TODO

Kenny Chiu

April 2, 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) [6] 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., 5, 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 [7] and explicit while the mixing distribution $q_\theta(\boldsymbol{\varepsilon})$ is required 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 [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) methods.

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

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

The first expectation can be estimated using samples from $q(\boldsymbol{\varepsilon})$ and $q(\mathbf{u})$ while the second expectation is more difficult as $\nabla_{\mathbf{z}} \log q_\theta(\mathbf{z})$ may not be computable if $q_\theta(\mathbf{z})$ 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) [7] 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]. TODO

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

Variational inference (VI) transform the problem of computing the posterior distribution $p(\mathbf{z}|\mathbf{x})$ into a problem of minimizing the KL divergence (or equivalently, maximizing the evidence lower bound (ELBO)) between the posterior distribution and a simpler variational distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ with variational parameters $\boldsymbol{\theta}$ belonging to some variational family $\mathcal{Q}_{\boldsymbol{\theta}}$ [6]. The performance of VI depends on the flexibility of the family $\mathcal{Q}_{\boldsymbol{\theta}}$ as well as the ability to optimize $\boldsymbol{\theta}$ over this family. Yin and Zhou [20] introduced *semi-implicit variational inference* (SIVI) in which components of the semi-implicit variational distribution could incorporate arbitrarily complex functions (e.g., deep networks) to expand its modeling capacity. However, SIVI relies on optimizing a lower bound of the ELBO objective which may lead to slower convergence rates.

Titsias and Ruiz [19] introduced *unbiased implicit variational inference* (UIVI) as an alternative method to SIVI. Like SIVI, UIVI posits a flexible semi-implicit variational family to approximate the true posterior distribution. However, unlike SIVI, UIVI uses an unbiased estimator of the ELBO gradient which theoretically should lead to faster convergence of the optimization. While the experiments conducted by Titsias and Ruiz [19] show promising results, theoretical analyses and guarantees are missing in their paper.

In this project, we discuss our attempts at analyzing the theoretical performance of UIVI. Although we are unable to make much progress due to time constraints, we at least suggest possible directions for continuing what is started in this work that are most likely to be fruitful. TODO

This report is organized as follows: Section 2.2 provides a brief overview of UIVI; Section 2.3 introduces other notation used in this report; Section 2.4 discusses our characterization of the quality of the UIVI approximation for a particular choice of the variational conditional and mixing distributions; Section 2.5 describes our attempts of analyzing the variance of the ELBO gradient in UIVI; and Section 2.6 summarizes our findings and the possible directions of future work.

2.2 Background

UIVI approximates the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ using a variational distribution of the form

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

where the variational conditional $q_\theta(\mathbf{z}|\boldsymbol{\varepsilon})$ is required to be reparameterizable and explicit but with the dependency on $\boldsymbol{\theta}$ to be arbitrarily complex, and where $q(\boldsymbol{\varepsilon})$ is some fixed auxiliary distribution. UIVI also requires that the log-gradient $\nabla_{\mathbf{z}} \log q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon})$ can be evaluated. Under these assumptions, the gradient of the evidence lower bound (ELBO) in UIVI can be written as

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

The first term in the gradient can be estimated using Monte Carlo samples from $q(\boldsymbol{\varepsilon})$ and $q(\mathbf{u})$. For the second term, as $\nabla_{\mathbf{z}} \log q_{\boldsymbol{\theta}}(\mathbf{z})$ may not necessarily be explicit, UIVI instead estimates its equivalent identity

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

using Monte Carlo samples drawn from the reverse conditional via Markov chain Monte Carlo (MCMC) methods. To avoid needing a burn-in phase, UIVI reuses samples from estimating the outer expectation as initial points in the MCMC.

2.3 Notation

We briefly introduce other notation used in this report in addition to the notation introduced in Section 2.2. Let λ denote the Lebesgue measure on the specified support of the mixing variable $\boldsymbol{\varepsilon}$. Let $\mathcal{B}$ denote the Borel σ-algebra of $\mathbb{R}^d$ where d is the dimension of the latent variable $\mathbf{z}$. Let ϕ_σ denote the density of a $\mathcal{N}(0, \sigma^2\mathbf{I}_d)$ distribution. Note that we often denote both a distribution and its density by q and let the context distinguish between which one is referred to. For densities f and q, let $f * q(z) = \int_{\mathcal{X}} f(z-x)q(x)dx$ denote the convolution of f and q.

2.4 Quality of approximation

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 aim to address this limitation of the paper by discussing particular cases of when UIVI is able or unable to theoretically recover the true distribution. We first show a simple example where the true posterior distribution is not in the variational family due to misspecification. We then show that under certain assumptions and choices of the reparameterization and mixing distribution, UIVI is able to approximate the true posterior arbitrarily closely.

2.4.1 Normal-normal 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 conditional and mixing distributions are still important in realizing the potential of UIVI.

Consider a univariate Gaussian posterior distribution $p(z|x)$ with mean $\mu_{Z|X}$ and known variance $\sigma^2_{Z|X}$. This is 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 the variational conditional reparameterized as

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

where $\boldsymbol{\varepsilon}$ and u are independent standard normal random variables and θ is the variational parameter to optimize. This reparameterization corresponds to $\mu_\theta(\boldsymbol{\varepsilon}) = \theta + \boldsymbol{\varepsilon}$ being a simple additive function. Then it is easy to see that

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

In other words, our variational family is the set of univariate Gaussian distributions $\left\{\mathcal{N}(\theta, \sigma^2_{Z|X} + 1) : \theta \in \mathbb{R}\right\}$. The true posterior distribution $\mathcal{N}(\mu_{Z|X}, \sigma^2_{Z|X})$ 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 well-illustrates that relying on the hierarchical structure alone is insufficient for setting up a flexible variational family.

2.4.2 Flexible variational family

We now show that under certain assumptions and a particular choice of the conditional and mixing distributions, the induced UIVI variational family can be set to include a distribution that approximates the true distribution 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 $\mathbf{z} \in \mathbb{R}^d$.

Consider a multivariate mixing distribution of the form $q(\boldsymbol{\varepsilon}) = \prod_{i=1}^{d} q(\varepsilon_i)$ where $q(\varepsilon_i) = \text{Unif}(0,1)$ for $i = 1, \ldots, d$. Let the variational conditional $q_{\boldsymbol{\theta},\sigma}(\mathbf{z}|\boldsymbol{\varepsilon})$ be multivariate Gaussian with mean $\mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})$ and covariance matrix $\Sigma_\sigma = \sigma^2 \mathbf{I}_d$ where $\mu_{\boldsymbol{\theta}} : [0,1]^d \to \mathbb{R}^d$ is some arbitrarily complex function. Note that we keep the variational parameters $\boldsymbol{\theta}$ and σ separate for reasons that will be clear shortly. This distribution is reparameterizable through the form

$$\mathbf{z} = h_{\boldsymbol{\theta},\boldsymbol{\sigma}}(\mathbf{u}; \boldsymbol{\varepsilon}) = \mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) + \Sigma_\sigma^{-\frac{1}{2}} \mathbf{u}$$

where $\mathbf{u} \sim \mathcal{N}(0, \mathbf{I}_d)$. Furthermore, the log-density of this distribution and its gradient are both evaluatable and so this distribution satisfies the UIVI requirements. Note that this choice of $h_{\boldsymbol{\theta},\sigma}$ resembles the NL-LVM studied by Plummer et al. [12], which allows us to apply their results with slight modifications.

To study the approximation capabilities of the NL-LVM, the key insight of Plummer et al. [12] is that the marginal density of the reparameterized variable induced by the NL-LVM has the form of a convolution with a Gaussian kernel. This insight applies to our UIVI model as well where $q_{\boldsymbol{\theta},\sigma}(\mathbf{z})$ can be rewritten as

$$\begin{aligned} q_{\boldsymbol{\theta},\sigma}(\mathbf{z}) &= \int_{[0,1]^m} q_{\boldsymbol{\theta},\sigma}(\mathbf{z}|\boldsymbol{\varepsilon}) q(\boldsymbol{\varepsilon}) \lambda(d\boldsymbol{\varepsilon}) \\ &= \int_{[0,1]^m} \phi_\sigma \left(\mathbf{z} - \mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\right) \lambda(d\boldsymbol{\varepsilon}) \\ &= \int_{\mathbb{R}^d} \phi_\sigma \left(\mathbf{z} - \mathbf{t}\right) \nu_{\mu_{\boldsymbol{\theta}}}(d\mathbf{t}) \end{aligned}$$

with $\nu_{\mu_{\boldsymbol{\theta}}}(B) = \lambda\left(\mu_{\boldsymbol{\theta}}^{-1}(B)\right)$, $B \in \mathcal{B}$, being the image measure of λ under $\mu_{\boldsymbol{\theta}}$. Using the approximation property of convolutions, we characterize the relationship between the true posterior distribution 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 $q_{\boldsymbol{\theta},\sigma}(\mathbf{z}|\boldsymbol{\varepsilon})$. Suppose that $\mu_{\boldsymbol{\theta}}(\mathbf{t}) = F^{-1}_{\mathbf{z}|\mathbf{x}}(\mathbf{t})$ for all $\mathbf{t} \in [0,1]^d$. Then $p(\mathbf{z}|\mathbf{x}) \in \mathcal{Q}_0$ (the limiting family as $\sigma \to 0$).*

Proof. If $\mu_{\boldsymbol{\theta}}(\mathbf{t}) = F^{-1}_{\mathbf{z}|\mathbf{x}}(\mathbf{t})$ for all $\mathbf{t} \in [0,1]$, then $q_{\boldsymbol{\theta},\sigma}(\mathbf{z}) = \phi_\sigma * p(\mathbf{z}|\mathbf{x})$ by the change of variables argument above. The result then immediately follows from the consistency property of convolutions that $\int_{\mathcal{Z}} |\phi_\sigma * p(\mathbf{z}|\mathbf{x}) - p(\mathbf{z}|\mathbf{x})| \, d\mathbf{z} \to 0$ as $\sigma \to 0$ [4]. □

Proposition 1 says that if the inverse CDF or quantile function $F^{-1}_{\mathbf{z}|\mathbf{x}}$ is in the set of functions $\{\mu_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ that can be modeled by $\mu_{\boldsymbol{\theta}}$, then the true posterior $p(\mathbf{z}|\mathbf{x})$ is a limiting member of the sequence of best approximations by the variational family as the bandwidth σ of the Gaussian kernel shrinks to zero. While this result also implies that the true posterior is not in the variation family for any $\sigma > 0$, it suggests 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 is not addresssed by Proposition 1 is the question of whether we are able to achieve the best approximation for a given σ in practice. This depends on the functional form of $\mu_{\boldsymbol{\theta}}$ being flexible enough such that $F^{-1}_{\mathbf{z}|\mathbf{x}} \in \{\mu_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and whether our optimization procedure is able to identify the correct $\boldsymbol{\theta}$ such that $\mu_{\boldsymbol{\theta}} = F^{-1}_{\mathbf{z}|\mathbf{x}}$. At least with regards to the existence of sufficiently flexible $\mu_{\boldsymbol{\theta}}$, a universal approximation theorem (Theorem 3.1) in the work by Daniels and Velikova [3] states that for any continuous monotone nondecreasing function on a compact subset of $\mathbb{R}^d$, there exists a feedforward neural network with at most d

hidden layers such that the pointwise error when approximating the function is within a desired tolerance.

We note that the work by Plummer et al. [12] includes other results that may be of interest–namely, a posterior contraction result for NL-LVM. However, the results do not appear to generalize easily to SIVI and UIVI. This is because the implicit model in NL-LVM is on the observable variable $\mathbf{x}$ as opposed to the latent variable $\mathbf{z}$ in SIVI and UIVI. Ultimately, it is also probably more interesting to understand the convergence of the UIVI optimization procedure to the true posterior distribution $p(\mathbf{z}|\mathbf{x})$ for a fixed $\mathbf{x}$ rather than how the approximation changes with increasing amounts of observations.

2.4.3 Directions of future work for quantifying quality

As mentioned in the previous section, Proposition 1 does not address the question of how *good* the quality of the UIVI approximation is, and the other results of Plummer et al. [12] for NL-LVM do not generalize to UIVI. We briefly discuss some possible approaches that may allow quantifying the quality of the convolution family approximation based on ideas from the previous section. We leave further exploration of these ideas for future work due to time constraints on this project.

One idea is to characterize the approximation quality in terms of the error from the Gaussian convolution for a given $\sigma > 0$. Such a result would quantify the error in the best approximation possible for a particular σ. The structure of such a result would look similar to those presented for kernel density estimation ([4, Section 9.5, Chapter 11]) where the ratio of expected empirical error and the theoretical error for some bandwidth is upper bounded by some constant that depends on the true distribution. Being able to do so would likely require additional assumptions on the smoothness of $p(\mathbf{z}|\mathbf{x})$ similar to those made by Plummer et al. [12] on the true density in their analysis of NL-LVM.

Another idea is to characterize the quality in terms of the error from estimating the quantile function with $\mu_{\boldsymbol{\theta}}$ for a particular $\boldsymbol{\theta}$. Such a result would be necessary in obtaining a convergence rate of the UIVI optimization procedure for this family. However, this approach appears to be more challenging than the previous idea in that smoothness assumptions will need to be made on the inverse CDF and the form of the function $\mu_{\boldsymbol{\theta}}$ will need to be specified. TODOA characterization of the convergence rate to the best approximation would likely require some combination of these two ideas.

2.5 Variance of gradient

It is known that the performance of standard SIVI suffers when the dimensionality d of the latent variable increases with Molchanova et al. [11] citing the reason being that the variational distribution $q_{\boldsymbol{\theta}}(\mathbf{z})$ is approximated with $K+1$ Gaussian distributions (where K is the mixture size) when estimating the ELBO gradient. As the number of dimensions increases, the mixture size K needs to increase exponentially in order to maintain a reasonable approximation. In this section, we attempt to show that UIVI shares a similar problem where the ELBO gradient is estimated using a finite mixture of Gaussians. TODO

An unbiased Monte Carlo estimator of the ELBO gradient in UIVI is given by

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

where $\mathbf{u}_{1:n}$ are n i.i.d. samples from $\mathcal{N}(0,\mathbf{I}_d)$, $\boldsymbol{\varepsilon}_{1:n}$ are n i.i.d. samples from $q(\boldsymbol{\varepsilon})$, and $\boldsymbol{\varepsilon}'_{1:m}$ are m MCMC samples drawn from $q(\boldsymbol{\varepsilon}|\mathbf{z})$. It is said that $n=1$ is used in practice and m is kept small ($m=5$ in the experiments by Titsias and Ruiz [19]). Intuitively, the problem of approximating the ELBO gradient with a Gaussian mixture described above appears in the second term where if the variational conditional $q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon})$ is taken to be Gaussian, then each sampled $\boldsymbol{\varepsilon}'_j$ corresponds to a single sampled Gaussian distribution. From this perspective, the number of samples m plays a similar role to the mixture size K in SIVI. However, although this interpretation of sampling Gaussians holds, it is not immediately obvious that this necessarily impacts the performance of UIVI in high dimensions as the contribution of each sampled Gaussian to the

gradient estimate is through their score function. Therefore, we aim to better understand the properties of this gradient estimator by studying its variance.

In particular, it is of interest to understand how the variance of this estimator behaves as the variational distribution approaches the true posterior distribution. If the true distribution lives in a high dimensional space, we may intuitively expect that as our variational distribution conforms to the true distribution, our sampled Gaussian distributions may also become more varied. Such an issue would pose a problem for convergence, and so we attempt to analyze the variance in this setting. In our analysis, we assume that $\boldsymbol{\theta}$ is fixed and is exactly or close to the value such that $q_{\boldsymbol{\theta}}(\mathbf{z}) = p(\mathbf{z}|\mathbf{x})$. As mentioned above, we assume that $q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon})$ is Gaussian.

Deriving the variance of the gradient estimator is not straightforward due to the various sampled components and the potentially complex mappings in $h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon})$. To simplify the analysis, we consider the case $n = 1$. We also follow the idea underlying Rao-Blackwellization [13] and define the functions

$$\begin{aligned} L_{\boldsymbol{\theta}}(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) &= \widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) \ , \\ L_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}'_{1:m}) &= \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}} \left[L_{\boldsymbol{\theta}}(\mathbf{u}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}'_{1:m}) | \boldsymbol{\varepsilon}'_{1:m} \right] \ . \end{aligned}$$

It then follows that

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

and so

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

The function $L_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}'_{1:m})$ is easier to analyze and examining its variance is equivalent to examining a lower bound of the variance of the full gradient estimator. The function has the form

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

Note that the first term is constant. Furthermore, the second term retains the interpretation of sampling individual Gaussians through sampling $\boldsymbol{\varepsilon}'_j$.

As the gradient of the ELBO is a vector, we consider the variance of the estimator for the i-th parameter given by

$$\begin{aligned} \mathrm{Var}(L_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}'_{1:m})_i) &= \frac{1}{m} \mathrm{Var} \left(\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}} \left[\nabla_{\mathbf{z}} \log q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon}') \big|_{\mathbf{z}=h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon})} \nabla_{\theta_i} h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon}) \right] \right) \\ &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\varepsilon}'} \left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}} \left[\nabla_{\mathbf{z}} \log q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon}') \big|_{\mathbf{z}=h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon})} \nabla_{\theta_i} h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon}) \right]^2 \right] \end{aligned}$$

using the fact that

$$\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'} \left[\nabla_{\mathbf{z}} \log q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon}') \big|_{\mathbf{z}=h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon})} \nabla_{\theta_i} h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon}) \right] = \mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}'} \left[\nabla_{\theta_i} \log q_{\boldsymbol{\theta}}(\mathbf{z}|\boldsymbol{\varepsilon}') \big|_{\mathbf{z}=h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon})} \right] \approx 0$$

under the assumption that $q_{\boldsymbol{\theta}}(\mathbf{z}) \approx p(\mathbf{z}|\mathbf{x})$. We can simplify the variance expression further by plugging in the Gaussian score function, which gives

$$\begin{aligned} \mathrm{Var}(L_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}'_{1:m})_i) &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\varepsilon}'} \left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}} \left[\left(\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')^{-1} (h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon}) - \mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')) \right)^\top \nabla_{\theta_i} h_\theta(\mathbf{u};\boldsymbol{\varepsilon}) \right]^2 \right] \\ &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\varepsilon}'} \left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}} \left[\left(\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')^{-1} ((\mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) + \Sigma_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon})\mathbf{u}) - \mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')) \right)^\top \nabla_{\theta_i} h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon}) \right]^2 \right] \\ &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\varepsilon}'} \left[\mathbb{E}_{\mathbf{u},\boldsymbol{\varepsilon}} \left[\left(\Sigma_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')^{-1} (\mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) - \mu_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')) + \Sigma_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}')^{-1} \Sigma_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) \mathbf{u} \right)^\top \nabla_{\theta_i} h_{\boldsymbol{\theta}}(\mathbf{u};\boldsymbol{\varepsilon}) \right]^2 \right] \ . \end{aligned}$$

At this point, we find that it is difficult to proceed without further assumptions on $\Sigma_{\boldsymbol{\theta}}$ and $\mu_{\boldsymbol{\theta}}$. One idea may be to impose a smoothness assumption on $\mu_{\boldsymbol{\theta}}$ such that the difference in evaluations at $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}'$ can be bounded. Another idea is to assume that $\mu_{\boldsymbol{\theta}_1}$ and $\Sigma_{\boldsymbol{\theta}_2}$ do not share parameters and so $\nabla_{\theta_i} h_{\boldsymbol{\theta}}(\mathbf{u}; \boldsymbol{\varepsilon})$ can be simplified in terms of only the function that θ_i is used in.

Though we are unable to progress further due to time constraints on this project, we can reason a few (perhaps obvious) conclusions from the above derivations. The lower bound to the variance of the original gradient estimator decreases as the number of samples m increases, which decreases the variation due to sampling the individual Gaussians. Also, the effect of the dimension of $\mathbf{z}$ can only come into play in this Gaussian sampling through the (squared) score. If the score function depends on d at a rate greater than that of $O(\sqrt{d})$, then m needs to scale faster than linear with d in order to maintain the lower bound. Otherwise, m can compensate for increases in dimension by scaling linearly with d.

2.6 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] Hennie Daniels and Marina Velikova. Monotone and partially monotone neural networks. *IEEE Transactions on Neural Networks*, 21(6):906–917, 2010.

[4] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

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

[6] 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.

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

[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.