Statistical Guarantees for Transformation Based Models with Applications to Implicit Variational Inference

Sean Plummer[1*] **Shuang Zhou**[2*] **Anirban Bhattacharya**[1] **David Dunson**[3] **Debdeep Pati**[1]
[1]Texas A&M University [2]Arizona State University [3]Duke University

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.

[*]Authors contributed equally to this work.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

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

$$\begin{aligned} y_i &= \mu(x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \\ x_i &\overset{iid}{\sim} g, \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$, where $y_i \in \mathbb{R}$ is a real-valued observed variable, μ 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 σ^2. For simplicity in exposition, we consider a very simple case to start but one can certain 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 μ, 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 μ, 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; Kruijer 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 $q_\phi(\theta) = \int q\{\theta \mid g_\phi(u)\}q(u)du$ where the density $q\{z \mid g_\phi(u)\}$ corresponds to a transformation-based model with transformation g_ϕ – typically taken to be a neural network with parameters ϕ. 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; Kruijer 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-parameteric inference using NL-LVM. In §4 we introduce GP-IVI. We then show that 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 $\alpha \in (0, 1)$.

1.1 Notation

We denote the Lesbesgue measure on $\mathbb{R}^p$ by λ. The supremum norm and L_1-norm are denoted by $\|\cdot\|_\infty$ and $\|\cdot\|_1$, respectively. For two density functions $p, q \in \mathcal{F}$, let h denote the Hellinger distance defined as $h^2(p, q) = \int (p^{1/2} - q^{1/2})^2 d\lambda$. Denote the Kullbeck-Leibler divergence between two probability densities p and q with respect to the Lebesgue measure by $D(p||q) = \int p \log(p/q) d\lambda$. We define the additional discrepancy measure $V(p||q) = \int p \log^2(p/q) d\lambda$, which will be referred to as the V-divergence. For a set A we use I_A to denote its indicator function. We denote the density of the normal distribution $N(t; 0, \sigma^2 I_d)$ by $\phi_\sigma(t)$. We denote the convolution of f and g by $f * g(y) = \int f(y - x)g(x)dx$. Absolute continuity of q with respect to p will be denoted $q \ll p$. We denote the set of all probability densities $f \ll \lambda$ by $\mathcal{F}$. The support of a density f is denoted by $\text{supp}(f)$. For a set $\mathcal{X}$, let $C(\mathcal{X})$ and $C^\beta(\mathcal{X})$, $\beta > 0$ denote the spaces of continuous functions and β-Hölder space, respectively. We write "$\precsim$" for inequality up to a constant multiple. For any $a > 0$ denote $\lfloor a \rfloor$ the largest integer that is no greater than a.

2 A specific transformation-based model

In this section, we focus on an NL-LVM model (Kundu & Dunson, 2014) in which the response variables are modeled as unknown functions (referred to as the transfer function) of uniformly distributed latent variables with an additive Gaussian error. We start from the model formulation and then present a general approximation result of 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 λ. We consider a non-linear latent variable model

$$\begin{aligned} Y_i &= \mu(\eta_i) + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2),\ i = 1, \ldots, n \\ \mu &\sim \Pi_\mu, \quad \sigma \sim \Pi_\sigma, \quad \eta_i \sim \text{U}(0, 1), \end{aligned} \tag{2}$$

where η_i's are latent variables, $\mu \in C[0, 1]$ is a *transfer function* relating the latent variables to the observed variables and ϵ_i is an idiosyncratic error. Marginalizing out the latent variable, we obtain the density of y conditional on the transfer function μ and scale σ

$$f(y; \mu, \sigma) \stackrel{\text{def}}{=} f_{\mu,\sigma}(y) = \int_0^1 \phi_\sigma(y - \mu(x))dx. \tag{3}$$

Remark 2.1. *While μ and η 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 μ without the need to identify the corresponding latent variable η. The presence of the latent variable η 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 $F_0\{F_0^{-1}(t)\} = t$ 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 σ. 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 $\tilde{\lambda}$ 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 ν_μ denote the induced measure on $(\mathbb{R}, \mathcal{B})$, then, for any Borel measurable set B, $\nu_\mu(B) = \tilde{\lambda}(\mu^{-1}(B))$. By the change of variable

theorem for induced measures,

$$\int_0^1 \phi_\sigma(y-\mu(x))dx = \int \phi_\sigma(y-t)d\nu_\mu(t), \tag{4}$$

so that $f_{\mu,\sigma}$ in (3) can be expressed as a kernel mixture form with mixing distribution ν_μ. It turns out that this mechanism of creating random distributions is very general. Depending on the choice of μ, one can create a large variety of mixing distributions based on this specification. For example, if μ is a strictly monotone function, then ν_μ is absolutely continuous with respect to the Lebesgue measure, while choosing μ to be a step function, one obtains a discrete mixing distribution.

2.2 Assumptions on true data density f_0

It is widely recognized that one needs certain smoothness assumptions and tail conditions on the true density f_0 to derive posterior convergence rates. We make the following assumptions:

Assumption F1 We assume $\log f_0 \in C^\beta[0,1]$. Let $l_j(x) = d^j/dx^j\{\log f_0(x)\}$ be the jth derivative for $j = 1,\ldots,r$ with $r = \lfloor\beta\rfloor$. For any $\beta > 0$, we assume that there exists a constant $L > 0$ such that

$$|l_r(x) - l_r(y)| \le L|x-y|^{\beta-r}, \text{ for all } x \ne y. \tag{5}$$

The smoothness assumption in the log scale will be used to obtain an optimal approximation error of the GP-transformation-based model to the true f_0, providing a key piece in managing the KL-divergence between the true and the model for posterior inference. Similar assumption on the local smoothness appeared in Kruijer et al. (2010), while in our case a global smoothness assumption is sufficient since f_0 is assumed to be compactly supported.

Assumption F2 We assume f_0 is compactly supported on $[0,1]$, 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 $f_0 * \phi_\sigma \ge C f_0$ 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 ν_μ. 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 $\phi_\sigma * f_0$ 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 $\beta \le 2$, the standard Taylor series expansion guarantees that $||\phi_\sigma * f_0 - f_0||_\infty = O(\sigma^\beta)$. 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 $\{f_j\}_{j\ge 0}$ by

$$f_{j+1} = f_0 - \triangle_\sigma f_j, \quad \triangle_\sigma f_j = \phi_\sigma * f_j - f_j, \quad j \ge 0. \tag{6}$$

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 $\phi_\sigma * f_\beta$ 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_β defined as from the iterative procedure* (6) *we have*

$$\|\phi_\sigma * f_\beta - f_0\|_\infty = O(\sigma^\beta),$$

and

$$\phi_\sigma * f_\beta(x) = f_0(x)(1 + D(x)O(\sigma^\beta)), \tag{7}$$

where

$$D(x) = \sum_{i=1}^{r} c_i |l_j(x)|^{\frac{\beta}{i}} + c_{r+1},$$

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 β-Hölder space $\mathcal{C}^\beta[0,1]^d$ 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 μ 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 Π on $\mathcal{F}$ via the mapping $f_{\mu,\sigma}$ by placing independent priors Π_μ and Π_σ 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 μ and an inverse-gamma prior on σ^2. Given the flexibility of $f_{\mu,\sigma}$ upon the choices of μ, placing a prior on μ 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 Π_μ on μ 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 Π_σ on σ 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 Π 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 Π_μ has full sup-norm support on $C[0,1]$ and Π_σ has full support on $[0,\infty)$, then the L_1 support of the induced prior Π on $\mathcal{F}$ 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 μ (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 μ_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 μ_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 Π_n associated with a prior Π on $\mathcal{F}$ is said to contract at a rate ϵ_n, if for a distance metric d_n on $\mathcal{F}$,

$$\mathbb{E}_{f_0} \Pi_n \{ d_n(f, f_0) > M\epsilon_n \mid Y^{(n)} \} \to 0 \tag{8}$$

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_β defined as in (6), then allow the GP prior on the transfer function to appropriately concentrate around μ_β, the inverse c.d.f. of the defined f_β. We first state our choices for the prior distributions Π_μ and Π_σ.

Assumption P1 We assume μ follows a centered and rescaled Gaussian process denoted by $\mathrm{GP}(0, c^A)$, where A denotes the rescaled parameter, and assume A has density g satisfying for $a > 0$,

$$C_1 a^p \exp\left(-D_1 a \log^q a\right) \le g(a) \le C_2 a^p \exp\left(-D_2 a \log^q a\right).$$

Assumption P2 We assume $\sigma \sim \mathrm{IG}(a_\sigma, b_\sigma)$.

Note that contrary to the usual conjugate choice of an inverse-gamma prior for σ^2, we have assumed an inverse-gamma prior for σ. This enables one to have slightly more prior mass near zero compared to an inverse-gamma prior for σ^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 **F1** and **F2** and the priors Π_μ and Π_σ are as in Assumptions **P1** and **P2** respectively, the best obtainable rate of posterior convergence relative to Hellinger metric h is*

$$\epsilon_n = n^{-\frac{\beta}{2\beta+1}}(\log n)^t, \tag{9}$$

where $t = \beta(2 \vee q)/(2\beta+1) + 1$.

We provide a sketch of the proof below, the full proof is deferred to the supplementary file section S2.4. It suffices to check sufficient conditions (prior thickness, sieve construction, entropy condition) for posterior contraction result in Ghosal et al. (2000) (See Theorem S1 in the supplementary file for details.) We first verify the prior thickness condition. From Lemma 8 of Ghosal & van der Vaart (2007), one has

$$\int f_0 \log\left(\frac{f_0}{f_{\mu,\sigma}}\right)^i \leq h^2(f_0, f_{\mu,\sigma})\left(1 + \log\left\|\frac{f_0}{f_{\mu,\sigma}}\right\|_\infty\right)^i,$$

for $i = 1, 2$. By Lemma S3.4, we have $\log\|f_0/f_{\mu,\sigma}\|_\infty \leq \|\mu - \mu_\beta\|_\infty/\sigma^2$, and by Lemma S3.1 and Lemma S3.8, we bound $h^2(f_0, f_{\mu,\sigma}) \precsim \|\mu - \mu_\beta\|_\infty/\sigma^2 + O(\sigma^{2\beta})$. Then we have

$$\{\sigma \in [\sigma_n, 2\sigma_n], \|\mu - \mu_\beta\|_\infty \precsim \sigma_n^{\beta+1}\} \subset \{D(f_0||f_{\mu,\sigma}) \precsim \sigma_n^{2\beta}, V(f_0||f_{\mu,\sigma}) \precsim \sigma_n^{2\beta}\}.$$

Under assumptions **P1** and **P2** the prior thickness is guarantee by upper bounding $\Pi\{\sigma \in [\sigma_n, 2\sigma_n], \|\mu - \mu_\beta\|_\infty \precsim \sigma_n^{\beta+1}\}$. We construct the sieve

$$\mathcal{F}_n = \{f_{\mu,\sigma} : \mu \in B_n, l_n < \sigma < h_n\}.$$

where B_n denotes the sieve for a GP prior on μ as defined in van der Vaart & van Zanten (2009). Further we calculate the entropy of $\mathcal{F}_n$; the logarithm of number of small balls in L_1 norm with radius at least ϵ_n covering $\mathcal{F}_n$; by observing that for $\sigma_2 > \sigma_1 > \sigma_2/2$,

$$\|f_{\mu_1,\sigma_1} - f_{\mu_2,\sigma_2}\|_1 \leq \left(\frac{2}{\pi}\right)^{1/2}\frac{\|\mu_1 - \mu_2\|_\infty}{\sigma_1} + \frac{3(\sigma_2 - \sigma_1)}{\sigma_1}.$$

The entropy condition can be verified by applying Lemma S3.9. Finally, the sieve compliment condition is easily verified by combining the results on GP priors in van der Vaart & van Zanten (2009) and tail properties of inverse-gamma distribution of σ.

4 Gaussian Process Implicit Variational Inference

Motivated 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 θ that admits a density p_θ over Θ. With a slight abuse of notation, we will use $p(Y^{(n)} \mid \theta)$ to denote $\mathbb{P}_\theta^{(n)}$ and its density function. We adopt a frequentist framework and assume a true data generating distribution $\mathbb{P}_{\theta^*}^{(n)}$ and a true parameter θ^*. 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

$$\ell_i(\theta, \theta^*) = \log[p(Y_i \mid \theta)/p(Y_i \mid \theta^*)]. \tag{10}$$

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

$$D(\theta^*||\theta) = -\mathbb{E}_{\theta^*}^{(n)}[\ell_1(\theta, \theta^*)],\ V_2(\theta^*||\theta) = \mathbb{E}_{\theta^*}^{(n)}[\ell_1(\theta, \theta^*)^2]. \tag{11}$$

Lastly denote the appropriate neighborhood around the true parameter θ^*,

$$B_n(\theta^*, \varepsilon) = \{\theta \mid D[p(Y^{(n)} \mid \theta^*)\|p(Y^{(n)} \mid \theta)] \leq n\varepsilon^2, V[p(Y^{(n)} \mid \theta^*)\|p(Y^{(n)} \mid \theta)] \leq n\varepsilon^2\}. \tag{12}$$

4.2 Gaussian Process Implicit Variational Inference

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

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

Marginalizing over the latent η gives us the implict variational distribution,

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

Together this defines the Gaussian process implict 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 θ^* using the model

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

in which μ_0, σ_0, σ are all known. Let $\overline{Y}_n, \mu_n, \sigma_n^2$ denote the sample mean, the posterior mean, and variance, respectively. Straight forward calculations show

$$D\left[N(\theta^*, n^{-1}\sigma^2)||N(\mu_n, \sigma_n^2)\right] \to \chi_1^2, \text{ weakly.}$$

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

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

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

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

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

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

Assumption B2 The variance bound σ_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 $\theta, \theta' \in \mathbb{R}^d$,

$$\begin{aligned} s_{max}\left(D^{(2)}(\theta^*||\theta) - D^{(2)}(\theta^*||\theta')\right) &\leq C\|\theta - \theta'\|_2^{\alpha_1}, \\ s_{max}\left(\mu_2^{(2)}(\theta^*||\theta) - \mu_2^{(2)}(\theta^*||\theta')\right) &\leq C\|\theta - \theta'\|_2^{\alpha_2}, \\ s_{max}\left(U^{(2)}(\theta) - U^{(2)}(\theta')\right) &\leq C\|\theta - \theta'\|_2^{\alpha_3}, \end{aligned}$$

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 Γ_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 $m_n^*(\mathcal{Q}_n) = \min_{q\in\mathcal{Q}_n}\{D[q||p(\cdot \mid Y^n)]\}$ is bounded in probability with respect to the data generating distribution $\mathbb{P}_{\theta^*}^{(n)}$. Formally, given any $\varepsilon > 0$, there exists $M_\varepsilon, N_\varepsilon > 0$ such that for $n \geq N_\varepsilon$, we have $\mathbb{P}_{\theta^*}^{(n)}(m_n^*(\mathcal{Q}_n) > M_\varepsilon) \leq \varepsilon$.*

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, $m_n^*(\mathcal{Q}_n) \leq D[q_n||p(\cdot \mid Y^{(n)})]$. We show $D[q_n||p(\cdot \mid Y^{(n)})]$ is $O_p(1)$ by showing that it is a sum of $O_p(1)$ terms. Letting $\mathbb{E}_n$ denote the expectation with respect to q_n, $D[q_n||p(\cdot \mid Y^{(n)})]$ can be broken into four parts $\mathbb{E}_n[\log q_n]$, $\log m(Y^{(n)})$, $\mathbb{E}_n[U(\theta)]$, and $\mathbb{E}_n\left[\sum_{i=1}^n \ell_i(\theta, \theta^*)\right]$. The first term $\mathbb{E}_n[\log q_n]$ is a constant, hence $O_p(1)$. Noting $\mathbb{E}_{\theta^*}^{(n)}[m(Y^{(n)})] = 1$, 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 θ^* and applying assumption **B4** and **B5** gives us the bounds

$$\begin{aligned} C_\ell(\sigma^2 + \sigma_n^2) \leq \mathbb{E}_n[D(\theta^*||\theta)] &\leq C_u(\sigma^2 + \sigma_n^2), \\ \mathbb{E}_n[\mu_2(\theta^*||\theta)] &\leq C_2(\sigma^2 + \sigma_n^2), \\ \mathbb{E}_n[U(\theta)] &\leq C_1(\sigma^2 + \sigma_n^2). \end{aligned} \tag{13}$$

Markov's inequality shows that $U(\theta)$ is $O_p(1)$. It remains to show $\mathbb{E}_n\left[\sum_{i=1}^n \ell_i(\theta, \theta^*)\right]$ is $O_p(1)$. Given $\varepsilon > 0$, choose $\delta = \left[C_2 c_0/(\varepsilon C_\ell)^2\right]^{1/2}$. Applying Chebychev's and Jensen's inequalities together with (13) we have,

$$\begin{aligned} \mathbb{P}_{\theta^*}^{(n)}\left\{\mathbb{E}_n\left[\sum_{i=1}^n \ell_i(\theta, \theta^*)\right] \leq -C_u(1+\delta)n(\sigma^2 + \sigma_n^2)\right\} \\ \leq \frac{\mathbb{E}_n[\mu_2(\theta^*||\theta)]}{\delta^2 n\left(E_n[D(\theta^*||\theta)]\right)^2} \leq \frac{C_2}{C_\ell \delta^2 n \sigma_n^2}. \end{aligned}$$

Finally by assumption **B2** we have $c_0 n \leq \sigma_n^{-2}$. Thus

$$\begin{aligned} \mathbb{P}_{\theta^*}^{(n)}\left\{\mathbb{E}_n\left[\sum_{i=1}^n \ell_i(\theta, \theta^*)\right] \leq -2C_u\left(1 + \left[C_2 c_0/(\varepsilon C_\ell)^2\right]^{1/2}\right)\right\} \\ \leq \varepsilon, \end{aligned}$$

which shows $\mathbb{E}_n\left[\sum_{i=1}^n \ell_i(\theta,\theta^*)\right]$ is $O_p(1)$. Combining the four bounds completes the proof.

4.4 α-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. α-variational Bayes (α-VB) (Yang et al., 2020) is a variational inference framework that aims to minimize the KL divergence between the variational density and the α-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 α-VB objective

$$\widehat{q}(\theta) = \underset{q\in\mathcal{Q}}{\operatorname{argmin}}\ D(q||p_\alpha(\cdot \mid Y^{(n)})) = \underset{q}{\operatorname{argmin}}\ \alpha\Psi(q), \quad (14)$$

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 α-VB framework comes from its ability to upper bound the variational Bayesian risk, the integral of $r(\theta,\theta^*) = n^{-1} D_\alpha[p_\theta^{(n)}||p_{\theta^*}^{(n)}]$ with respect to $\widehat{q}(\theta)$, 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 Θ. Minimizing the α-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 $\alpha^{-1}\log[\mathbb{P}_\theta\{B_n(\theta^*,\varepsilon)\}^{-1}]$, 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 $\widehat{q}_{\mu,\sigma}$ satisfies* (14) *and $\widehat{q}_{\mu,\sigma} \ll p_\theta$. It holds with $\mathbb{P}_{\theta^*}^{(n)}$-probability at least $1 - 2/[(D-1)^2 n(1+n^{-2})\varepsilon^2]$ that,*

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

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 there is a member of the GP-IVI family $\mathcal{Q}_{GP}$ such that the model-fit term is of order $O_p(n\varepsilon^2)$ and the regularization term is proportional to the local Bayesian complexity. We leverage the approximation properties from §3 to construct an approximation that achieve this balance. We construct this variational distribution as follows.

Let the prior distribution of θ is given by the density $p_\theta(\theta) = f_0(\theta) \in C^\beta[0,1]$, $\beta \in (2j, 2j+2]$. Let $f_\beta = f_j$ be the density constructed as in (6) satisfying $\|\phi_\sigma * f_\beta - f_0\|_\infty = O(\sigma^\beta)$. Define the density function

$$\widetilde{f}_\beta(t) = \frac{f_\beta(t) I_{B_n(\theta^*,\varepsilon)}}{\int_{B_n(\theta^*,\varepsilon)} f_\beta(t)dt} \quad (15)$$

and its corresponding variational density

$$q_{\widetilde{f}_\beta,\sigma}(\theta) = \int_{-\infty}^{\infty} \phi_\sigma(\theta - t)\widetilde{f}_\beta(t)dt. \quad (16)$$

The model-fit term is bounded in high probability using a straight forward application of Chebychev's inequality. Using (7), we bound the regularization term proportional to the local Bayesian complexity. Combining these and using Theorem 3.2 of Yang et al. (2020) finishes the proof.

Assumption A1 Prior density p_θ satisfies $\log[\mathbb{P}_\theta\{B_n(\theta^*,\varepsilon)\}^{-1}] \leq -n\varepsilon^2$.

Remark 4.1. *Let $\{p_\theta, \theta \in \Theta\}$ be a parametric family of densities. Assume for $\theta, \theta_1, \theta_2$, there exists $\alpha > 0$ such that $D(\theta^*\|\theta) \precsim \|\theta^* - \theta\|^{2\alpha}$, $\mu_2(\theta^*\|\theta) \precsim \|\theta^* - \theta\|^{2\alpha}$, and $\|\theta_1 - \theta_2\|^\alpha \precsim h(\theta_1,\theta_2) \precsim \|\theta_1 - \theta_2\|^\alpha$. Then if the prior measure possesses a density that is uniformly bounded away from zero and infinity on Θ, 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_θ satisfies Assumption* **A1** *and $\widehat{q}$ satisfies* (14)*. It holds with probability tending to one as $n \to \infty$ that,*

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

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 inline 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

Bhattacharya, A., Pati, D., & Yang, Y. (2019, 02). Bayesian fractional posteriors. *The Annals of Statistics*, *47*(1), 39–66.

Dasgupta, S., Pati, D., & Srivastava, A. (2017, 01). A geometric framework for density modeling. *Statistica Sinica*.

Devroye, L. (1992). A note on the usefulness of superkernels in density estimation. *The Annals of Statistics*, 2037–2056.

Escobar, M., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, *2*(4), 615–629.

Ferris, B., Fox, D., & Lawrence, N. (2007). Wifi-slam using Gaussian process latent variable models. In *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 2480–2485).

Figurnov, M., Mohamed, S., & Mnih, A. (2018). Implicit reparameterization gradients. In *Advances in neural information processing systems* (pp. 441–452).

Ghosal, S., Ghosh, J., & Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, *27*(1), 143–158.

Ghosal, S., Ghosh, J., & van der Vaart, A. (2000, 04). Convergence rates of posterior distributions. *Ann. Statist.*, *28*(2), 500–531.

Ghosal, S., & van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, *35*(2), 697–723.

Huszár, F. (2017). Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*.

Jankowiak, M., & Obermeyer, F. (2018). Pathwise derivatives beyond the reparameterization trick. *arXiv preprint arXiv:1806.01851*.

Kingma, D., Salimans, T., & Welling, M. (2015). Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems* (pp. 2575–2583).

Kingma, D., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kruijer, W., Rousseau, J., & van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, *4*, 1225–1257.

Kundu, S., & Dunson, D. B. (2014). Latent factor models for density estimation. *Biometrika*, *101*(3), 641–654.

Lawrence, N. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems 16: proceedings of the 2003 conference* (Vol. 16, p. 329).

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, *6*, 1783–1816.

Lawrence, N., & Moore, A. (2007). Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on machine learning* (pp. 481–488).

Lenk, P. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association*, *83*(402), 509–516.

Lenk, P. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, *78*(3), 531–543.

MacEachern, S. (1999). Dependent nonparametric processes. In *Proceedings of the section on Bayesian statistical science* (pp. 50–55).

Molchanov, D., Kharitonov, V., Sobolev, A., & Vetrov, D. (2019). Doubly semi-implicit variational inference. In *The 22nd international conference on artificial intelligence and statistics* (pp. 2593–2602).

Müller, P., Erkanli, A., & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, *83*(1), 67–79.

Shi, J., Sun, S., & Zhu, J. (2017). Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*.

Titsias, M., & Lawrence, N. (2010). Bayesian Gaussian process latent variable model. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 844–851).

Titsias, M., & Ruiz, F. (2019). Unbiased implicit variational inference. In *The 22nd international conference on artificial intelligence and statistics* (pp. 167–176).

Tokdar, S. (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics*, *16*(3), 633-655.

Tokdar, S., Zhu, Y., & Ghosh, J. (2010, 06). Bayesian density regression with logistic Gaussian process and subspace projection. , *5*(2), 319–344.

van der Vaart, A., & van Zanten, J. (2007). Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics*, *1*, 433–448.

van der Vaart, A., & van Zanten, J. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, *36*(3), 1435–1463.

van der Vaart, A., & van Zanten, J. (2009). Adaptive Bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, *37*(5B), 2655–2675.

Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. CRC Press.

Yang, Y., Pati, D., & Bhattacharya, A. (2020). α-Variational inference with statistical guarantees. *The Annals of Statistics*, *48*(2), 886–905.

Yin, M., & Zhou, M. (2018). Semi-implicit variational inference. In *International conference on machine learning* (pp. 5660–5669).