

Abstract

Transformation-based methods have been an attractive approach in non-parametric inference for problems such as unconditional and conditional density estimation due to their unique hierarchical structure that models the data as flexible transformation of a set of common latent variables. More recently, transformation-based models have been used in variational inference (VI) to construct flexible implicit families of variational distributions. However, their use in both non-parametric inference and variational inference lacks theoretical justification. We provide theoretical justification for the use of non-linear latent variable models (NL-LVMs) in non-parametric inference by showing that the support of the transformation induced prior in the space of densities is sufficiently large in the $L_1$ sense. We also show that, when a Gaussian process (GP) prior is placed on the transformation function, the posterior concentrates at the optimal rate up to a logarithmic factor. Adopting the flexibility demonstrated in the non-parametric setting, we use the NL-LVM to construct an implicit family of variational distributions, deemed GP-IVI. We delineate sufficient conditions under which GP-IVI achieves optimal risk bounds and approximates the true posterior in the sense of the Kullback–Leibler divergence. To the best of our knowledge, this is the first work on providing theoretical guarantees for implicit variational inference.

1 Introduction

Transformation-based models are a powerful class of latent variable models, which rely on a hierarchical generative structure for the data. In their simplest form, these models have the following structure

$$y_i = \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

$$x_i \overset{iid}{\sim} g,$$

for $i = 1, \ldots, n$, where $y_i \in \mathbb{R}$ is a real-valued observed variable, $\mu$ is the ‘transformation’ function, $x_i$ is a latent (unobserved) variable underlying $y_i$, $g$ is a known density of the latent data (e.g., uniform or standard normal), and we include a Gaussian measurement error with variance $\sigma^2$. For simplicity in exposition, we consider a very simple case to start but one can certainly include multivariate $x_i$ and $y_i$ and other elaborations.

Model (1) and its elaborations include many popular methods in the literature. If we choose a Gaussian process (GP) prior for the function $\mu$, then we obtain a type of GP Latent Variable Model (GP-LVM) (Lawrence, 2004, 2005; Lawrence & Moore, 2007). We can also obtain kernel mixtures as a special case; for example, by choosing a discrete distribution for $g$. The extremely popular Variational Auto-Encoder (VAE) is based on choosing a deep neural network for $\mu$, and then obtaining a particular variational approximation relying on a separate encoder and decoder neural network (Kingma & Welling, 2013). Refer also to the non-linear latent variable model (NL-LVM) framework of (Kundu & Dunson, 2014) for a nonparametric Bayesian perspective on models related to (1).

Providing theoretical justification for ‘transformation’ based models of the form in (1) rests on the answers to the following two questions: 1) Can this framework be used to approximate any density with an arbitrarily high degree of accuracy? 2) Does the accuracy improve with sample size as the optimal rate for density estimation or conditional density estimation (given fixed covariates) problems?

These types of questions have been answered elegantly
for many nonparametric Bayes and frequentist density estimation methods, especially for the models constructed via model (1) with a discrete distribution $g$ of the latent variable. For example, Dirichlet process mixture models (DPMMs) have been very widely applied (Escobar & West, 1995; Ferguson, 1973, 1974; MacEachern, 1999; Müller et al., 1996) and studied in terms of their optimality properties asymptotically (Ghosal et al., 1999, 2000; Ghosal & van der Vaart, 2007; Krujier et al., 2010).

When using a continuous distribution $g$, model (1) leads to a specific class of continuous transformation-based model such as the NL-LVM models. Here a GP prior is a natural choice for the unknown transformation (Dasgupta et al., 2017; Kundu & Dunson, 2014; Lenk, 1988, 1991; Tokdar, 2007; Tokdar et al., 2010). These models can be written as Gaussian convolution of a continuous mixing measure. Unfortunately the algorithms developed for discrete mixing measures are not readily adaptable to their continuous analogs. The alternative approach uses Markov chain Monte Carlo methods, which come with theoretical guarantees, but suffer from computational instability owing to a lack of conjugacy. This instability propagates through the posterior distribution of the unknown transformation requiring expert parameter tuning and vigilance for guaranteed performance. To mitigate some of these issues associated with a full-blown MCMC, approximate Bayesian methods including the variational inference (VI) are proposed (Titsias & Lawrence, 2010). The success of VI depends largely on two things: 1) the flexibility of the variational family and 2) the algorithm used to perform the optimization.

Development of flexible variational families using the reparametrization trick (Figurnov et al., 2018; Jankowiak & Obermeyer, 2018; Kingma et al., 2015; Kingma & Welling, 2013) have emerged as a powerful idea over the last decade and continues to flourish, often in parallel with latest developments in generative deep-learning methods. While the overarching goal of this trick is to find unbiased estimates of the gradient of the objective function (evidence lower bound in variational inference), one cannot but notice its connection with non-linear latent variable methods. A similar idea is explored as Implicit variational inference (Huszár, 2017; Shi et al., 2017) to construct an implicit distribution, a distribution that cannot be analytically specified but can be sampled from. Such a construction brings in certain computational challenges stemming from density ratio estimation. More recently, implicit VI was extended to semi-implicit VI (Molchanov et al., 2019; Titsias & Ruiz, 2019; Yin & Zhou, 2018) which avoids density ratio estimation by using a semi-implicit variational distribution $q_{\phi}(\theta) = \int q(\theta | g_{\phi}(u))q(u)du$ where the density $q(z | g_{\phi}(u))$ corresponds to a transformation-based model with transformation $g_{\phi}$ — typically taken to be a neural network with parameters $\phi$. Although VI approaches have shown significant improvements in computational speed their theoretical properties are largely a mystery.

Thus the aim of this work is to address one of the fundamental questions in latent variable transformation methods, namely, under what conditions are these methods “flexible” enough? The central idea is to recognize that such models can be written as Gaussian convolution of a continuous mixing measure. Such a construction serves as a flexible family for inference in either the latent variable semi-parametric density estimation setting or density estimation using implicit variational inference. The traditional approach to the density estimation problem is through the use of discrete mixtures, whose approximation properties have been well-studied (Ghosal et al., 1999, 2000; Ghosal & van der Vaart, 2007; Krujier et al., 2010). However, the well-known transformation based methods such as GP-LVM and IVI, are based off of continuous mixtures rather than discrete ones. Unfortunately, the existing tools for studying properties of these models for discrete mixtures do not readily extend to the continuous mixture case which requires different techniques to quantify the accuracy of approximation. Because of this, there has been, to the best of our knowledge, no results pertaining to properties of continuous mixture models in either the non-parametric or variational settings. There are no results that specify for which class of functions $\mathcal{F}$ these continuous mixture models are capable of estimating the true data distribution $f_0 \in \mathcal{F}$ arbitrarily well. Similarly, there are no results pertaining to risk bounds or convergence properties of any implicit variational inference framework. The closest related works in either case are those that address these questions for discrete mixture models. Lastly, we have chosen to exclude detailed empirical illustration, but provide a sketch of the algorithm in the supplementary material, as there is a relatively large body of existing work delineating algorithms and demonstrating the empirical performance of these continuous mixture models in both the non-parametric setting using GP-LVM (Ferris et al., 2007; Lawrence, 2004, 2005; Lawrence & Moore, 2007) and the variational setting using IVI (Huszár, 2017; Molchanov et al., 2019; Shi et al., 2017; Titsias & Ruiz, 2019; Yin & Zhou, 2018).

A summary of our contributions. Our results are the first to provide a concrete theoretical framework for transformation-based models widely used in Bayesian inference and machine learning. By establishing a connection between NL-LVM with implicit family of dis-
tributions, we provide statistical guarantees for implicit variational inference. Motivated by our findings, transformation-based models have the potential to provide machine learning with a rich class of implicit variational inference methods that come with strong theoretical guarantees.

We close the section by defining some notations in §1.1 used throughout the paper. In §2 we present an overview of the NL-LVM model as well as several properties of the model. In section §3 we discuss our two main results for non-parametric inference using NL-LVM. In §4 we introduce GP-IVI. We then show that the KL divergence between the variational posterior and the true posterior is stochastically bounded and argue why this is optimal from a statistical perspective. Inspired by Yang et al. (2020), we additionally present parameter risk bounds of a version of implicit variational inference, which we term as α-GP-IVI which is obtained by raising the likelihood to a fractional power $\alpha \in (0, 1)$. §3 we discuss our two main results for non-parameteric inference using NL-LVM model to the true density under mild regularity conditions. A review of the necessary background material for this section can be found in the supplementary file section S1.

2.1 The NL-LVM model

Suppose we have IID observations $Y_i \in \mathbb{R}$ for $i = 1, \ldots, n$ with density $f_0 \in \mathcal{F}$, the set of all densities on $\mathbb{R}$ absolutely continuous with respect to the Lebesgue measure $\lambda$. We consider a non-linear latent variable model

$$
Y_i = \mu(\eta_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \ i = 1, \ldots, n
$$

$$
\mu \sim \Pi_\mu, \quad \sigma \sim \Pi_\sigma, \quad \eta_i \sim U(0, 1),
$$

(2)

where $\eta_i$‘s are latent variables, $\mu \in C[0, 1]$ is a transfer function relating the latent variables to the observed variables and $\epsilon_i$ is an idiosyncratic error. Marginalizing out the latent variable, we obtain the density of $y$ conditional on the transfer function $\mu$ and scale $\sigma$

$$
f(y; \mu, \sigma) \overset{\text{def}}{=} f_{\mu, \sigma}(y) = \int_0^1 \phi_{\sigma}(y - \mu(x)) dx.
$$

(3)

Remark 2.1. While $\mu$ and $\eta$ are not identifiable in (2), our goal is to estimate $f_0$ using $f_{\mu, \sigma}$ which is an identifiable quantity itself. The flexibility of the induced model is guaranteed via the GP prior over the transformation function $\mu$ without the need to identify the corresponding latent variable $\eta$. The presence of the latent variable $\eta$ simply ensures flexibility of the induced density and allows for straightforward computation via Gibbs sampler or variational techniques.

It is not immediately clear whether the class of densities $\{f_{\mu, \sigma}\}$ encompasses a large subset of the density space. The following intuition relates the above class with continuous convolutions which plays a key role in studying theoretical properties for models related to NL-LVMs. Within the support of a continuous density $f_0$, its cumulative distribution function $F_0$ is strictly monotone and hence has an inverse $F_0^{-1}$ satisfying $F_0^{-1}(t) = t$ for all $t \in \text{supp}(f_0)$. Now letting $\mu_0(x) = F_0^{-1}(x)$, one obtains $f_{\mu_0, \sigma}(y) = \phi_{\sigma} * f_0$, the convolution of $f_0$ with a normal density having mean 0 and standard deviation $\sigma$. This provides a way to approximate $f_0$ by the NL-LVM with optimal approximation accuracy. We summarize the approximation result in section 2.3.

Let $\lambda$ denote the Lebesgue measure on $[0, 1]$ and denote the Borel sigma-field of $\mathbb{R}$ by $\mathcal{B}$. For any measurable function $\mu : [0, 1] \to \mathbb{R}$, let $\nu_\mu$ denote the induced measure on $(\mathbb{R}, \mathcal{B})$, then, for any Borel measurable set $B$, $\nu_\mu(B) = \lambda(\mu^{-1}(B))$. By the change of variable
The smoothness assumption in the log scale will be in Kruijer et al. (2010), while in our case a global assumption on the local smoothness appeared providing a key piece in managing the KL-divergence be-

\[ \exists a constant \ F_2 \ \text{Assumption} \]

bounded away from 0 on \( [a, b] \), and that there exists some interval \( (a, b) \subset [0,1] \) such that \( f_0 \) is non-decreasing on \( [0, a] \), bounded away from 0 on \( [a, b] \) and non-increasing on \( [b, 1] \).

Assumption F2 guarantees that for every \( \delta > 0 \), there exists a constant \( C > 0 \) such that \( f_0 \ast \phi_\sigma \geq C f_0 \) for every \( \sigma < \delta \). Also see Ghosal et al. (1999) for similar assumption in density estimation.

2.3 Approximation property

As mentioned above, the flexibility of \( f_{\mu, \sigma} \) comes from a large class of the induced density measure \( \nu_{\mu} \). Now we quantify the approximation of \( f_{\mu, \sigma} \) to the true \( f_0 \) by utilizing its equivalent form as a convolution with a Gaussian kernel. It is well known that the convolution \( \phi_\sigma \ast f_0 \) can approximate \( f_0 \) arbitrary closely as the bandwidth \( \sigma \to 0 \). For Hölder-smooth functions, the order of approximation can be characterized in terms of the smoothness. If \( f_0 \in C^\beta[0, 1] \) with \( \beta \leq 2 \), the standard Taylor series expansion guarantees that \( \| \phi_\sigma \ast f_0 - f_0 \|_\infty = O(\sigma^\beta) \). However, for \( \beta > 2 \), it requires higher order kernels for the convolution to remain the optimal error (Devroye, 1992; Wand & Jones, 1994).

Kruijer et al. (2010) proposed an iterative procedure to construct a sequence of functions \( \{f_j\}_{j \geq 0} \) by

\[ f_{j+1} = f_0 - \triangle_\sigma f_j, \quad \triangle_\sigma f_j = \phi_\sigma \ast f_j - f_j, \quad j \geq 0. \]

We define \( f_\beta = f_j \) with integer \( j \) such that \( \beta \in (2j, 2j + 2] \). Under such construction, for \( f_0 \in C^\beta[0, 1] \) the convolution \( \phi_\sigma \ast f_\beta \) preserves the optimal error \( O(\sigma^\beta) \) (Lemma 1 in Kruijer et al. (2010)). We state a similar result in the following.

**Proposition 2.1.** For \( f_0 \in C^\beta[0, 1] \) with \( \beta \in (2j, 2j + 2] \) satisfying Assumptions F1 and F2, for \( f_\beta \) defined as from the iterative procedure (6) we have

\[ \| \phi_\sigma \ast f_\beta - f_0 \|_\infty = O(\sigma^\beta), \]

and

\[ \phi_\sigma \ast f_\beta(x) = f_0(x)(1 + D(x)O(\sigma^\beta)), \]

where

\[ D(x) = \sum_{i=1}^{r} c_i |l_j(x)|^\delta + c_{r+1}, \]

for non-negative constants \( c_i, i = 1, \ldots, r + 1 \), and for any \( x \in [0, 1] \).

The proof can be found in the supplementary file section S2.2. The ability to represent the model in terms proportional to true density plays an important role in bounding the KL-divergence between \( f_{\mu, \sigma} \) and \( f_0 \).

**Remark 2.2.** The approximation result can be extended to the isotropic \( \beta \)-H"older space \( C^\beta[0, 1]^d \) under similar regularity assumptions. The extended approximation result can be applied to more general cases.

3 Posterior inference for NL-LVM

Most of the existing literature on non-parametric Bayesian approaches to the density estimation problem are centered around DP mixture priors (Ferguson, 1973, 1974), which are simply transformation-based models with a discrete distribution for the latent variables. On the other hand, the theoretical properties of continuous transformation-based models remain largely unknown.

In this section, we provide theoretical results for posterior inference of the transformation-based model for
unconditioned density estimation in the context of NL-LVM. Our results are two-fold: (1) We first show that a large class of transfer function \( \mu \) leads to \( L_1 \) large support of the space of densities induced by the NL-LVM; (2) We obtain the optimal frequentist rate up to a logarithmic factor under standard regularity conditions on the true density using the transformation-based approach with induced GP priors.

3.1 \( L_1 \) large support

One can induce a prior \( \Pi \) on \( F \) via the mapping \( f_{\mu,\sigma} \) by placing independent priors \( \Pi_{\mu} \) on \( C[0,1] \) and \( [0,\infty) \) respectively, as \( \Pi = (\Pi_{\mu} \otimes \Pi_{\sigma}) \circ f_{\mu,\sigma}^{-1} \). Kundu & Dunson (2014) assumes a Gaussian process prior with squared exponential covariance kernel on \( \mu \) and an inverse-gamma prior on \( \sigma^2 \). Given the flexibility of \( f_{\mu,\sigma} \) upon the choices of \( \mu \), placing a prior on \( \mu \) supported on the space of continuous functions \( C[0,1] \) without further restrictions is convenient and Theorem 3.1 assures us that this specification leads to large \( L_1 \) support on the space of densities.

Suppose the prior \( \Pi_{\mu} \) on \( \mu \) has full sup-norm support on \( C[0,1] \) so that \( \Pi_{\mu}(\|\mu - \mu^*\|_{\infty} < \epsilon) > 0 \) for any \( \epsilon > 0 \) and \( \mu^* \in C[0,1] \), and the prior \( \Pi_{\sigma} \) on \( \sigma \) has full support on \( [0,\infty) \). If \( f_0 \) is compactly supported, so that the quantile function \( \mu_0 \in C[0,1] \), then it can be shown that under mild conditions, the induced prior \( \Pi_{\mu} \) assigns positive mass to arbitrarily small \( L_1 \) neighborhoods of any density \( f_0 \). We summarize the above discussion in the following theorem, with a proof provided in the section S2.3 of supplementary file.

**Theorem 3.1.** If \( \Pi_{\mu} \) has full sup-norm support on \( C[0,1] \) and \( \Pi_{\sigma} \) has full support on \( [0,\infty) \), then the \( L_1 \) support of the induced prior \( \Pi \) on \( F \) contains all densities \( f_0 \) which have a finite first moment and are non-zero almost everywhere on their support.

**Remark 3.1.** The conditions of Theorem 3.1 are satisfied for a wide range of Gaussian process priors on \( \mu \) (for example, a GP with a squared exponential or Matérn covariance kernel).

**Remark 3.2.** When \( f_0 \) has full support on \( \mathbb{R} \), the quantile function \( \mu_0 \) is unbounded near 0 and 1, so that \( \|\mu_0\|_{\infty} = \infty \). However, \( \int_0^1 |\mu_0(t)| \, dt = \int_{\mathbb{R}} |x| f_0(x) \, dx \), which implies that \( \mu_0 \) can be identified as an element of \( L_1[0,1] \) if \( f_0 \) has finite first moment. Since \( C[0,1] \) is dense in \( L_1[0,1] \), the previous conclusion regarding \( L_1 \) support can be shown to hold in the non-compact case too.

3.2 Posterior contraction results

Gaussian process priors have been widely used in non-parametric Bayesian inference as well as machine learning due to their modeling advantages and proper theoretical grounding (van der Vaart & van Zanten, 2007, 2008, 2009). Considering a Gaussian process as the transfer function over the latent variable, the transformation-based model essentially aligns with a Gaussian process latent variable model (GP-LVM) (Ferris et al., 2007; Lawrence, 2004, 2005; Lawrence & Moore, 2007). Theoretical work of GP-LVM such as Kundu & Dunson (2014) showed a KL large support of the induced prior process, and also showed the posterior consistency to the true density function. However a straightforward description of the space of densities induced by the proposed model is not clear. Additionally, the posterior contraction rate of the proposed model, an important property characterizing how fast the posterior distribution concentrates around the truth, is still unknown for finite data.

We now present the posterior contraction result for transformation-based model with NL-LVM. To that end, we first review its definition, more details are deferred to the supplementary file section S1. Given independent and identically distributed observations \( Y^{(n)} = (Y_1, \ldots, Y_n) \) from a true density \( f_0 \), a posterior \( \Pi_{\mu} \) associated with a prior \( \Pi \) on \( F \) is said to contract at a rate \( \epsilon_n \), if for a distance metric \( d_n \) on \( F \),

\[
\mathbb{E}_{f_0,\Pi_{\mu}}(d_n(f, f_0) > M \epsilon_n \mid Y^{(n)}) \to 0 \quad (8)
\]

for a suitably large integer \( M > 0 \). Unlike the treatment in discrete mixture models (Ghosal & van der Vaart, 2007) where a compactly supported density is approximated with a discrete mixture of normals, the main idea is to first approximate the true density \( f_0 \) by a Gaussian convolution with \( f_{\beta} \) defined as in (6), then allow the GP prior on the transfer function to appropriately concentrate around \( \mu_{\beta} \), the inverse c.d.f. of the defined \( f_{\beta} \). We first state our choices for the prior distributions \( \Pi_{\mu} \) and \( \Pi_{\sigma} \).

**Assumption P1** We assume \( \mu \) follows a centered and rescaled Gaussian process denoted by GP(0, \( cA \)), where \( A \) denotes the rescaled parameter, and assume \( A \) has density \( g \) satisfying for \( a > 0 \),

\[
C_1 a^p \exp(-D_1 a \log^q a) \leq g(a) \leq C_2 a^p \exp(-D_2 a \log^q a).
\]

**Assumption P2** We assume \( \sigma \sim IG(a_\sigma, b_\sigma) \).

Note that contrary to the usual conjugate choice of an inverse-gamma prior for \( \sigma^2 \), we have assumed an inverse-gamma prior for \( \sigma \). This enables one to have slightly more prior mass near zero compared to an inverse-gamma prior for \( \sigma^2 \), leading to the optimal rate of posterior convergence. Refer also to Kruijer et al. (2010) for a similar prior choice for the bandwidth of the kernel in discrete location-scale mixture priors.
for densities.

**Theorem 3.2.** If \( f_0 \) satisfies Assumptions \( \mathbf{F1} \) and \( \mathbf{F2} \) and the priors \( \Pi_\mu \) and \( \Pi_\sigma \) are as in Assumptions \( \mathbf{P1} \) and \( \mathbf{P2} \) respectively, the best obtainable rate of posterior contraction result in Ghosal et al. (2000) (See Theorem 1 in the supplementary file for details.) We first motivate by the flexibility we have demonstrated for transformation-based models in the non-parametric setting, we construct a flexible implicit variational family of distributions, deemed Gaussian process implicit variational inference (GP-IVI). We provide sufficient conditions under which GP-IVI achieves optimal risk bounds and approximates the true posterior in the sense of the Kullback–Leibler divergence. We begin by defining common terminology used throughout the section and defining GP-IVI.

### 4.1 Preliminaries

We consider IID observations \( Y_i \in \mathbb{R}^p \), for \( i = 1, \ldots, n \). Let \( \mathbb{P}_\theta^{(n)} \) be the distribution of the observations with parameter \( \theta \in \Theta \subset \mathbb{R}^d \) that admits a density \( p_\theta^{(n)} \) relative to the Lebesgue measure. Let \( \mathbb{P}_\theta \) denote the prior distribution of \( \theta \) that admits a density \( p_\theta \) over \( \Theta \). With a slight abuse of notation, we will use \( p(Y^{(n)} \mid \theta) \) to denote \( p_\theta^{(n)} \) and its density function. We adopt a frequentist framework and assume a true data generating distribution \( p_\theta^{(n)} \) and a true parameter \( \theta^* \). Denote the negative log prior \( U(\theta) = -\log p_\theta(\theta) \) and the log-likelihood ratio of \( Y_i \), for \( i = 1, \ldots, n \), by

\[
\ell_i(\theta, \theta^*) = \log \frac{p(Y_i \mid \theta)}{p(Y_i \mid \theta^*)}.
\]

We denote the first two moments of the log-likelihood by

\[
D(\theta^* \mid \theta) = -\mathbb{E}_\theta^{(n)}[\ell_1(\theta, \theta^*)], \quad \mu_2(\theta^* \mid \theta) = \mathbb{E}_\theta^{(n)}[\ell_1(\theta, \theta^*)]^2.
\]

Lastly denote the appropriate neighborhood around the true parameter \( \theta^* \),

\[
B_n(\theta^*, \varepsilon) = \{ \theta \mid D[p(Y^{(n)} \mid \theta^*)]p(Y^{(n)} \mid \theta) \leq n\varepsilon^2, V[p(Y^{(n)} \mid \theta^*)p(Y^{(n)} \mid \theta) \leq n\varepsilon^2] \}.
\]

### 4.2 Gaussian Process Implicit Variational Inference

Using the NL-LVM model, we can define the variational family of \( \theta \) conditioned on the latent variable \( \eta \), with parameters \( \mu \in C[0, 1] \) and \( \sigma \in (0, \infty) \),

\[
q_{\mu, \sigma}(\theta_i \mid \eta_i) = \phi_\sigma(\theta_i - \mu(\eta_i)) \quad \eta_i \sim U(0, 1), \quad i = 1, \ldots, d.
\]

Marginalizing over the latent \( \eta \) gives us the implicit variational distribution,

\[
q_{\mu, \sigma}(\theta) = \int_0^1 \phi_\sigma(\theta - \mu(\eta))d\eta.
\]

Together this defines the Gaussian process implicit variational inference (GP-IVI) family,

\[
\mathcal{Q}_{GP} = \left\{ q_{\mu, \sigma}(\theta) = \int_0^1 \phi_\sigma(\theta - \mu(\eta))d\eta \mid \mu \in C[0, 1], \sigma > 0 \right\}.
\]
4.3 Approximation Quality of GP-IVI

In this section, we show that KL divergence between the true posterior and its optimal GP-IVI approximation is $O_p(1)$. Using a simple example, we show that without further assumptions this bound cannot be improved. We begin the section with said example.

Consider the following one-dimensional Gaussian-Gaussian Bayesian model for inference of an unknown true mean $\theta^*$ using the model

$$Y_1, \ldots, Y_n \sim N(\theta, \sigma^2), \quad \theta \sim N(\mu_0, \sigma_0^2)$$

in which $\mu_0, \sigma_0, \sigma$ are all known. Let $Y_n, \mu_n, \sigma_n^2$ denote the sample mean, the posterior mean, and variance, respectively. Straight forward calculations show

$$D \left[ N(\theta^*, n^{-1}\sigma^2) \| N(\mu_n, \sigma_n^2) \right] \rightarrow \chi^2_1, \quad \text{weakly.}$$

Even in the simple case of a normal-normal model, we see that the KL divergence between the true data generating distribution and the true posterior does not converge weakly to 0 but instead converges weakly to a stochastically bounded random variable.

The $O_p(1)$ bound is achieved over a rather small subfamily of GP-IVI. Define the restricted Gaussian family

$$\Gamma_n = \{ N(\mu, \sigma^2 I_d) \mid \|\mu\|_2 \leq M, 0 \leq \sigma_n \leq \tau \leq c_0^{1/2} \sigma_n \},$$

and let $\mu_f$ denote the quantile function corresponding to $f \in \Gamma_n$. We define the corresponding small bandwidth convolution Gaussian (variational) family

$$Q_n = \left\{ q_{\mu, \sigma}(\theta) \mid q_{\mu, \sigma}(\theta) = \int_0^1 \phi_\sigma(\theta - \mu_f(\eta)) d\eta, \quad f \in \Gamma_n \right\}.$$ 

The following assumptions are required to show the $O_p(1)$ bound for the KL-divergence.

**Assumption B1** The true parameter $\theta^*$ satisfies $\|\theta^*\|_2 \leq M$.

**Assumption B2** The variance bound $\sigma_n$ satisfies $0 \leq \sigma_n \leq n^{-1/2} \leq c_0^{1/2} \sigma_n$, for all $n \geq 1$.

**Assumption B3** The quantities $D(\theta^*||\theta)$ and $\mu_2(\theta^*||\theta)$ are finite for all $\theta \in \mathbb{R}^d$.

**Assumption B4** The matrices of the second derivatives, $D(2)(\theta^*||\theta)$, $\mu_2(2)(\theta^*||\theta)$, $U(2)(\theta)$ exist on $\mathbb{R}^d$ and satisfy for any $\theta, \theta' \in \mathbb{R}^d$,

$$s_{\max} \left( D(2)(\theta^*||\theta) - D(2)(\theta^*||\theta') \right) \leq C \|\theta - \theta'\|_2^{\alpha_1},$$

$$s_{\max} \left( \mu_2(2)(\theta^*||\theta) - \mu_2(2)(\theta^*||\theta') \right) \leq C \|\theta - \theta'\|_2^{\alpha_2},$$

$$s_{\max} \left( U(2)(\theta) - U(2)(\theta') \right) \leq C \|\theta - \theta'\|_2^{\alpha_3},$$

for some $\alpha_1, \alpha_2, \alpha_3 > 0$. Here $s_{\max}$ denotes the maximum eigenvalue of the matrix.

**Assumption B5** $D(\theta^*||\theta) \geq C \|\theta - \theta^*\|_2$.

Assumption B1 is needed so that a normal distribution centered at the true parameter is contained in $\Gamma_n$. Assumptions B2-B4 are technical assumptions needed in order to achieve convergence of certain bounds used in the proof. Assumption B5 is a standard identifiability condition.

**Theorem 4.1.** Under assumptions B1 through B5 it holds that $m_n^*(\mathcal{Q}_n) = \min_{\theta \in \mathcal{Q}_n} \{ D[q'||p(\cdot|Y^n)] \}$ is bounded in probability with respect to the data generating distribution $p_{\mathcal{Q}_n}(\theta)$. Formally, given any $\varepsilon > 0$, there exists $M_\varepsilon, N_\varepsilon > 0$ such that for $n \geq N_\varepsilon$, we have

$$P_{\mathcal{Q}_n}(m_n^*(\mathcal{Q}_n) > M_\varepsilon) \leq \varepsilon.$$

Again, we provide a sketch of the proof below and provide a full proof in section S2.5 of the supplementary file. Under assumptions B1-B2, $q_n(\theta) = N(\theta;\theta^*, \sigma^2 + \sigma_n^2)$ belongs to $\mathcal{Q}_n$. By definition, $m_n^*(\mathcal{Q}_n) \leq D[q_n||p(\cdot|Y^n)]$. We show $D[q_n||p(\cdot|Y^n)]$ is $O_p(1)$ by showing that it is a sum of $O_p(1)$ terms. Letting $E_n$ denote the expectation with respect to $q_n$, $D[q_n||p(\cdot|Y^n)]$ can be broken into four parts $E_n[\log q_n]$, $\log m(Y^n)$, $E_n(U(\theta))$, and $E_n[\sum_{i=1}^n \ell_i(\theta, \theta^*)]$. The first term $E_n[\log q_n]$ is a constant, hence $O_p(1)$. Noting $E_n[m(Y^n)] = 1$, an application of Markov’s inequality shows that $\log m(Y^n)$ is $O_p(1)$. Taking a (multivariate) Taylor expansion of the functions $U(\theta)$, $D(\theta^*||\theta)$, and $\mu_2(\theta^*||\theta)$ about $\theta^*$ and applying assumption B4 and B5 gives us the bounds

$$C_\ell(\sigma^2 + \sigma_n^2) \leq E_n[D(\theta^*||\theta)] \leq C_u(\sigma^2 + \sigma_n^2),$$

$$E_n[\mu_2(\theta^*||\theta)] \leq C_2(\sigma^2 + \sigma_n^2),$$

$$E_n[U(\theta)] \leq C_1(\sigma^2 + \sigma_n^2).$$

Markov’s inequality shows that $U(\theta)$ is $O_p(1)$. It remains to show $E_n[\sum_{i=1}^n \ell_i(\theta, \theta^*)]$ is $O_p(1)$. Given $\varepsilon > 0$, choose $\delta = \left[ C_2/c_{\alpha_3}(1/\varepsilon C_2)^{1/2} \right]$. Applying Chebychev’s and Jensen’s inequalities together with (13) we have,

$$E_n \left[ \sum_{i=1}^n \ell_i(\theta, \theta^*) \right] \leq -C_u(1 + \delta) n(\sigma^2 + \sigma_n^2) \leq \frac{1}{C_1 \delta^2 n} \left( C_2 \frac{E_n[U(\theta)]}{E_n[D(\theta^*||\theta)]^2} \right) \leq \frac{1}{C_1 \delta^2 n} \frac{C_2}{\delta^2 n} \leq \frac{C_2}{\delta^2 n} \leq \varepsilon.$$
which shows $\mathbb{E}_n[\sum_{i=1}^{n} \ell_i(\theta, \theta^*)] = O_p(1)$. Combining the four bounds completes the proof.

### 4.4 \(\alpha\)-Variational Bayes Risk Bound for GP-IVI

In developing risk bounds for parameter estimation, we use a slight variation of the standard variational objective function for technical simplicity. \(\alpha\)-variational Bayes (\(\alpha\)-VB) (Yang et al., 2020) is a variational inference framework that aims to minimize the KL divergence between the variational density and the fractional posterior (Bhattacharya et al., 2019), defined as

$$\mathbb{P}_\alpha(\theta \in B \mid Y^{(n)}) = \frac{\int_B [p(Y^{(n)} \mid \theta)]^\alpha p_\theta(\theta) d\theta}{\int_{\Theta} [p(Y^{(n)} \mid \theta)]^\alpha p_\theta(\theta) d\theta}.$$

This leads to the following \(\alpha\)-VB objective

$$\tilde{q}(\theta) = \arg\min_{\varphi \in \mathcal{Q}} D(q \mid p_\theta(\cdot \mid Y^{(n)})) = \arg\min_q \alpha \Psi(q), \quad (14)$$

where

$$\Psi(q) = \int_{\Theta} q(\theta) \log \left[ \frac{p(Y^{(n)} \mid \theta^*)}{p(Y^{(n)} \mid \theta)} \right] d\theta - \alpha^{-1} D[q || p_\theta].$$

The variational expected log-likelihood ratio will be hence referred to as the model-fit term and the remaining KL term will be hence referred to as the regularization term.

The importance of the \(\alpha\)-VB framework comes from its ability to upper bound the variational Bayesian risk, the integral of $r(\theta, \theta^*) = n^{-1} D_{p_\theta} \left[ p_\theta^{(n)}(\cdot) \mid | p_\theta^{(n)}(\cdot) \right]$ with respect to $\tilde{q}(\theta)$, by the variational objective $\Psi(q)$. Minimizing the variational objective in turn minimizes the variational risk.

Before proceeding we motivate the form of our optimal risk bound. Consider preforming VI over the unrestricted class of densities over $\Theta$. Minimizing the \(\alpha\)-VB risk bound is achieved by balancing the two terms in terms in $\Psi(q)$. By choosing

$$q(\theta) = \frac{p_\theta(\theta) I_{B_n(\theta^*, \varepsilon)}(\theta)}{\mathbb{P}_\theta \left[ B_n(\theta^*, \varepsilon) \right]},$$

where $B_n(\theta^*, \varepsilon)$ is defined in (12), the model-fit term can be shown to be of order $O_p(n\varepsilon^2)$ and the regularization term can be shown to be $\alpha^{-1} \log \mathbb{P}_\theta \left[ B_n(\theta^*, \varepsilon) \right]^{-1}$, a multiple of the local Bayesian complexity. This is the optimal risk bound for variational inference considering the class of all distributions as the variational family (Yang et al., 2020). We summarize this in the theorem below.

**Theorem 4.2.** Assume $\tilde{q}_{\mu, \sigma}$ satisfies (14) and $\tilde{q}_{\mu, \sigma} \ll p_\theta$. It holds with $\mathbb{P}_\theta^{(n)}$ probability at least $1 - 2/((D - 1)^2 n(1 + n^{-2}) \varepsilon^2)$, that,

$$\int \frac{1}{n} D_{p_\theta} \left[ \theta^*, \theta \right] \tilde{q}_{\mu, \sigma}(\theta) d\theta \leq \frac{D_\alpha}{1 - \alpha} \varepsilon^2 + \frac{1}{n(1 - \alpha)} \log \left\{ \mathbb{P}_\theta \left[ B_n(\theta^*, \varepsilon) \right]^{-1} \right\} + O(n^{-1}).$$

We provide a sketch of the proof below. The full proof can be found in section S2.6 of the supplementary file.

Following our above motivation, we aim to show that a term in $\text{Assumption A1}$.

**Remark 4.1.** Let $\{p_\theta, \theta \in \Theta\}$ be a parametric family of densities. Assume for $\theta_1, \theta_2$, there exists $\alpha > 0$ such that $D(\theta^* || \theta) \lesssim \| \theta^* - \theta \|^{2\alpha}$, $\mu_2(\theta^* || \theta) \lesssim \| \theta^* - \theta \|^{2\alpha}$, and $\| |\theta_1 - \theta_2| \|^{(2\alpha)} \approx h(\theta_1, \theta_2) \approx \| \theta_1 - \theta_2 \|. \quad (16)$

Then if the model measure possesses a density that is uniformly bounded away from zero and infinity on $\Theta$, then Assumption A1 is satisfied. Assumptions of this form are common in the literature; refer to pg 517 (Ghosal et al., 2000).

**Corollary 4.1.** Suppose the prior density $p_\theta$ satisfies Assumption A1 and $\tilde{q}$ satisfies (14). It holds with probability tending to one as $n \to \infty$ that,

$$\left\{ \int h^2[p(\cdot \mid \theta) || p(\cdot \mid \theta^*)] \tilde{q}_{\mu, \sigma}(\theta) d\theta \right\}^{1/2} \leq O(n^{-1}),$$

Assumption A1 Prior density $p_\theta$ satisfies

$$\log \mathbb{P}_\theta \left[ B_n(\theta^*, \varepsilon) \right]^{-1} \leq -n\varepsilon^2.$$
demonstrating that the risk bound is parametric even when a flexible class of variational approximation is used.

5 Conclusion

To summarize, we have provided theoretical properties of transformation-based models in non-parametric and variational inferences in the context of NL-LVM. Further work is needed to generalize some of our results to higher dimensional models as several of the technical lemmas in the appendix hold only for dimension $d = 1$. A natural follow-up to this work would be to study the asymptotic distribution of the parameters of interest or a finite dimensional functional of densities arising from the estimates. These results would be in line with Bernstein-von Mises type theorems for the GP-LVM and GP-IVI.

Acknowledgements

Pati and Bhattacharya acknowledge support from NSF DMS (1854731, 1916371). In addition, Bhattacharya acknowledges the NSF CAREER 1653404 award for supporting this project.

References


Statistical Guarantees for Transformation Based Models


