrary finite endpoints.

Under Assumptions 1 and 2, Plummer et al. [12] follow the work of Kruijer et al. [7] and consider a sequence of functions $\{p_j\}_{j \geq 0}$ constructed through an iterative procedure given by

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

with $p_0(z|x) = p(z|x)$. The quality of the approximation is then characterized in terms of the error between the convolution TODO

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

TODOhigh dimension

unbiased ELBO gradient estimator

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

$n = 1$ used in practice, would like to study variance as $q_\theta(\mathbf{z})$ approaches $p(\mathbf{x},\mathbf{z})$ but arbitrary function complicates analysis. Assume $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})$ Gaussian

Let $L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) = \widehat{\nabla}_\theta \mathcal{L}(\theta; \mathbf{u}_{1:n}, \boldsymbol{\varepsilon}_{1:n}, \boldsymbol{\varepsilon}'_{1:m})$ and let $L_\theta(\boldsymbol{\varepsilon}'_{1:m}) = \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})|\boldsymbol{\varepsilon}'_{1:m}\right]$. Then following the idea underlying Rao-Blackwellization [13],

$$\mathrm{Var}\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m})\right) = \mathrm{Var}\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right) + \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'}\left[\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) - L_\theta(\boldsymbol{\varepsilon}'_{1:m})\right)^2\right]$$

and so

$$\mathrm{Var}\left(L_\theta(\boldsymbol{\varepsilon}'_{1:m})\right) \leq \mathrm{Var}\left(L_\theta(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m})\right) \ .$$

The variance is lower bounded and we will see that the lower bound is more interpretable. We have

$$\begin{aligned} L_\theta(\boldsymbol{\varepsilon}'_{1:m}) &= \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x},\mathbf{z})\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}'_j)\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] \\ &= \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_{\mathbf{z}} \log p(\mathbf{x},\mathbf{z})\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\nabla_\theta h_\theta(\mathbf{u};\boldsymbol{\varepsilon})\right] - \frac{1}{m}\sum_{j=1}^{m}\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}}\left[\nabla_\theta \log q_\theta(\mathbf{z}|\boldsymbol{\varepsilon}'_j)\big|_{\mathbf{z}=h_\theta(\mathbf{u};\boldsymbol{\varepsilon})}\right] \ . \end{aligned}$$

Note that the first term is constant. Given a particular $\boldsymbol{\varepsilon}'_j$, the second term can be interpreted as the expected score function of a sampled Gaussian distribution evaluated at some fixed θ. Suppose that $q(\mathbf{z}) \approx p(\mathbf{x},\mathbf{z})$ and so TODOθ

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] Subhashis Ghosal, Jayanta K Ghosh, and RV Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

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

[5] 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.

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

[7] 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.

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

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

[10] 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.

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

[12] 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.

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

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

[15] 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.

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

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

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

[19] 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.

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