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 $\bar{\epsilon}_n, \tilde{\epsilon}_n \to 0$ with $n \min\{\bar{\epsilon}_n^2, \tilde{\epsilon}_n^2\} \to \infty$ 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,

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

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

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

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 $c_i, i = 1, \dots, r+1$, 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 $\phi_\sigma * f_0(x)$, $\phi_\sigma^{(2)} * f_0(x)$, $\dots$, $\phi_\sigma^{(j)} * f_0(x)$, 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, $\dots$, 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 $\delta = \epsilon\,\sigma_0/4$,

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

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,

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

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

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

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

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

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 $\tilde{\epsilon}_n^2 = \sigma_n^{2\beta}$,

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

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 $\sigma_2 > \sigma_1 > \sigma_2/2$,

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

Taking $\kappa_n = \min\{\bar{\epsilon}_n/6, 1\}$ and $\sigma_m^n = l_n(1+\kappa_n)^m, m \ge 0$, we obtain a partition of $[l_n, h_n]$ as $l_n = \sigma_0^n < \sigma_1^n < \cdots < \sigma_{m_n-1}^n < h_n \le \sigma_{m_n}^n$ 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 $3(\sigma_m^n - \sigma_{m-1}^n)/\sigma_{m-1}^n = 3\kappa_n \le \bar{\epsilon}_n/2$. 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

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

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

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

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

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

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

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

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

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

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

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 $m_n^*(\mathcal{Q}_n) = \min_{q\in\mathcal{Q}_n}\{D[q||p(\cdot \mid Y^{(n)})]\}$ 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 $-\mathbb{E}_{\theta^*}^{(n)}\{L_n(\theta,\theta^*)\} = nD(\theta^*||\theta)$, we have

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

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

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

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 $\mathbb{P}_{\theta^*}^{(n)}$-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 $\mathbb{P}_{\theta^*}^{(n)}$-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 $\mathbb{P}_{\theta^*}^{(n)}$-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 $1 - 1/[(D-1)^2 K_\beta(\sigma)n\varepsilon^2]$

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

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

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

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

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

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

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

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 $\epsilon < 1/2$,*

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

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 $|b - a_i| \leq 1$, so that $\cup_{i\leq k}\mathcal{F}^{a_i}$ forms an ϵ-net over $\cup_{a\leq r}\mathbb{H}_1^a$. Since the covering number of $\cup_{i\leq k}\mathcal{F}^{a_i}$ is bounded by summation of covering number of $\mathcal{F}^{a_i}$, we obtain

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

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 $\cup_{i\leq m}B_i$, 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 $\#(\mathcal{F}^a) \leq \#(\mathcal{F}^b)$ if $a \leq b$, 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 $|h|^2 \leq C^2$.*

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 $X(t)$, $0 \leq t \leq 1$, with covariance function

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

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

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

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 $(\sigma, \xi_1, \dots, \xi_N)$,

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

and its gradient

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

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.