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).

Supplementary Document to "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

S1 A brief introduction to nonparametric Bayes

S1.1 Posterior contraction in nonparametic setting

We first give a brief review of the contraction rate of a posterior distribution under a general nonparametric regression setting. Given independently and identically distributed samples $Y^{(n)}$ generated from the true density f_0, a regular nonparametric model considers $Y_i \mid f \overset{i.i.d.}{\sim} f(\cdot)$ for some unknown density $f \in \mathcal{F}$, where $\mathcal{F}$ denotes a suitable class of the density functions that are absolutely continuous with respect to the Lebesgue measure. Assigning a nonparametric prior $\Pi(\cdot)$ over the set $\mathcal{F}$ and multiplying it with the likelihood denoted by $P(Y^{(n)} \mid f)$ produces the posterior distribution $\Pi_n(\cdot \mid Y^{(n)})$ defined as

$$\Pi_n(f \in B \mid Y^{(n)}) = \frac{\int_B P(Y^{(n)} \mid f) d\Pi(f)}{\int P(Y^{(n)} \mid f) d\Pi(f)},$$

for any set $B \subset \mathcal{F}$. As the posterior distribution is a random measure conditioning on the given data, we are interested in studying frequentist properties of such posterior distribution such as the consistency and convergence rate to the true data generating function f_0. In particular, the convergence rate characterizes how fast a posterior distribution concentrates on the true density f_0 as n increases, measured by the decreasing rate of the radius of a neighborhood centered at the true f_0 that received posterior probability converging to 1. We define the posterior distribution contracts at a rate ϵ_n to the true function f_0 with respect to certain metric $d(\cdot, \cdot)$ almost surely under the true probability measure denoted by E_{f_0}, if

$$E_{f_0}\{\Pi_n(d(f, f_0) > M\epsilon_n \mid Y^{(n)})\} \to 0, \quad \text{as} \quad n \to \infty,$$

for some sufficiently large integer $M > 0$. Ghosal et al. (2000) derived a general approach to obtain the optimal rate (up to a logarithmic factor) by verifying sufficient conditions regarding the prior measure and the considered density space $\mathcal{F}$. We now restate Theorem 2.1 of Ghosal et al. (2000).

Theorem S1. If there exist sequences $\bar{\epsilon}_n, \widetilde{\epsilon}_n \to 0$ with $n \min\{\bar{\epsilon}_n^2, \widetilde{\epsilon}_n^2\} \to \infty$ such that there exist constants $C_1, C_2, C_3, C_4 > 0$ and a sequence of sieve $\mathcal{F}_n \subset \mathcal{F}$ so that,

$$\text{(Entropy condition)} \qquad \log N(\bar{\epsilon}_n, \mathcal{F}_n, d) \le C_1 n \bar{\epsilon}_n^2, \tag{S1.1}$$

$$\text{(Sieve condition)} \qquad \Pi(\mathcal{F}_n^c) \le C_3 \exp\{-n\widetilde{\epsilon}_n^2 (C_2 + 4)\}, \tag{S1.2}$$

$$\text{(Prior thickness condition)} \quad \Pi\left(f : \int f_0 \log \frac{f_0}{f} \le \widetilde{\epsilon}_n^2, \int f_0 \log \left(\frac{f_0}{f}\right)^2 \le \widetilde{\epsilon}_n^2\right) \ge C_4 \exp\{-C_2 n \widetilde{\epsilon}_n^2\}. \tag{S1.3}$$

then we have

$$E_{f_0}\{\Pi_n(d(f, f_0) > M\epsilon_n \mid Y^{(n)})\} \to 0, \quad \text{a.s.} \quad \text{as} \ n \to \infty,$$

for some sufficiently large constant $M > 0$.

S1.2 Gaussian process and its reproducing kernel Hilbert space

We first review the definition of Gaussian process. A Gaussian process defined on a probability space $(\Omega, \mathcal{U}, P)$ is a collection of random variables $\{X(t), t \in T\}$ indexed by some arbitrary set T such that each finite dimensional subset of random variables has a joint multivariate normal distribution with mean function $\mu(t) = E(X(t))$ and convaraince kernel function $K(s,t) = \mathrm{Cov}(X(s), X(t))$. For some univaraite function $f : \mathbb{R} \to \mathbb{R}$, we endow it with a Gaussian process prior denoted by $f \sim GP(\mu(\cdot), K(\cdot,\cdot))$ with $\mu(x) = E(f(x))$ and $K(x,x') = \mathrm{Cov}(f(x), f(x'))$ for any $x, x' \in \mathbb{R}$. The mean function reflects the expected center of realizations and the covariance kernel function controls the smoothness of the realizations and correlations of the realization across covariates. Refer to Rasmussen (2003) for a detailed introduction to Gaussian processes.

We now briefly recall the definition of the reproducing kernel Hilbert space of a Gaussian process prior; a detailed review can be found in van der Vaart & van Zanten (2008). A Borel measurable random element W with values in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ (e.g., $C[0,1]$) is called Gaussian if the random variable b^*W is normally distributed for any element $b^* \in \mathbb{B}^*$, the dual space of $\mathbb{B}$. The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to a zero-mean Gaussian process W is defined as the completion of the linear space of functions $t \mapsto EW(t)H$ relative to the inner product

$$\langle \mathrm{E}W(\cdot)H_1; \mathrm{E}W(\cdot)H_2 \rangle_{\mathbb{H}} = \mathrm{E}H_1H_2,$$

where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i W(s_i)$ with $a_i \in \mathbb{R}$ and s_i in the index set of W.

Let $W = (W_t : t \in \mathbb{R})$ be a Gaussian process with squared exponential covariance kernel. The spectral measure m_w of W is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}$ with the Radon-Nikodym derivative given by

$$\frac{dm_w}{d\lambda}(x) = \frac{1}{2\pi^{1/2}} e^{-x^2/4}.$$

Define a scaled Gaussian process $W^a = (W_{at} : t \in [0,1])$, viewed as a map in $C[0,1]$. Let $\mathbb{H}^a$ denote the RKHS of W^a, with the corresponding norm $\|\cdot\|_{\mathbb{H}^a}$. The unit ball in the RKHS is denoted $\mathbb{H}^a_1$.

S2 Proofs of results in the main document

S2.1 Conventions

Equations in the main document are cited as (1), (2) etc., retaining their numbers, while new equations defined in this document are numbered (S1), (S2) etc. In this section we collect the proof of Proposition 2.1, Theorems 3.1, 3.2, 4.1 and 4.2.

S2.2 Proof of Proposition 2.1

In this section we prove the results in Proposition 2.1.

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

$$\phi_\sigma * f_\beta(x) = f_0(x)(1 + D(x)O(\sigma^\beta)), \tag{S2.1}$$

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 $x \in [0,1]$.

Proof. We now show equation (S2.1). Following the proof of Lemma 1 in Kruijer et al. (2010), for any $x, y \in [0,1]$,

$$\log f_0(y) \leq \log f_0(x) + \sum_{i=1}^{r} \frac{l_j(x)}{j!}(y-x)^j + L|y-x|^\beta,$$
$$\log f_0(y) \geq \log f_0(x) + \sum_{i=1}^{r} \frac{l_j(x)}{j!}(y-x)^j - L|y-x|^\beta.$$

Define

$$B^u_{f_0,r}(x,y) = \sum_{i=1}^{r} \frac{l_j(x)}{j!}(y-x)^j + L|y-x|^\beta,$$
$$B^l_{f_0,r}(x,y) = \sum_{i=1}^{r} \frac{l_j(x)}{j!}(y-x)^j - L|y-x|^\beta.$$

Then we have

$$e^{B^u_{f_0,r}} \leq 1 + B^u_{f_0,r} + \frac{1}{2!}(B^u_{f_0,r})^2 + \cdots + M|B^u_{f_0,r}|^{r+1},$$
$$e^{B^l_{f_0,r}} \geq 1 + B^l_{f_0,r} + \frac{1}{2!}(B^l_{f_0,r})^2 + \cdots - M|B^l_{f_0,r}|^{r+1}.$$

where

$$M = \frac{1}{(r+1)!} \exp\left\{ \sup_{x,y\in[0,1],x\neq y} \left(\left| \sum_{j=1}^{r} \frac{l_j(x)}{j!}(y-x)^j \right| + L|y-x|^\beta \right) \right\}.$$

Note that f_0 is bounded on $[0,1]$, we consider the convolution on the whole real line by extending f_0 analytically outside $[0,1]$. For $\beta \in (1,2], r = 1$ and $x \in (0,1)$,

$$\begin{aligned} \phi_\sigma * f_0(x) &\leq f_0(x) \int e^{B^u_{f_0,r}(x,y)} \phi_\sigma(y-x)dy \\ &\leq f_0(x) \int_{\mathbb{R}} \phi_\sigma(y-x)[1 + L|y-x|^\beta + M\{l_1^2(x)(y-x)^2 + 2Ll_1(x)(y-x)|y-x|^\beta + L^2|y-x|^{2\beta}\}]dy. \end{aligned} \tag{S2.2}$$

Since $l_j(x)$'s are all continuous on $[0,1]$, there exist finite constants M_j such that $|l_j| \leq M_j$ and $|y-x| \leq 1$. The integral in the last inequality in (S2.2) can be bounded by

$$\int_{\mathbb{R}} \phi_\sigma(y-x)[1 + L|y-x|^\beta + M\{M_1^{2-\beta}|l_1(x)(y-x)|^\beta + (L^2 + 2M_1)|y-x|^\beta\}]dy$$

Therefore,

$$\phi_\sigma * f_0(x) \leq f_0(x)\{1 + (r_1|l_1(x)|^\beta + r_2)\sigma^\beta\},$$

where $r_1 = MM_1^{2-\beta}\mu'_\beta$, $r_2 = L(1 + ML + 2MM_1)\mu'_\beta$, and $\mu'_\beta = \mathbb{E}\{|y-x|^\beta\}$.
In the other direction,

$$\phi_\sigma * f_0(x) \geq f_0(x) \int \phi_\sigma(y-x)[\{1 - L|y-x|^\beta - M\{l_1^2(x)(y-x)^2 - 2Ll_1(x)(y-x)|y-x|^\beta + L^2|y-x|^{2\beta}\}]dy.$$

Thus we achieve expression of $\phi_\sigma * f_\beta$ in Proposition 2.1.

For any $\beta > 2$ and the integer j such that $\beta \in (2j, 2j+2]$. We define $\phi^{(i)} * f$ as the i-folded convolution of ϕ with f for any integer $i \geq 1$. First we calculate $\phi_\sigma * f_0(x)$, $\phi_\sigma^{(2)} * f_0(x)$, ..., $\phi_\sigma^{(j)} * f_0(x)$, and by Lemma S3.5 we get $\phi_\sigma * f_j(x)$. The calculation of $\phi_\sigma^{(i)} * f_0(x)$ is the same as that of $\phi_\sigma * f_0(x)$ except taking the convolution with $\phi_{\sqrt{i}\sigma}$. The terms σ^2, σ^4, ..., σ^{2j} caused by the factors containing $|y-x|^k$ for $k < \beta$ in $\phi_\sigma^{(i)} * f_0$ can be canceled out by Lemma S3.5. For terms containing $|y-x|^k$ for $k \geq \beta$, we take out $|y-x|^\beta$ and bound the rest by a certain power of $|l_j(x)|$ or some constant. Following an induction in Kruijer et al. (2010), we can guarantee the approximation error of $\phi_\sigma * f_\beta$ is at the order of $O(\sigma^\beta)$. □

S2.3 Proof of Theorem 3.1

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.

Proof. Let f_0 be a density with quantile function μ_0 that satisfies the conditions of Theorem 3.1. Observe that $\|\mu_0\|_1 = \int_{t=0}^1 |\mu_0(t)|\, dt = \int_{-\infty}^{\infty} |z|\, f_0(z) dz < \infty$ since f_0 has a finite first moment, and thus $\mu_0 \in \mathrm{L}_1[0,1]$. Fix $\epsilon > 0$. We want to show that $\Pi\{B_\epsilon(f_0)\} > 0$, where $B_\epsilon(f_0) = \{f \ : \ \|f - f_0\|_1 < \epsilon\}$.

Note that $\mu_0 \notin C[0,1]$, so that $\mathbb{P}(\|\mu - \mu_0\|_\infty < \epsilon)$ can be zero for small enough ϵ. The main idea is to find a continuous function $\widetilde{\mu}_0$ close to μ_0 in L_1 norm and exploit the fact that the prior on μ places positive mass to arbitrary sup-norm neighborhoods of $\widetilde{\mu}_0$. The details are provided below.

Since $\|\phi_\sigma * f_0 - f_0\|_1 \to 0$ as $\sigma \to 0$, find σ_1 such that $\|\phi_\sigma * f_0 - f_0\|_1 < \epsilon/2$ for $\sigma < \sigma_1$. Pick any $\sigma_0 < \sigma_1$. Since $C[0,1]$ is dense in $\mathrm{L}_1[0,1]$, for any $\delta > 0$, we can find a continuous function $\widetilde{\mu}_0$ such that $\|\mu_0 - \widetilde{\mu}_0\|_1 < \delta$. Now, $\|f_{\mu,\sigma} - f_{\widetilde{\mu}_0,\sigma}\|_1 \leq C \|\mu - \widetilde{\mu}_0\|_1 / \sigma$ for a global constant C. Thus, for $\delta = \epsilon\, \sigma_0/4$,

$$\left\{f_{\mu,\sigma} \ : \ \sigma_0 < \sigma < \sigma_1, \|\mu - \widetilde{\mu}_0\|_\infty < \delta\right\} \subset \left\{f_{\mu,\sigma} \ : \ \|f_0 - f_{\mu,\sigma}\|_1 < \epsilon\right\},$$

since $\|f_0 - f_{\mu,\sigma}\|_1 < \|f_0 - f_{\mu_0,\sigma}\|_1 + \|f_{\mu_0,\sigma} - f_{\widetilde{\mu}_0,\sigma}\|_1 + \|f_{\widetilde{\mu}_0,\sigma} - f_{\mu,\sigma}\|_1$ and $f_{\mu_0,\sigma} = \phi_\sigma * f_0$. Thus, $\Pi\{B_\epsilon(f_0)\} > \Pi_\mu(\|\mu - \widetilde{\mu}_0\|_\infty < \delta)\, \Pi_\sigma(\sigma_0 < \sigma < \sigma_1) > 0$, since Π_μ has full sup-norm support and Π_σ has full support on $[0,\infty)$. □

S2.4 Proof of Theorem 3.2

In this section we will give a detailed proof for the adaptive posterior contraction rate result for the NL-LVM models.

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

$$\epsilon_n = n^{-\frac{\beta}{2\beta+1}} (\log n)^t, \tag{S2.3}$$

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

Proof. Following Ghosal et al. (2000), to obtain the posterior convergence rate we need to find sequences $\bar{\epsilon}_n, \widetilde{\epsilon}_n \to 0$ with $n \min\{\bar{\epsilon}_n^2, \widetilde{\epsilon}_n^2\} \to \infty$ such that there exist constants $C_1, C_2, C_3, C_4 > 0$ and sets $\mathcal{F}_n \subset \mathcal{F}$ so that,

$$\log N(\bar{\epsilon}_n, \mathcal{F}_n, d) \leq C_1 n \bar{\epsilon}_n^2, \tag{S2.4}$$

$$\Pi(\mathcal{F}_n^c) \leq C_3 \exp\{-n\widetilde{\epsilon}_n^2 (C_2 + 4)\}, \tag{S2.5}$$

$$\Pi\left(f_{\mu,\sigma} : \int f_0 \log \frac{f_0}{f_{\mu,\sigma}} \leq \widetilde{\epsilon}_n^2,\ \int f_0 \log \left(\frac{f_0}{f_{\mu,\sigma}}\right)^2 \leq \widetilde{\epsilon}_n^2\right) \geq C_4 \exp\{-C_2 n \widetilde{\epsilon}_n^2\}. \tag{S2.6}$$

Then we can conclude that for $\epsilon_n = \max\{\bar{\epsilon}_n, \widetilde{\epsilon}_n\}$ and sufficiently large $M > 0$, the posterior probability

$$\Pi_n(f_{\mu,\sigma} : d(f_{\mu,\sigma}, f_0) > M\epsilon_n | Y_1, \ldots, Y_n) \to 0 \text{ a.s. } P_{f_0},$$

where P_{f_0} denotes the true probability measure whose the Radon-Nikodym density is f_0. To proceed, we consider the Gaussian process $\mu \sim W^A$ given A, with A satisfying Assumption **P1**.

We will first verify (S2.6) along the lines of Ghosal & van der Vaart (2007). Recall f_β is defined as from (6), by Lemma S3.7 we guarantee that f_β is a well-defined density. Denote by $\mu_\beta = F_\beta^{-1}$ the quantile function of f_β, then we have $f_{\mu_\beta,\sigma} = \phi_\sigma * f_\beta$. Note that

$$h^2(f_0, f_{\mu,\sigma}) \precsim h^2(f_0, f_{\mu_\beta,\sigma}) + h^2(f_{\mu_\beta,\sigma}, f_{\mu,\sigma}). \tag{S2.7}$$

Under Assumptions **F1** and **F2** and by Lemma S3.8, one obtains

$$h^2(f_0, f_{\mu_\beta,\sigma}) \leq \int f_0 \log\left(\frac{f_0}{f_{\mu_\beta,\sigma}}\right) \precsim O(\sigma^{2\beta}). \tag{S2.8}$$

From Lemma S3.1 and the following remark, we obtain

$$h^2(f_{\mu_\beta,\sigma}, f_{\mu,\sigma}) \precsim \frac{\|\mu - \mu_\beta\|_\infty^2}{\sigma^2}. \tag{S2.9}$$

From Lemma 8 of Ghosal & van der Vaart (2007), one has

$$\int f_0 \log\left(\frac{f_0}{f_{\mu,\sigma}}\right)^i \leq h^2(f_0, f_{\mu,\sigma})\left(1 + \log\left\|\frac{f_0}{f_{\mu,\sigma}}\right\|_\infty\right)^i, \tag{S2.10}$$

for $i = 1, 2$.

From (S2.7)-(S2.10), for any $b \geq 1$ and $\tilde{\epsilon}_n^2 = \sigma_n^{2\beta}$,

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

Since $\mu_\beta \in C^{\beta+1}[0,1]$, from Section 5.1 of van der Vaart & van Zanten (2009),

$$\Pi_\mu(\|\mu - \mu_\beta\|_\infty \leq 2\delta_n) \geq C_4 \exp\left\{-C_5(1/\delta_n)^{\frac{1}{\beta+1}} \log\left(\frac{1}{\delta_n}\right)^{2\vee q}\right\}(C_6/\delta_n)^{(p+1)/(\beta+1)},$$

for $\delta_n \to 0$ and constants $C_4, C_5, C_6 > 0$. Letting $\delta_n = \sigma_n^{\beta+1}$, we obtain

$$\Pi_\mu(\|\mu - \mu_\beta\|_\infty \leq 2\delta_n) \geq \exp\left\{-C_7\left(\frac{1}{\sigma_n}\right)\log\left(\frac{1}{\sigma_n^{\beta+1}}\right)^{2\vee q}\right\},$$

for some constant $C_7 > 0$. Since $\sigma \sim IG(a_\sigma, b_\sigma)$, we have

$$\begin{aligned}
\Pi_\sigma(\sigma \in [\sigma_n, 2\sigma_n]) &= \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \int_{\sigma_n}^{2\sigma_n} x^{-(a_\sigma+1)} e^{-b_\sigma/x} dx \\
&\geq \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \int_{\sigma_n}^{2\sigma_n} e^{-2b_\sigma/x} dx \\
&\geq \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \sigma_n \exp\{-b_\sigma/\sigma_n\} \\
&\geq \exp\{-C_8/\sigma_n\},
\end{aligned}$$

for some constant $C_8 > 0$. Hence

$$\begin{aligned}
\Pi\{\sigma \in [\sigma_n, 2\sigma_n], \|\mu - \mu_\beta\|_\infty \precsim \sigma_n^{\beta+1}\} &\geq \exp\left\{-C_7\left(\frac{1}{\sigma_n}\right)\log\left(\frac{1}{\sigma_n^{\beta+1}}\right)^{2\vee q}\right\}\exp\{-C_8/\sigma_n\} \\
&\geq \exp\left\{-2C_9\left(\frac{1}{\sigma_n}\right)\log\left(\frac{1}{\sigma_n^{\beta+1}}\right)^{2\vee q}\right\}.
\end{aligned}$$

Then (S2.6) will be satisfied with $\widetilde{\epsilon}_n = n^{-\beta/(2\beta+1)} \log^{t_1}(n)$, where $t_1 = \beta(2 \vee q)/(2\beta+1)$ and some $C_9 > 0$. Next we construct a sequence of subsets $\mathcal{F}_n$ such that (S2.4) and (S2.5) are satisfied with $\bar{\epsilon}_n = n^{-\beta/(2\beta+1)} \log^{t_2} n$ and $\widetilde{\epsilon}_n$ for some global constant $t_2 > 0$.

Now we construct the sieves for $\mathcal{F}$. Letting $\mathbb{H}_1^a$ denote the unit ball of RKHS of the Gaussian process with rescaled parameter a and $\mathbb{B}_1$ denote the unit ball of $C[0,1]$ and given positive sequences M_n, r_n, define

$$B_n = \cup_{a<r_n}(M_n \mathbb{H}_1^a) + \bar{\delta}_n \mathbb{B}_1,$$

as in van der Vaart & van Zanten (2009), with $\bar{\delta}_n = \bar{\epsilon}_n l_n / K_1, K_1 = 2(2/\pi)^{1/2}$ and let

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

First we need to calculate $N(\bar{\epsilon}_n, \mathcal{F}_n, \|\cdot\|_1)$. Observe that for $\sigma_2 > \sigma_1 > \sigma_2/2$,

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

Taking $\kappa_n = \min\{\bar{\epsilon}_n/6, 1\}$ and $\sigma_m^n = l_n(1+\kappa_n)^m, m \geq 0$, we obtain a partition of $[l_n, h_n]$ as $l_n = \sigma_0^n < \sigma_1^n < \cdots < \sigma_{m_n-1}^n < h_n \leq \sigma_{m_n}^n$ with

$$m_n = \left(\log \frac{h_n}{l_n}\right) \frac{1}{\log(1+\kappa_n)} + 1. \tag{S2.11}$$

One can show that $3(\sigma_m^n - \sigma_{m-1}^n)/\sigma_{m-1}^n = 3\kappa_n \leq \bar{\epsilon}_n/2$. Let $\{\widetilde{\mu}_k^n, k = 1, \ldots, N(\bar{\delta}_n, B_n, \|\cdot\|_\infty)\}$ be a $\bar{\delta}_n$-net of B_n. Now consider the set

$$\{(\widetilde{\mu}_k^n, \sigma_m^n) : k = 1, \ldots, N(\bar{\delta}_n, B_n, \|\cdot\|_\infty),\ 0 \leq m \leq m_n\}. \tag{S2.12}$$

Then for any $f = f_{\mu,\sigma} \in \mathcal{F}_n$, we can find $(\widetilde{\mu}_k^n, \sigma_m^n)$ such that $\|\mu - \widetilde{\mu}_k^n\|_\infty < \bar{\delta}_n$. In addition, if one has $\sigma \in (\sigma_{m-1}^n, \sigma_m^n]$, then

$$\left\|f_{\mu,\sigma} - f_{\mu_k^n,\sigma_m^n}\right\|_1 \leq \bar{\epsilon}_n.$$

Hence the set in (S2.12) is an $\bar{\epsilon}_n$-net of $\mathcal{F}_n$ and its covering number is given by

$$m_n N(\bar{\delta}_n, B_n, \|\cdot\|_\infty).$$

From the proof of Theorem 3.1 in van der Vaart & van Zanten (2009), for any M_n, r_n with $r_n > 0$, we obtain

$$\log N(2\bar{\delta}_n, B_n, \|\cdot\|_\infty) \leq K_2 r_n \left(\log\left(\frac{M_n}{\bar{\delta}_n}\right)\right)^2. \tag{S2.13}$$

Again from the proof of Theorem 3.1 in van der Vaart & van Zanten (2009), for $r_n > 1$ and for $M_n^2 > 16 K_3 r_n (\log(r_n/\bar{\delta}_n))^2$, we have

$$\mathbb{P}(W^A \notin B_n) \leq \frac{K_4 r_n^p e^{-K_5 r_n \log^q r_n}}{K_5 \log^q r_n} + \exp\{-M_n^2/8\}, \tag{S2.14}$$

for constants $K_3, K_4, K_5 > 0$.

Next we calculate $\mathbb{P}(\sigma \notin [l_n, h_n])$. Observe that

$$\begin{aligned}
\mathbb{P}(\sigma \notin [l_n, h_n]) &= \mathbb{P}(\sigma^{-1} < h_n^{-1}) + \mathbb{P}(\sigma^{-1} > l_n^{-1}) \\
&\leq \sum_{k=\alpha_\sigma}^{\infty} \frac{e^{-b_\sigma h_n^{-1}} (b_\sigma h_n^{-1})^k}{k!} + \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \int_{l_n^{-1}}^{\infty} e^{-b_\sigma x/2} dx \\
&\leq e^{-a_\sigma \log(h_n)} + \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} e^{-b_\sigma l_n^{-1}/2}.
\end{aligned} \tag{S2.15}$$

Thus with $h_n = O(\exp\{n^{1/(2\beta+1)}(\log n)^{2t_1}\}), l_n = O(n^{-1/(2\beta+1)}(\log n)^{-2t_1}), r_n = O(n^{1/(2\beta+1)}(\log n)^{2t_1}), M_n = O(n^{1/(2\beta+1)}(\log n)^{t_1+1})$, (S2.14) and (S2.15) implies

$$\Pi(\mathcal{F}_n^c) = \exp\{-K_6 n\widetilde{\epsilon}_n^2\},$$

for some constant $K_6 > 0$, which guarantees that (S2.5) is satisfied with $\widetilde{\epsilon}_n = n^{-\beta/(2\beta+1)}(\log n)^{t_1}$.

Also with $\bar{\epsilon}_n = n^{-\beta/(2\beta+1)}(\log n)^{t_1+1}$, it follows from (S2.11) and (S2.13) that

$$\log N(\bar{\epsilon}_n, \mathcal{F}_n, \|\cdot\|_1) \leq K_7 n^{1/(2\beta+1)}(\log n)^{2t_1+2},$$

for some constant $K_7 > 0$. Hence $\max\{\bar{\epsilon}_n, \widetilde{\epsilon}_n\} = n^{-\beta/(2\beta+1)}(\log n)^{t_1+1}$. □

S2.5 Proof of Theorem 4.1

In this section, we present the detailed proof of the high probability bound for KL divergence between the true posterior and its α-VB approximation in the case of the GP-IVI.

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

The objective $m_n^*(\mathcal{Q}_n)$ can be bounded above by $D[q||p(Y^{(n)} \mid \theta)]$ for any $q \in \mathcal{Q}_n$. Choosing q as a particular univariate Gaussian centered at the true parameter with variance satisfying our assumptions **B1**-**B5** allows us to bound the KL divergence between the true posterior $p(Y^{(n)} \mid \theta)$ in high $\mathbb{P}_{\theta^*}^{(n)}$-probability.

Proof. It follows from the definition of $m_n^*(\mathcal{Q}_n)$ that for any $q \in \mathcal{Q}_n$

$$m_n^*(\mathcal{Q}_n) \leq D(q||p(\cdot \mid Y^{(n)})).$$

Choose μ_n to be the quantile function of the distribution $N(\theta^*, \sigma_n^2)$. Define the variational distribution

$$q_n(\theta) = \int \phi_\sigma(\theta - \mu_n(u))du,$$

where σ_n satisfies assumption **B2**. By change of measure,

$$\int \phi_\sigma(\theta - \mu_n(u))du = \int \phi_\sigma(\theta - t)\phi_{\sigma_n}(t - \theta^*)dt = N(\theta; \theta^*, \sigma^2 + \sigma_n^2).$$

Therefore $q_n(\theta) = N(\theta; \theta^*, \sigma^2 + \sigma_n^2) \in \mathcal{Q}_n$. Denote by $\mathbb{E}_n$ the mean respect to q_n. Expanding $D(q_n||p(Y^{(n)} \mid \theta))$,

$$\mathbb{E}_n\left[\log \frac{q_n(\theta)}{p(Y^{(n)} \mid \theta)(\theta)}\right] = \mathbb{E}_n[\log q_n] + \mathbb{E}_n[U(\theta)] + \log m(Y^{(n)}) - \mathbb{E}_n\left[L_n(\theta, \theta^*)\right],$$

where $L_n(\theta, \theta^*) = \sum_{i=1}^n \ell_i(\theta, \theta^*)$. Since the sum of $O_p(1)$ terms is $O_p(1)$, it suffices to show that each of the terms in the above sum is $O_p(1)$. The first term $\mathbb{E}_n[\log q_n]$, the differential entropy of q_n, is a constant and is $O_p(1)$. A straight forward application of Markov's inequality along with the fact that $\mathbb{E}_{\theta^*}^{(n)}[m(Y^{(n)})] = 1$ shows that $\log m(Y^{(n)})$ is $O_p(1)$.

Next, expand each of the functions $D(\theta^*||\theta)$, $\mu_2(\theta^*||\theta)$, and $U(\theta)$ using a multivariate Taylor expansion around θ^*. Applying assumptions **B4** and **B5** shows

$$\mathbb{E}_n[U(\theta)] \leq C_1(\sigma^2 + \sigma_n^2),$$
$$\mathbb{E}_n[\mu_2(\theta^*||\theta)] \leq C_2(\sigma^2 + \sigma_n^2), \tag{S2.16}$$
$$\mathbb{E}_n[D(\theta^*||\theta)] \leq C_u(\sigma^2 + \sigma_n^2), \tag{S2.17}$$
$$\mathbb{E}_n[D(\theta^*||\theta)] \geq C_\ell(\sigma^2 + \sigma_n^2). \tag{S2.18}$$

Markov's inequality shows that $U(\theta)$ is $O_p(1)$. We will use Chebychev's inequality to show $\mathbb{E}_n\left[\sum_{i=1}^n \ell_i(\theta,\theta^*)\right]$ is $O_p(1)$. Given $\varepsilon > 0$, choose $\delta = \left[C_2 c_0/(\varepsilon C_\ell)^2\right]^{1/2}$. Using (S2.16)-(S2.18) and noting that $-\mathbb{E}^{(n)}_{\theta^*}\{L_n(\theta,\theta^*)\} = nD(\theta^*||\theta)$, we have

$$\begin{aligned}
\mathbb{P}^{(n)}_{\theta^*}\left\{\mathbb{E}_n[L_n(\theta,\theta^*)] \le -C_u(1+\delta)n(\sigma^2+\sigma_n^2)\right\} &\le \mathbb{P}^{(n)}_{\theta^*}\left\{\mathbb{E}_n[L_n(\theta,\theta^*)] \le -(1+\delta)n\mathbb{E}_n[D(\theta^*||\theta)]\right\} \\
&\le \mathbb{P}^{(n)}_{\theta^*}\left\{\frac{1}{\sqrt{n}}\mathbb{E}_n[L_n(\theta,\theta^*) - \mathbb{E}^{(n)}_{\theta^*}\{L_n(\theta^*,\theta)\}] \le -\delta\sqrt{n}\mathbb{E}_n[D(\theta^*||\theta)]\right\} \\
&\le \frac{\mathrm{Var}^{(n)}_{\theta^*}\left(\mathbb{E}_n[\ell_1(\theta,\theta^*)]\right)}{\delta^2 n\left(\mathbb{E}_n[D(\theta^*||\theta)]\right)^2} \le \frac{\mathbb{E}_n[\mu_2(\theta^*||\theta^*)]}{\delta^2 n\left(\mathbb{E}_n[D(\theta^*||\theta)]\right)^2} \\
&\le \frac{C_2(\sigma^2+\sigma_n^2)}{\delta^2 n C_\ell(\sigma^2+\sigma_n^2)^2} \le \frac{C_2}{\delta^2 n C_\ell^2(\sigma^2+\sigma_n^2)} \le \frac{C_2}{\delta^2 n C_\ell^2 \sigma_n^2}.
\end{aligned}$$

Applying assumption **B2** we have $c_0^{-1/2}n^{-1/2} \le \sigma_n \le n^{-1/2}$. This gives

$$\mathbb{P}^{(n)}_{\theta^*}\left\{\int L_n(\theta,\theta^*)q_n(\theta)d\theta \le -2C_u(1+(C_2c_0/(\varepsilon C_\ell^2))^{1/2})\right\} \le \mathbb{P}^{(n)}_{\theta^*}\left\{\int L_n(\theta,\theta^*)q_n(\theta)d\theta \le -C_u(1+\delta)n(\sigma^2+\sigma_n^2)\right\} \le \varepsilon.$$

Thus $\mathbb{E}_n[L_n(\theta,\theta^*)]$ is $O_p(1)$. This completes the proof. □

S2.6 Proof of Theorem 4.2

In this section, we present the detailed proof of the Bayesian risk bound for α-variational inference in the case of the GP-IVI model. We also present a proof of the corollary for the Hellinger risk bound. The main theorem and the lemmas are restated here for convenience. Our risk bound is based of the following theorem,

Theorem S2.1 (Yang et al. (2020))**.** *For any $\zeta \in (0,1)$, it holds with $\mathbb{P}^{(n)}_{\theta^*}$-probability at least $(1-\zeta)$ that for any probability measure $q \in \mathcal{Q}$ with $q \ll p_\theta$,*

$$\int \frac{1}{n} D_\alpha[p_\theta^{(n)}||p_{\theta^*}^{(n)}]\widehat{q}(\theta)d\theta \le \frac{\alpha\Psi(q) + \log(1/\zeta)}{n(1-\alpha)}.$$

The GP-IVI risk bound is stated as follows.

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

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

The desired risk bound follows from bounding the right hand side of Theorem 3.2 of Yang et al. (2020)

$$\frac{\alpha}{n(1-\alpha)}\Psi(q_{\mu,\sigma}) := \frac{\alpha}{n(1-\alpha)}\left[\int q_{\mu,\sigma}(\theta)\log\frac{p(Y^{(n)} \mid \theta^*)}{p(Y^{(n)} \mid \theta)}d\theta + \frac{1}{\alpha}D(q_{\mu,\sigma}||p_\theta)\right]$$

in high $\mathbb{P}^{(n)}_{\theta^*}$-probability in terms of the local Bayesian complexity $\log \mathbb{P}_\theta(B_n(\theta^*,\varepsilon))$. By choosing a particular member of the variational family we can bound both the likelihood ratio integral as well as the KL divergence between the prior and the variational approximation. The relation between the variational distribution and the local Bayesian complexity come from the KL divergence term.

Proof. We will construct a special choice of μ as follows. Denote $p_\theta(\theta) = f_0(\theta)$. Let $B_n(\theta^*,\varepsilon)$ be as in (12). Define the truncated densities

$$\widetilde{f}_0(t) = \frac{f_0(t)I_{B_n(\theta^*,\varepsilon)}(t)}{\int_{B_n(\theta^*,\varepsilon)} f_0(u)du} = \frac{f_0(t)I_{B_n(\theta^*,\varepsilon)}(t)}{\mathbb{P}_\theta(B_n(\theta^*,\varepsilon))}, \quad \widetilde{f}_\beta(t) = \frac{f_\beta(t)I_{B_n(\theta^*,\varepsilon)}(t)}{\int_{B_n(\theta^*,\varepsilon)} f_\beta(u)du},$$

where f_β is constructed by procedure (6) such that $\|\phi_\sigma * f_\beta - f_0\|_\infty = O(\sigma^\beta)$ along with its associated distribution functions

$$\widetilde{F}_0(t) = \int_{(-\infty,t]\cap B_n(\theta^*,\varepsilon)} \widetilde{f}_0(t)dt, \quad \widetilde{F}_\beta(t) = \int_{(-\infty,t]\cap B_n(\theta^*,\varepsilon)} \widetilde{f}_\beta(t)dt.$$

Define the quantile function of $\widetilde{F}_\beta$ as $\widetilde{\mu}(t) = \widetilde{F}_\beta^{-1}(t)$. This can be used to define the variational density

$$q_{\widetilde{f}_\beta,\sigma}(\theta) = \int_{[0,1]} \phi_\sigma(\theta - \widetilde{\mu}(\eta))d\eta = \int_{-\infty}^{\infty} \phi_\sigma(\theta - t)\widetilde{f}_\beta(t)dt = \phi_\sigma * \widetilde{f}_\beta(\theta),$$

with $\sigma > 0$ a bandwidth that will be specified later in the proof. The main tool for the proof will be from Proposition 2.1

$$q_{\widetilde{f}_\beta,\sigma}(\theta) = \phi_\sigma * \widetilde{f}_\beta(\theta) \leq \widetilde{f}_0(\theta)(1 + D(\theta)O(\sigma^\beta)). \tag{S2.19}$$

Denote $M_D = \sup_{B_n(\theta^*,\varepsilon)} D(\theta)$ and $K_\beta(\sigma) = 1 + M_D O(\sigma^\beta)$. We will now bound the model-fit term. Denote the random variable

$$H(Y^{(n)}, \widetilde{f}_\beta, \sigma) = \int q_{\widetilde{f}_\beta,\sigma}(\theta) \log[p(Y^{(n)} \mid \theta^*)/p(Y^{(n)} \mid \theta)]d\theta.$$

The mean and variance (with respect to the data generating distribution) of the model-fit term are bounded by applying (S2.19),

$$\begin{aligned}
\mathbb{E}_{\theta^*}^{(n)}[H(Y^{(n)}, \widetilde{f}_\beta, \sigma)] &= \int D[p(Y^{(n)} \mid \theta^*)||p(Y^{(n)} \mid \theta)]q_{\widetilde{f}_\beta,\sigma}(\theta)d\theta \\
&\leq \int D[p(Y^{(n)} \mid \theta^*)||p(Y^{(n)} \mid \theta)]\widetilde{f}_0(\theta)(1 + D(\theta)O(\sigma^\beta))d\theta \\
&\leq K_\beta(\sigma) \int_{B(\theta^*,\varepsilon)} D[p(Y^{(n)} \mid \theta^*)||p(Y^{(n)} \mid \theta)] \frac{f_0(\theta)}{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}d\theta \\
&\leq K_\beta(\sigma) n\varepsilon^2,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}_{\theta^*}^{(n)}[H(Y^{(n)}, \widetilde{\mu}, \sigma)] &\leq \int V[p(Y^{(n)} \mid \theta^*)||p(Y^{(n)} \mid \theta)]q_{\widetilde{f}_\beta,\sigma}(\theta)d\theta \\
&\leq \int V[p(Y^{(n)} \mid \theta^*)||p(Y^{(n)} \mid \theta)]\widetilde{f}_0(\theta)(1 + D(\theta)O(\sigma^\beta))d\theta \\
&\leq K_\beta(\sigma) \int_{B(\theta^*,\varepsilon)} V[p(Y^{(n)} \mid \theta^*)||p(Y^{(n)} \mid \theta)] \frac{f_0(\theta)}{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}d\theta \\
&\leq K_\beta(\sigma) n\varepsilon^2.
\end{aligned}$$

It follows from Chebyshev's inequality that with $\mathbb{P}_{\theta^*}^{(n)}$-probability at least $1 - 1/[(D-1)^2 K_\beta(\sigma) n\varepsilon^2]$

$$\int q_{\widetilde{f}_\beta,\sigma}(\theta) \log\left[\frac{p(Y^{(n)} \mid \theta^*)}{p(Y^{(n)} \mid \theta)}\right] d\theta \leq D K_\beta(\sigma) n\varepsilon^2.$$

Next we will bound the regularization in terms of the local Bayesian complexity. Using (S2.19) we can bound the KL divergence,

$$D[q_{\widetilde{f}_\beta,\sigma}||p_\theta] = \int q_{\widetilde{f}_\beta,\sigma}(\theta) \log\left[\frac{q_{\widetilde{f}_\beta,\sigma}(\theta)}{f_0(\theta)}\right] d\theta \leq \int \log\left[\frac{\widetilde{f}_0(\theta)(1 + O(D(\theta)\sigma^\beta))}{f_0(\theta)}\right] \widetilde{f}_0(\theta)(1 + O(D(\theta)\sigma^\beta))d\theta.$$

Expanding $\widetilde{f}_0(\theta)$ and making use of the convention $I_{B_n(\theta^*,\varepsilon)}(\theta)\log(I_{B_n(\theta^*,\varepsilon)}(\theta)) = 0$ for $\theta \notin B_n(\theta^*,\varepsilon)$ we have

$$\begin{aligned}
&\int \log\left[\frac{f_0(\theta)I_{B_n(\theta^*,\varepsilon)}(1+O(D(\theta)\sigma^\beta))}{f_0(\theta)\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}\right]\frac{f_0(\theta)I_{B_n(\theta^*,\varepsilon)}}{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}(1+O(D(\theta)\sigma^\beta))d\theta \\
&\quad= \int_{B_n(\theta^*,\varepsilon)} \log\left[\frac{(1+O(D(\theta)\sigma^\beta))}{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}\right]\frac{f_0(\theta)}{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}(1+O(D(\theta)\sigma^\beta))d\theta \\
&\quad\leq K_\beta(\sigma)\log\left[\frac{K_\beta(\sigma)}{\mathbb{P}_\theta(B_n(\theta^*,\varepsilon))}\right]\int_{B_n(\theta^*,\varepsilon)}\frac{f_0(\theta)}{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]}d\theta \\
&\quad= K_\beta(\sigma)\log\left[\frac{K_\beta(\sigma)}{\mathbb{P}_\theta(B_n(\theta^*,\varepsilon))}\right].
\end{aligned}$$

Combining the bounds from both parts, we have with probability at least $1 - 1/[(D-1)^2K_\beta(\sigma)n\varepsilon^2]$ that

$$\Psi(q_{\widetilde{f}_\beta,\sigma}) \leq DK_\beta(\sigma)n\varepsilon^2 + \alpha^{-1}K_\beta(\sigma)\log K_\beta(\sigma) + \alpha^{-1}K_\beta(\sigma)\log\left\{\mathbb{P}_\theta[B_n(\theta^*,\varepsilon)]^{-1}\right\}.$$

Choosing $\zeta = 1/[(D-1)^2K_\beta(\sigma)n\varepsilon^2]$. It follows from the union bound for probabilities, we have with probability at least $1 - 2/[(D-1)^2K_\beta(\sigma)n\varepsilon^2]$ that

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

Recall that $K_\beta(\sigma) = 1 + O(\sigma^\beta)$. Choosing $\sigma = n^{-2/\beta}$ gives

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

□

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.

Proof. For IID data $n^{-1}D_\alpha^{(n)}(\theta,\theta^*) = D_\alpha[p_\theta||p_{\theta^*}]$. Applying Theorem 4.2 with $\varepsilon = n^{-1}$ and Assumption **A1** yields,

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

Combining the above with the fact that $\max\{1,(1-\alpha)^{-1}\alpha\}h^2(p,q) \leq D_\alpha[p||q]$ competes the proof. □

S3 Auxiliary results

In this section, we summarize results used in the proofs of main theorems in the main document. First to guarantee that the model (2) leads to the optimal rate of convergence, we start from deriving sharp bounds for the Hellinger distance between f_{μ_1,σ_1} and f_{μ_2,σ_2} for $\mu_1, \mu_2 \in C[0,1]$ and $\sigma_1, \sigma_2 > 0$. We summarize the result in the following Lemma S3.1.

Lemma S3.1. *For $\mu_1, \mu_2 \in C[0,1]$ and $\sigma_1, \sigma_2 > 0$,*

$$h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\Big\{ - \frac{\|\mu_1-\mu_2\|_\infty^2}{4(\sigma_1^2+\sigma_2^2)} \Big\}. \tag{S3.1}$$

Proof. Note that by Hölder's inequality,

$$f_{\mu_1,\sigma_1}(y) f_{\mu_2,\sigma_2}(y) \geq \Big\{ \int_0^1 \sqrt{\phi_{\sigma_1}(y-\mu_1(x))}\sqrt{\phi_{\sigma_2}(y-\mu_2(x))}dx \Big\}^2.$$

Hence,

$$\begin{aligned} h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq \int \Big[\int_0^1 \phi_{\sigma_1}(y-\mu_1(x))dx + \int_0^1 \phi_{\sigma_2}(y-\mu_2(x))dx \\ - 2\int_0^1 \sqrt{\phi_{\sigma_1}(y-\mu_1(x))}\sqrt{\phi_{\sigma_2}(y-\mu_2(x))}dx \Big] dy. \end{aligned}$$

By changing the order of integration (applying Fubini's theorem since the function within the integral is jointly integrable) we get

$$\begin{aligned} h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) &\leq \int_0^1 h^2(f_{\mu_1(x),\sigma_1}, f_{\mu_2(x),\sigma_2})dx \\ &= \int_0^1 \Big[1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\Big\{ - \frac{(\mu_1(x)-\mu_2(x))^2}{4(\sigma_1^2+\sigma_2^2)} \Big\} \Big] dx \\ &\leq 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\Big\{ - \frac{\|\mu_1-\mu_2\|_\infty^2}{4(\sigma_1^2+\sigma_2^2)} \Big\}. \end{aligned}$$

□

Remark S3.2. *When $\sigma_1 = \sigma_2 = \sigma$, $h^2(f_{\mu_1,\sigma}, f_{\mu_2,\sigma}) \leq 1 - \exp\big\{ \|\mu_1-\mu_2\|_\infty^2 / 8\sigma^2 \big\}$, which implies that $h^2(f_{\mu_1,\sigma}, f_{\mu_2,\sigma}) \precsim \|\mu_1-\mu_2\|_\infty^2 / \sigma^2$.*

Remark S3.3. *The standard inequality $h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq \|f_{\mu_1,\sigma_1} - f_{\mu_2,\sigma_2}\|_1$ relating the Hellinger distance to the total variation distance leads to the cruder bound*

$$h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq C_1 \frac{\|\mu_1-\mu_2\|_\infty}{(\sigma_1 \wedge \sigma_2)} + C_2 \frac{|\sigma_2-\sigma_1|}{(\sigma_1 \wedge \sigma_2)},$$

which is linear in $\|\mu_1-\mu_2\|_\infty$. This bound is less sharp than what is obtained in Lemma S3.1 and does not suffice for obtaining the optimal rate of convergence.

In order to apply Lemma 8 in Ghosal & van der Vaart (2007) to control the Kullback–Leibler divergence between the true density f_0 and the model $f_{\mu,\sigma}$, we derive an upper bound for $\log \|f_0/f_{\mu,\sigma}\|_\infty$ in Lemma S3.4.

Lemma S3.4. *If f_0 satisfies Assumption **F2**,*

$$\log \Big\| \frac{f_0}{f_{\mu,\sigma}} \Big\|_\infty \leq C + \frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2} \tag{S3.2}$$

for some constant $C > 0$.

Proof. Note that

$$\begin{aligned} f_{\mu,\sigma}(y) &= \frac{1}{\sqrt{2\pi}\sigma}\int_0^1 \exp\left\{-\frac{(y-\mu(x))^2}{2\sigma^2}\right\}dx \\ &\geq \frac{1}{\sqrt{2\pi}\sigma}\int_0^1 \exp\left\{-\frac{(y-\mu_0(x))^2}{\sigma^2}\right\}dx \exp\left\{-\frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2}\right\} \\ &\geq C\phi_{\sigma/\sqrt{2}} * f_0(y)\exp\left\{-\frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2}\right\} \\ &\geq Cf_0(y)\exp\left\{-\frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2}\right\}, \end{aligned}$$

where the last inequality follows from Lemma 6 of Ghosal & van der Vaart (2007) since f_0 is compactly supported by Assumption **F2**. This provides the desired inequality. □

Lemma S3.5. *Let $j \geq 0$ be the integer such that $\beta \in (2j, 2j+2]$, and the sequence of f_j is constructed by the procedure in (6). Then we have $f_\beta = \sum_{i=0}^{j}(-1)^i\binom{j+1}{i+1}\phi_\sigma^{(i)} * f_0$, where $\phi_\sigma^{(i)} * f_0 = \phi_\sigma * \cdots * \phi_\sigma * f_0$, the i-fold convolution of ϕ_σ with f_0.*

Proof. Consider f_j constructed by (6). When $j = 1$, $f_1 = 2f_0 - \phi_\sigma * f_0$, so the form holds. By induction, suppose this form holds for $j > 1$, then

$$\begin{aligned} f_{j+1} &= f_0 - (\phi_\sigma * f_j - f_j) \\ &= f_0 + \sum_{i=0}^{j}(-1)^{i+1}\binom{j+1}{i+1}\phi_\sigma^{(i+1)} * f_0 + \sum_{i=0}^{j}(-1)^i\binom{j+1}{i+1}\phi_\sigma^{(i)} * f_0 \\ &= (j+2)f_0 + \sum_{i=1}^{j+1}(-1)^i\binom{j+1}{i+1}\phi_\sigma^{(i)} * f_0 + \sum_{i=1}^{j}(-1)^i\binom{j+1}{i}\phi_\sigma^{(i)} * f_0 \\ &= (j+2)f_0 + \sum_{i=1}^{j}(-1)^i\left(\binom{j+1}{i+1} + \binom{j+1}{i}\right)\phi_\sigma^{(i)} * f_0 + (-1)^{j+1}\phi_\sigma^{(i+1)} * f_0 \\ &= (j+2)f_0 + \sum_{i=1}^{j}(-1)^i\binom{j+2}{i+1}\phi_\sigma^{(i)} * f_0 + (-1)^{j+1}\phi_\sigma^{(i+1)} * f_0 \\ &= \sum_{i=0}^{j+1}(-1)^i\binom{j+2}{i+1}\phi_\sigma^{(i)} * f_0. \end{aligned}$$

It holds for $j+1$, which completes the proof. □

Lemma S3.6. *Let f_0 satisfy Assumptions **F1** and **F2**. With $A_\sigma = \{x : f_0(x) \geq \sigma^H\}$, we have*

$$\int_{A_\sigma^c} f_0(x)dx = O(\sigma^{2\beta}), \quad \int_{A_\sigma^c} \phi_\sigma * f_j(x)dx = O(\sigma^{2\beta}), \tag{S3.3}$$

for all non-negative integer j, sufficiently small σ and sufficiently large H.

Proof. Under Assumption **F2** there exists $(a, b) \subset [0, 1]$ such that $A_\sigma^c \subset [0, a) \cup (b, 1]$ if we choose σ sufficiently small, so that $f_0(x) \leq \sigma^H$ for $x \in A_\sigma^c$. Therefore, $\int_{A_\sigma^c} f_0(x) \leq \sigma^H \leq O(\sigma^{2\beta})$ if we choose $H \geq 2\beta$. Using Proposition 2.1,

$$\int_{A_\sigma^c} \phi_\sigma * f_j(x)dx = \int_{A_\sigma^c} f_0(x)\{1 + O(D(x)\sigma^\beta)\} \leq O(\sigma^H).$$

With bounded $D(x)$ and $H \geq 2\beta$ it is easy to bound the second integral in (S3.3) by $O(\sigma^{2\beta})$. □

Lemma S3.7. *Suppose f_0 satisfies Assumptions **F1** and **F2**. For $\beta > 2$ and the integer j such that $\beta \in (2j, 2j+2]$, f_β is a density function.*

Proof. To show f_β is a density function, it suffices to show f_β is non-negative, since a simple calculation shows that $\int f_\beta = 1$ for $j \geq 0$. Following the proof of Lemma 2 in Kruijer et al. (2010), we treat $\log f_0$ as a function in $C^2[0,1]$ and obtain the same form of $\phi_\sigma * f_0$ as in (S2.1). For small enough σ we can find $\rho_1 \in (0,1)$ very close to 0 such that

$$\phi_\sigma * f_0(x) = f_0(x)(1 + O(D^{(2)}(x)\sigma^2)) < f_0(x)(1+\rho_1),$$

where $D^{(2)}$ contains $|l_1(x)|$ and $|l_2(x)|$ to certain power, so $D^{(2)}$ is bounded. Then we have

$$f_1(x) = 2f_0(x) - K_\sigma f_0(x) > 2f_0(x) - f_0(x)(1+\rho_1) = f_0(x)(1-\rho_1).$$

Then we treat $\log f_0$ as a function with $\beta = 4$, $j = 1$. Similarly, we can get

$$\phi_\sigma * f_1(x) = f_0(x)(1 + O(D^{(4)}(x)\sigma^4)),$$

where $D^{(4)}$ contains $|l_1(x)|, \ldots, |l_4(x)|$. We can find $0 < \rho_2 < \rho_1$ such that $\phi_\sigma * f_1(x) < f_0(x)(1+\rho_2)$, then can get

$$f_2(x) = f_0(x) - (\phi_\sigma * f_1(x) - f_1(x)) > f_0(x)(1-\rho_1-\rho_2) > f_0(x)(1-2\rho_1).$$

Continuing this procedure, we can get $f_j(x) > f_0(x)(1-j\rho_1)$ with sufficiently small σ and $1 - j\rho_1 \in (0,1)$ and it is close to 1. Then we show f_j is non-negative.

□

Lemma S3.8. *Let f_0 satisfy Assumptions **F1** and **F2** and let j be the integer such that $\beta \in (2j, 2j+2]$. Then we show that the density f_β obtained by (6) satisfies*

$$\int f_0(x) \log \frac{f_0(x)}{\phi_\sigma * f_\beta(x)} = O(\sigma^{2\beta}), \tag{S3.4}$$

for sufficiently small σ and all $x \in [0,1]$.

Proof. Again consider the set $A_\sigma = \{x : f_0(x) \geq \sigma^H\}$ with arbitrarily large H. We separate the Kullback–Leibler divergence into

$$\begin{aligned}\int_{[0,1]} f_0 \log \frac{f_0}{\phi_\sigma * f_\beta} &= \int_{[0,1]\cap A_\sigma} f_0 \log \frac{f_0}{\phi_\sigma * f_\beta} + \int_{[0,1]\cap A_\sigma^c} f_0 \log \frac{f_0}{\phi_\sigma * f_\beta} \\ &\leq \int_{A_\sigma} \frac{(f_0 - \phi_\sigma * f_\beta)^2}{\phi_\sigma * f_\beta} + \int_{A_\sigma^c} (\phi_\sigma * f_\beta - f_0) + \int_{A_\sigma^c} f_0 \log \frac{f_0}{\phi_\sigma * f_\beta}. \end{aligned} \tag{S3.5}$$

Under Assumption **F2** and by Remark 3 in Ghosal et al. (1999), for small enough σ there exists a constant C such that $\phi_\sigma * f_0 \geq C f_0$ for all $x \in [0,1]$. Especially, f_0 satisfies $\phi_\sigma * f_0 \geq f_0/3$ for $x \in A_\sigma^c$. Also in the proof of Lemma S3.7 we can find $\rho \in (0,1)$ such that $f_\beta > \rho f_0$. Then, on set A_σ with sufficiently small σ, we have

$$\phi_\sigma * f_j \geq \rho \phi_\sigma * f_0 \geq K f_0,$$

where $K = \min\{\rho/3, \rho C\}$. Applying (S2.1), the first integral on the r.h.s. of (S3.5) can be bounded by

$$\begin{aligned}\int_{A_\sigma} \frac{(f_0 - \phi_\sigma * f_j)^2}{\phi_\sigma * f_j} &\leq \int_{A_\sigma} \frac{[f_0(x) - f_0(x)(1 + O(D(x)\sigma^\beta))]^2}{K f_0(x)} \\ &\precsim \int_{A_\sigma} f_0(x) O(D^2(x)\sigma^{2\beta}) = O(\sigma^{2\beta}).\end{aligned}$$

To bound the second integral of r.h.s in (S3.5), according to Remark 3 in Ghosal et al. (1999) we get $\phi_\sigma * f_j \geq \rho f_0/3$, then we can find a constant $C < 1$ such that $\phi_\sigma * f_j \geq C f_0$. The second and third term in (S3.5) can be bounded by $O(\sigma^{2\beta})$ based on Lemma S3.6. □

Lemma S3.9. *Let $\mathbb{H}_1^a$ denote the unit ball of RKHS of the Gaussian process with rescaled parameter a and $\mathbb{B}_1$ be the unit ball of $C[0,1]$. For $r > 1$, there exists a constant K, such that for $\epsilon < 1/2$,*

$$\log N(\epsilon, \cup_{a\in[0,r]}\mathbb{H}_1^a, \|\cdot\|_\infty) \le Kr\left(\log\frac{1}{\epsilon}\right)^2. \tag{S3.6}$$

Proof. Since we can write any element of $\mathbb{H}_1^a$ as a function of $\mathrm{Re}(z)$ by Lemma 4.5 in van der Vaart & van Zanten (2009), and an ϵ-net denoted by $\mathcal{F}^a$ over $\mathbb{H}_1^a$ is constructed through a finite set of piece-wise polynomial functions, and according to Lemma 4.4 and Lemma 4.5 in Bhattacharya et al. (2014), $\mathcal{F}^a$ also forms an ϵ-net over $\mathbb{H}_1^b$ as long as a is sufficiently close to b. Thus we can find one set $\Gamma = \{a_i, i = 1, \ldots, k\}$ with $k = \lfloor r \rfloor + 1$ and $a_k = r$, such that for any $b \in [0, r]$ there exists some a_i satisfying $|b - a_i| \le 1$, so that $\cup_{i\le k}\mathcal{F}^{a_i}$ forms an ϵ-net over $\cup_{a\le r}\mathbb{H}_1^a$. Since the covering number of $\cup_{i\le k}\mathcal{F}^{a_i}$ is bounded by summation of covering number of $\mathcal{F}^{a_i}$, we obtain

$$\log N\big(\epsilon, \cup_{a\in[0,r]}\mathbb{H}_1^a, \|\cdot\|_\infty\big) \le \log\left(\sum_{i=1}^{k}\#(\mathcal{F}^{a_i})\right) \le \log(k\cdot\#(\mathcal{F}^r)) \le Kr\left(\log\frac{1}{\epsilon}\right)^2.$$

Here we write $\#(A)$ to denote the cardinality of any arbitrary set A. To prove the second inequality above, note that the piece-wise polynomials are constructed on the partition over $[0,1]$, denoted by $\cup_{i\le m}B_i$, where B_i's are disjoint interval with length R that can be considered as a non-increasing function of a, so the total number of polynomials is non-decreasing in a. Also we find that when building the mesh grid of the coefficients of polynomials in each B_i, both the approximation error and tail estimate are invariant to interval length R, therefore we have $\#(\mathcal{F}^a) \le \#(\mathcal{F}^b)$ if $a \le b$, for $a, b \in [0, r]$. □

Remark S3.10. *With larger a we need a finer partition on $[0,1]$ while the grid of coefficients of piece-wise polynomial remains the same except the range and the meshwidth will change together along with a. Since we can see the element h of RKHS ball as a function of it and with Cauchy formula we can bound the derivatives of h by C/R^n, where $|h|^2 \le C^2$.*

S4 GP-IVI Algorithm

In this section we outline an algorithm to train GP-IVI based on the Karhunen–Loéve representation of a Gaussian process; details on the Karhunen–Loéve representation of a stochastic process can be found in either Jin (2014) or Le Maître & Knio (2010).

S4.1 Karhunen–Loéve representation of a Gaussian process

For a mean zero Guassian process $X(t)$, $0 \le t \le 1$, with covariance function

$$K(s,t) = \mathbb{E}[X(t)X(s)], \text{ for } 0 \le s, t \le 1.$$

The Karhunen–Loéve expansion is given by

$$X(t) = \sum_{k=1}^{\infty}\sqrt{\lambda_k}e_k(t)\xi_k,$$

where $\{(\lambda_k, e_k)\}$ are the eigenvalue eigenfunction pairs to the Fredholm integral equation

$$\lambda_k e_k(t) = \int_0^1 K(s,t)e_k(s)ds, \text{ for } 0 \le t \le 1,$$

and ξ_k are IID $N(0,1)$ random variables. For computational purposes, we need work with the finite approximation

$$X_N(t) = \sum_{k=1}^{N}\sqrt{\lambda_k}\xi_k e_k(t).$$

S4.2 Algorithm

Recall the GP-IVI family consists of distributions of the form,

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

Substituting in the truncated Karhunen–Loéve expansion in place of $\mu(\eta)$ we can equivalently define $q_{\mu,\sigma}(\theta) = \mathbb{E}_\eta[N(\theta; \mu(\eta), \sigma^2)]$ using the reparameterization trick

$$\theta = \sum_{k=1}^{N} \sqrt{\lambda_k}\xi_k e_k(\eta) + \sigma\varepsilon \tag{S4.1}$$

$$q_{\mu,\sigma}(\theta) = \mathbb{E}_\eta \left[\exp\left\{ -\frac{1}{2\sigma^2} \left(\theta - \sum_{k=1}^{N} \sqrt{\lambda_k}\xi_k e_k(\eta) \right)^2 \right\} \right], \tag{S4.2}$$

where $\xi_k \overset{iid}{\sim} N(0,1)$ for $1 \leq k \leq N$, $\varepsilon \sim N(0,1)$, and $\eta \sim U(0,1)$. This allows us to define the joint ELBO in $(\sigma, \xi_1, \ldots, \xi_N)$,

$$\text{ELBO}(\sigma, \xi_1, \ldots, \xi_N) = \mathbb{E}_{q_{\mu,\sigma}(\theta)}[\log p(\theta, Y^{(n)}) - \log q_{\mu,\sigma}(\theta)] \tag{S4.3}$$

and its gradient

$$\nabla_{\sigma,\xi_1,\ldots,\xi_N} \text{ELBO}(\sigma, \xi_1, \ldots, \xi_N).$$

At this point we can compute the ELBO and its gradient using Monte Carlo techniques and maximize the ELBO using a gradient-based optimization technique.

References

Bhattacharya, A., Pati, D., & Dunson, D. (2014). Anisotropic function estimation using multi-bandwidth gaussian processes. *Annals of statistics*, *42*(1), 352.

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.

Jin, S. (2014). *Gaussian processes: Karhunen-loeve expansion, small ball estimates and applications in time series models* (Unpublished doctoral dissertation). University of Delaware.

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

Le Maître, O., & Knio, O. (2010). *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media.

Rasmussen, C. (2003). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63–71).

van der Vaart, A., & van Zanten, J. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections*, *3*, 200–222.

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.

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