Asymptotic Properties for Bayesian Neural Network in Besov Space

Kyeongwon Lee
Department of Statistics
Seoul National University
Seoul, Republic of Korea 08826
lkw1718@snu.ac.kr

Jaeyong Lee
Department of Statistics
Seoul National University
Seoul, Republic of Korea 08826
leejyc@gmail.com

A Posterior consistency

Lemma 1 (Lemma 2 in Ghosal and Van Der Vaart [2007])**.** *Suppose that observation $X^{(n)} = (X_1, \cdots, X_n)$ of independent observations X_i. Assume that the distribution $P_{\theta,i}$ of the ith component X_i has a density $p_{\theta,i}$ relative to a σ-finite measure μ_i. Let $P_\theta^{(n)} = \bigotimes_{i=1}^n P_{\theta,i}$ be the product measures and average Hellinger distance*

$$d_n^2(\theta_0, \theta) = \frac{1}{n}\sum_{i=1}^n \int (\sqrt{p_{\theta_0,i}} - \sqrt{p_{\theta,i}})^2 d\mu_i.$$

Then there exists test ϕ_n such that $P_{\theta_0}^{(n)}\phi_n \le e^{-\frac{1}{2}nd_n^2(\theta_0,\theta_1)}$ and $P_\theta^{(n)}(1-\phi_n) \le e^{-\frac{1}{2}nd_n^2(\theta_0,\theta_1)}$ for all $\theta \in \Theta$ such that $d_n(\theta,\theta_1) \le \frac{1}{18}d_n(\theta_0,\theta_1)$.

Proof. See Ghosal and Van Der Vaart [2007]. □

Let define divergences

$$K(f,g) = \mathbb{E}^f\left[\log\frac{f(X)}{g(X)}\right] = \int f\log\frac{f}{g}d\mu, \tag{1}$$

$$V_k(f,g) = \mathbb{E}^f\left[\left|\log\frac{f(X)}{g(X)}\right|^k\right], \tag{2}$$

$$V_{k,0}(f,g) = \mathbb{E}^f\left[\left|\log\frac{f(X)}{g(X)} - K(f,g)\right|^k\right] \tag{3}$$

and set

$$B_n^*(\theta_0,\epsilon;k) = \left\{\theta\in\Theta : \frac{1}{n}\sum_{i=1}^n K(P_{\theta_0,i}, P_{\theta,i}) \le \epsilon^2,\ \frac{1}{n}\sum_{i=1}^n V_{k,0}(P_{\theta_0,i}, P_{\theta,i}) \le C_k\epsilon^k\right\}.$$

Here, the C_k is the constant satisfying

$$\mathbb{E}\Big[\big|\bar{X}_n - \mathbb{E}\big[\bar{X}_n\big]\big|^k\Big] \le C_k n^{-k/2}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[|X_i|^k\Big]$$

for $k \ge 2$. The existence of C_k is guaranteed by Marcinkiewiz-Zygmund inequality. The following lemma gives a sufficient condition for obtaining posterior consistency.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Lemma 2 (Theorem 4 of Ghosal and Van Der Vaart [2007]). *Let $P_\theta^{(n)}$ be product measures and $d_n(\theta_0, \theta) = \frac{1}{n}\sum_{i=1}^n \int (\sqrt{p_{\theta_0,i}} - \sqrt{p_{\theta,i}})^2 d\mu_i$. Suppose that for a sequence $\epsilon_n \to 0$ such that $n\epsilon_n^2$ is bounded away from zero, some $k > 1$, all sufficiently large j and sets $\Theta_n \subset \Theta$ which satisfies following conditions:*

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, \{\theta \in \Theta_n : d_n(\theta, \theta_0) < \epsilon\}, d_n) \leq n\epsilon_n^2, \tag{4}$$

$$\frac{\Pi(\Theta - \Theta_n)}{\Pi(B_n^*(\theta_0, \epsilon_n; k))} = o\left(e^{-2n\epsilon_n^2}\right), \tag{5}$$

$$\frac{\Pi(\theta \in \Theta_n : j\epsilon_n < d_n(\theta, \theta_0) \leq 2j\epsilon_n)}{\Pi(B_n^*(\theta_0, \epsilon_n; k))} \leq e^{n\epsilon_n^2 j^2/4} \tag{6}$$

Then $P_{\theta_0}^{(n)}\left(\Pi(\theta : d_n(\theta, \theta_0) \geq M_n\epsilon_n | \mathbb{D}_n) \to 0\right.$ for any sequence $M_n \to \infty$.

Proof. See Ghosal and Van Der Vaart [2007]. □

In Ghosal and Van Der Vaart [2007], they mentioned that it can be replaced by any other distance d_n for which the conclusion of Lemma 1 holds. Moreover, for $k = 2$, Lemma 2 work with the smaller neighborhood

$$\overline{B}_n(\theta_0, \epsilon) = \left\{\theta \in \Theta : \frac{1}{n}\sum_{i=1}^n K(P_{\theta_0,i}, P_{\theta,i}) \leq \epsilon^2,\ \frac{1}{n}\sum_{i=1}^n V_{2,0}(P_{\theta_0,i}, P_{\theta,i}) \leq \epsilon^2\right\} \tag{7}$$

instead of $B_n^*(\theta_0, \epsilon_n; k)$.

The following is an adapted version of Lemma 2, which was introduced in Polson and Ročková [2018].

Lemma 3. *Assume model (1). Suppose that $\mathcal{F}$ is uniformly bounded. Let*

$$A_{\epsilon,M} := \{f \in \mathcal{F} : \|f - f_0\|_n \leq M\epsilon\}.$$

If there exist $C > 2/\sigma^2$ and $\mathcal{F}_n \subset \mathcal{F}$ such that

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \leq n\epsilon_n^2, \tag{8}$$

$$\Pi(A_{\epsilon_n,1}) \geq e^{-Cn\epsilon_n^2}, \tag{9}$$

$$\Pi(\mathcal{F} - \mathcal{F}_n) = o\left(e^{-(C\sigma^2+2)n\epsilon_n^2}\right) \tag{10}$$

for any $\epsilon_n \to 0$, $n\epsilon_n^2 \to \infty$,

$$\Pi(A_{\epsilon_n,M_n}^c | \mathbb{D}_n) \to 0$$

in $P_{f_0}^{(n)}$-probability as $n \to \infty$ for any $M_n \to \infty$.

Proof. See appendix C.1. □

B Consistency of neural network

The upper bound of the covering number of the neural network space is given by as follows.

Lemma 4 (Lemma 3 in Suzuki [2018]). *$\forall\ \epsilon > 0$,*

$$\log N(\epsilon, \Phi(L, W, S, B), \|\cdot\|_{L^\infty}) \leq (S+1)\log\left(2\epsilon^{-1}L(B \vee 1)^L(W+1)^{2L}\right) \tag{11}$$

for $W,\ L \geq 3$.

Proof. See appendix C.2. □

For $L,\ W,\ S \in \mathbb{N}$ and $B,\ a > 0$, define a function space

$$\Phi(L, W, S, B, a) = \{f_\theta : \theta \in \Theta(L, W, S, B, a)\}, \tag{12}$$

where

$$\Theta(L, W, S, B, a) = \left\{\theta : (\theta_i I(|\theta_i| > a))_{i=1}^{T_n} \in \Theta(L, W, S, B)\right\}.$$

Lemma 5. $\forall \epsilon \geq 2aL(B \vee 1)^{L-1}(W+1)^L$,

$$\log N(\epsilon, \Phi(L, W, S, B, a), \|\cdot\|_{L^\infty}) \leq (S+1)\log\big(2\epsilon^{-1}L(B \vee 1)^L(W+1)^{2L}\big) \tag{13}$$

for $W,\ L \geq 3$.

Proof. See appendix C.3. □

The following lemma shows that for any function f_0 in the Besov space, there exist neural networks which is closed enough to f_0.

Lemma 6 (Proposition 1 in Suzuki [2018])**.** *Suppose that* $0 < p, q, r \leq \infty$, $\omega := d(1/p - 1/r)_+ < s < \infty$ *and* $\nu = (s-\omega)/(2\omega)$. *Assume that* $N \in \mathbb{N}$ *is sufficiently large and* $m \in \mathbb{N}$ *satisfies* $0 < s < \min\{m, m-1+1/p\}$. *Let* $W_0 = 6dm(m+2) + 2d$. *Then,*

$$\sup_{f_0 \in U(B^s_{p,q}([0,1]^d))} \inf_{f \in \Phi(L,W,S,B)} \|f_0 - f\|_{L^r} \lesssim N^{-s/d} \tag{14}$$

for

$$\begin{aligned} L &= 3 + 2\lceil \log_2 \left(\frac{3^{d \vee m}}{\tau(N) c_{(d,m)}} \right) + 5 \rceil \lceil \log_2(d \vee m) \rceil, \quad W = NW_0, \\ S &= (L-1)W_0^2 N + N, \quad B = O\left(N^{(\nu^{-1}+d^{-1})(1 \vee (d/p-s)_+)} \right), \end{aligned} \tag{15}$$

where $U(\mathcal{H})$ *is the unit ball of a quasi-Banach space* $\mathcal{H}$, $c_{(d,m)} = \left(1 + 2de\frac{(2e)^m}{\sqrt{m}}\right)^{-1}$ *and* $\tau(N) = N^{-s/d-(\nu^{-1}+d^{-1})(d/p-s)_+}(\log N)^{-1}$.

Proof. See Suzuki [2018]. □

C Proof of theorems

C.1 Proof of Lemma 3

Proof. Let $\overline{d}_n(f_0, f) = \|f_0 - f\|_n$. As in Ghosal and Van Der Vaart [2007], page 214, we may use the norm $\overline{d}_n$ instead of the average Hellinger distance d_n and

$$K(P_{f_0,i}, P_{f,i}) = \frac{1}{2\sigma^2}|f_0(x_i) - f(x_i)|^2 = \frac{1}{2}V_{2,0}(P_{\theta_0,i}, P_{\theta,i}) \tag{16}$$

for all $i = 1, 2, \cdots, n$. Then,

$$\overline{B}_n(f_0, \epsilon) = \left\{ f \in \mathcal{F} : \overline{d}_n^2(f_0, f) \leq \sigma^2\epsilon^2 \right\} = A_{\sigma\epsilon_n, 1} \tag{17}$$

and it is enough to show that

$$\frac{\Pi(\mathcal{F} - \mathcal{F}_n)}{\Pi(A_{\sigma\epsilon_n,1})} = o\left(e^{-2n\epsilon_n^2}\right), \tag{18}$$

$$\frac{\Pi(A_{2j\epsilon_n,1} - A_{j\epsilon_n,1})}{\Pi(A_{\sigma\epsilon_n,1})} \leq e^{n\epsilon_n^2 j^2/4} \tag{19}$$

for all sufficiently large j.

$$\frac{\Pi(\mathcal{F} - \mathcal{F}_n)}{\Pi(A_{\sigma\epsilon_n,1})} \leq e^{Cn\sigma^2\epsilon_n^2}\Pi(\mathcal{F} - \mathcal{F}_n) = o\left(e^{-2n\epsilon_n^2}\right) \tag{20}$$

and

$$\frac{\Pi(A_{\epsilon_n,2j} - A_{\epsilon_n,j})}{\Pi(A_{\sigma\epsilon_n,1})} \leq e^{Cn\sigma^2\epsilon_n^2}\Pi(A_{\epsilon_n,2j} - A_{\epsilon_n,j}) \leq e^{n\epsilon_n^2 j^2/4} \tag{21}$$

for all sufficiently large j. □

C.2 Proof of Lemma 4

Proof. Let denote $f \in \Phi(L, W, S, B)$ as

$$f(x) = (W^{(L)}\zeta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(1)}x + b^{(1)}),$$

$$\begin{aligned} \mathcal{A}_k^+(f)(x) &= \zeta \circ (W^{(k-1)}\zeta(\cdot) + b^{(k-1)}) \circ \cdots \circ (W^{(1)}x + b^{(1)}), \\ \mathcal{A}_k^-(f)(x) &= (W^{(L)}\zeta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(k)}x + b^{(k)}) \end{aligned} \tag{22}$$

for $k = 2, \cdots, L$, and $\mathcal{A}_{L+1}^-(f)(x) = \mathcal{A}_1^+(f)(x) = x$. Then,

$$f(x) = \mathcal{A}_{k+1}^-(f) \circ (W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^+(f)(x).$$

Since $f \in \Phi(L, W, S, B)$,

$$\begin{aligned} \left\|\mathcal{A}_k^+(f)(x)\right\|_\infty &\le \max_j \left\|W_{j,:}^{(k-1)}\right\|_1 \left\|\mathcal{A}_{k-1}^+(f)(x)\right\|_\infty + \left\|b^{(k-1)}\right\|_\infty \\ &\le WB\left\|\mathcal{A}_{k-1}^+(f)(x)\right\|_\infty + B \\ &\le (W+1)(B \vee 1)\left\|\mathcal{A}_{k-1}^+(f)(x)\right\|_\infty \\ &\le (W+1)^{k-1}(B \vee 1)^{k-1}, \end{aligned} \tag{23}$$

for all x, where $A_{j,:}$ is the j-th row of the matrix A. Similarly,

$$\left|\mathcal{A}_k^-(f)(x_1) - \mathcal{A}_k^-(f)(x_2)\right| \le (BW)^{L-k+1}\|x_1 - x_2\|_\infty. \tag{24}$$

Fix $\epsilon > 0$ and $\theta \in \Theta(L, W, S, B)$. For any $\theta^* \in \Theta(L, W, S, B)$ which satisfies $\|\theta - \theta^*\|_\infty < \epsilon$,

$$\begin{aligned} &|f_\theta(x) - f_{\theta^*}(x)| \\ &= \left|\sum_{k=1}^{L} \mathcal{A}_{k+1}^-(f_{\theta^*}) \circ (W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^+(f_\theta)(x) - \mathcal{A}_{k+1}^-(f_{\theta^*}) \circ (W^{(k)^*} \cdot + b^{(k)^*}) \circ \mathcal{A}_k^+(f_\theta)(x)\right| \\ &\le \sum_{k=1}^{L} (BW)^{L-k}\left\|(W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^+(f_\theta)(x) - (W^{(k)^*} \cdot + b^{(k)^*}) \circ \mathcal{A}_k^+(f_\theta)(x)\right\|_{L^\infty} \\ &\le \sum_{k=1}^{L} (BW)^{L-k}\epsilon\left[W(B \vee 1)^{k-1}(W+1)^{k-1} + 1\right] \\ &\le \sum_{k=1}^{L} (BW)^{L-k}\epsilon(B \vee 1)^{k-1}(W+1)^k \\ &\le \epsilon L(B \vee 1)^{L-1}(W+1)^L. \end{aligned} \tag{25}$$

Let s be the number of nonzero components of θ, $s \le S$. Consider a subspace $\Theta_\theta(L, W, S, B)$ of the parameter space $\Theta(L, W, S, B)$ consists of parameters which have s nonzero component. Choose $\theta_1, \cdots, \theta_M$ from each grid divided with length $\epsilon^* = \frac{\epsilon}{L(B\vee 1)^{L-1}(W+1)^L}$ over $\Theta_\theta(L, W, S, B)$. Then

$$f_\theta \in \bigcup_{m=1}^{M} \{f : \|f - f_{\theta_m}\|_{L^\infty} < \epsilon\}.$$

Note

$$T = |\Theta(L, W, S, B)| \le \sum_{l=1}^{L} (W+1)W \le (L+1)(W+1)^2 \le (W+1)^L$$

for $W,\ L \ge 3$, and the number of cases of choose s nonzero components are

$$\binom{T}{s} = \frac{T(T-1)\cdots(T-s+1)}{s!} \le T^s \le (W+1)^{Ls}.$$

Thus,

$$\begin{aligned} N(\epsilon, \Phi(L,W,S,B), \|\cdot\|_{L^\infty}) &\le \sum_{s^*\le S} \binom{T}{s^*} (2B\epsilon^{-1}(L(B\vee 1)^{L-1}(W+1)^L)^{s^*} \\ &\le \sum_{s^*\le S} (2\epsilon^{-1}L(B\vee 1)^L(W+1)^{2L})^{s^*} \\ &\le (2\epsilon^{-1}L(B\vee 1)^L(W+1)^{2L})^{S+1} \end{aligned} \tag{26}$$

and we get desired results by taking the logarithm of both sides. □

C.3 Proof of Lemma 5

Proof. Let $\tilde{\theta}(a) = (\theta_i I(|\theta_i| > a))$. Then $\left\|\theta - \tilde{\theta}(a)\right\|_\infty \le a$ and

$$\tilde{\Theta}(L,W,S,B,a) = \left\{\tilde{\theta}(a) : \theta \in \Theta(L,W,S,B,a)\right\} = \Theta(L,W,S,B).$$

As in the proof of Lemma 4,

$$\left\|f_\theta - f_{\tilde{\theta}(a)}\right\|_{L^\infty} \le aL(B\vee 1)^{L-1}(W+1)^L \le \epsilon/2 \tag{27}$$

for any $f_\theta,\ f_{\tilde{\theta}(a)} \in \Phi(L,W,S,B,a,F)$. Let $s \le S$ be the number of nonzero components of $\tilde{\theta}(a)$. Consider a subspace $\tilde{\Theta}_\theta(L,W,S,B,a)$ of $\tilde{\Theta}(L,W,S,B,a)$ consisting of parameters which have s nonzero component. Choose $\theta_1, \cdots, \theta_M$ from each grid divided with length $\epsilon^* = \frac{\epsilon/2}{L(B\vee 1)^{L-1}(W+1)^L}$ over $\tilde{\Theta}_\theta(L,W,S,B,a)$. Then

$$f_\theta \in \bigcup_{m=1}^{M} \{f : \|f - f_{\theta_m}\|_{L^\infty} < \epsilon\}$$

from triangular inequality. We get desired results in a similar way as in the proof of Lemma 4. □

C.4 Proof of Theorem 1

Proof. Let $\mathcal{F} = \Phi \cap \mathcal{UB}(F)$ and

$$A_{\epsilon,M} := \{f \in \mathcal{F} : \|f - f_0\|_n \le M\epsilon\}.$$

From Lemma 3, it is enough to show that there exist a constant $C > 2/\sigma^2$ and $\mathcal{F}_n \subset \mathcal{F}$ which satisfy

(a) $\sup_{\epsilon>\epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \le n\epsilon_n^2$

(b) $-\log \Pi(A_{\epsilon_n,1}) \le Cn\epsilon_n^2$

(c) $\Pi(\mathcal{F} - \mathcal{F}_n) = o\left(e^{-(C\sigma^2+2)n\epsilon_n^2}\right)$

for sufficiently large n. Let $\mathcal{F}_n = \Phi(L_n, W_n, S_n, B_n) \cap \mathcal{UB}(F)$. First, (c) is trivial from (16). From

$$\{f \in \mathcal{F}_n : \|f\|_{L^\infty} \le \epsilon\} \subset \{f \in \mathcal{F}_n : \|f\|_n \le \epsilon\}$$

and Lemma 4,

$$\begin{aligned} &\sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n\right) \\ &\le \sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\ &\le \sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\ &\le \log N\left(\frac{\epsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\ &\le (S_n+1)\left[\log L_n + L_n \log\big((B_n\vee 1)(W_n+1)^2\big) - \log\frac{\epsilon_n}{72}\right] \\ &\lesssim N_n(\log n)^3 \\ &= n\epsilon_n^2 \end{aligned} \tag{28}$$

for sufficiently large n. Thus, (a) holds. Here, the last inequality holds from

$$L_n = O(\log n),\ W_n = O(N_n),\ S_n = O(N_n \log n) \tag{29}$$

Next, from Lemma 6, there is a constant $C_1 > 0$ and $\hat{f}_n = f_{\hat{\theta}} \in \mathcal{F}$ such that

$$\left\| \hat{f}_n - f_0 \right\|_{L^2} \leq C_1 \|f_0\|_{B^s_{p,q}([0,1]^d)} N_n^{-s/d} \leq \epsilon_n/4. \tag{30}$$

for sufficiently large n. Moreover, by the assumptions and the strong law of large numbers,

$$\|f - f_0\|_n^2 \leq 2\|f - f_0\|^2_{L^2(P_X)} \leq 4\|f - f_0\|^2_{L^2} \tag{31}$$

for sufficiently large n $P_{f_0}^{(n)}$-almost surely. Let $\hat{\gamma}$ and $\hat{\theta}_{\hat{\gamma}}$ be index and value of nonzero components of $\hat{\theta}$ respectively. Let $\Theta(\hat{\gamma}; L_n, W_n, S_n, B_n)$ be a subset of parameter space $\Theta(L_n, W_n, S_n, B_n)$ consists of parameters which have nonzero components at $\hat{\gamma}$ only and $\mathcal{F}_n(\hat{\gamma}) = \Phi(\hat{\gamma}; L_n, W_n, S_n, B_n) \cap \mathcal{UB}(F)$ be an uniformly bounded neural network space generated by $\Theta(\hat{\gamma}; L_n, W_n, S_n, B_n)$. Note

$$\begin{aligned}
\Pi(A_{\epsilon_n,1}) &= \Pi(f \in \mathcal{F}_n : \|f - f_0\|_n \leq \epsilon_n) \\
&\geq \Pi(f \in \mathcal{F}_n : \|f - f_0\|_{L^2} \leq \epsilon_n/2) \\
&\geq \Pi\left(f \in \mathcal{F}_n : \left\|f - \hat{f}_n\right\|_{L^2} \leq \epsilon_n/4\right) \\
&\geq \Pi\left(f \in \mathcal{F}_n : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/4\right) \\
&\geq \Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/4\right)
\end{aligned} \tag{32}$$

for sufficiently large n. As in the proof of Lemma 4,

$$\begin{aligned}
&\Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/4\right) \\
&\geq \Pi\left(\theta \in \mathbb{R}^{T_n} : \theta_{\hat{\gamma}^c} = 0,\ \|\theta\|_\infty \leq B_n,\ \left\|\hat{\theta} - \theta\right\|_\infty \leq \frac{\epsilon_n}{4(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n - 1}}\right) \\
&\geq \left(\frac{\epsilon_n}{4B_n(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n-1}}\right)^{S_n} \binom{T_n}{S_n}^{-1} \\
&\geq \left(\frac{\epsilon_n}{4B_n(W_n+1)^{2L_n} L_n (B_n \vee 1)^{L_n-1}}\right)^{S_n} \\
&= \exp\left(-S_n \log\left(\frac{4B_n(W_n+1)^{2L_n} L_n (B_n \vee 1)^{L_n-1}}{\epsilon_n}\right)\right)
\end{aligned} \tag{33}$$

Thus,

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n,1}) &\leq S_n \log\left(\frac{4B_n(W_n+1)^{2L_n} L_n (B_n \vee 1)^{L_n-1}}{\epsilon_n}\right) \\
&\leq S_n \left[L_n \log\big((W_n+1)^2 (B_n \vee 1)\big) + \log 2L_n - \log \epsilon_n\right] \\
&\lesssim N_n (\log n)^3 \\
&= n\epsilon_n^2
\end{aligned} \tag{34}$$

for sufficiently large n. □

C.5 Proof of Theorem 2

Proof. Let $\mathcal{F} = \Phi \cap \mathcal{UB}(F)$. From Lemma 3, it is enough to show that there exist a constant $C' > 2/\sigma^2$ and $\mathcal{F}_n \subset \mathcal{F}$ which satisfy

(a) $\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \leq n\epsilon_n^2$

(b) $-\log \Pi(A_{\epsilon_n,1}) \leq C' n \epsilon_n^2$

(c) $\Pi(\mathcal{F} - \mathcal{F}_n) = o\left(e^{-(C'\sigma^2+2)n\epsilon_n^2}\right)$

for sufficiently large n. From (29), we can choose

$$H_0 > \sup_n\{L_n/\log n,\ W_n/N_n,\ S_n/(N_n \log n),\ \Xi\}. \tag{35}$$

Let $\tilde{N}_n = C_N N_n$,

$$\mathcal{F}_n = \mathcal{UB}(F) \cap \left(\bigcup_{N=1}^{\tilde{N}_n} \Phi(\tilde{L}_n(H_0),\ \tilde{W}_n(H_0,N),\ \tilde{S}_n(H_0,N),\ \tilde{B}_n(H_0,N)) \right)$$

for sufficiently large $C_N > 0$ and $\pi_N(N)$ be a density function of N. First, show that (a) holds. From Lemma 4,

$$\begin{aligned}
&N\left(\frac{\epsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\le \sum_{N=1}^{\tilde{N}_n} \left(\frac{72}{\epsilon_n} \tilde{L}_n(H_0) (\tilde{B}_n(H_0,N) \vee 1)^{\tilde{L}_n(H_0)} (\tilde{W}_n(H_0,N)+1)^{2\tilde{L}_n(H_0)} \right)^{\tilde{S}_n(H,N)+1} \\
&\le \tilde{N}_n \left(\frac{72}{\epsilon_n} \tilde{L}_n(H_0) (\tilde{B}_n(H_0,\tilde{N}_n) \vee 1)^{\tilde{L}_n(H_0)} (\tilde{W}_n(H_0,\tilde{N}_n)+1)^{2\tilde{L}_n(H_0)} \right)^{\tilde{S}_n(H_0,\tilde{N}_n)+1}
\end{aligned} \tag{36}$$

and

$$\begin{aligned}
&\sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n\right) \\
&\le \sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\le \sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\le \log N\left(\frac{\epsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\le \log \tilde{N}_n \\
&\quad + \left[\tilde{S}_n(H_0,\tilde{N}_n)+1\right] \log\left(\frac{72}{\epsilon_n} \tilde{L}_n(H_0) (\tilde{B}_n(H_0,\tilde{N}_n) \vee 1)^{\tilde{L}_n(H_0)} (\tilde{W}_n(H_0,\tilde{N}_n)+1)^{2\tilde{L}_n(H_0)} \right) \\
&\lesssim \tilde{N}_n (\log n)^3 \\
&\lesssim n\epsilon_n^2.
\end{aligned} \tag{37}$$

for sufficiently large n. Next, show that (b) holds. Note $N_n(\log n)^3 \lesssim n\epsilon_n^2$ and

$$L_n \le \tilde{L}_n(H_n),\ W_n \le \tilde{W}_n(H_n,N_n),\ S_n \le \tilde{S}_n(H_n,N_n),\ B_n \le \tilde{B}_n(H_n,\tilde{N}_n) \tag{38}$$

for $N_n,\ L_n,\ W_n,\ S_n,\ B_n$ in Theorem 1 and sufficiently large n. Thus, there exists a constant $D > 0$ such that

$$\pi_N(N_n) \gtrsim \exp\left(-N_n (\log n)^2 \log \frac{N_n}{\lambda_N} \right) \gtrsim \exp\left(-Dn\epsilon_n^2\right) \tag{39}$$

and

$$\begin{aligned}
&\Pi(f_\theta \in \mathcal{F}_n : \|f - f_0\|_n \le \epsilon_n) \\
&\ge \pi_N(N_n) \Pi(f_\theta \in \Phi(L_n,W_n,S_n,B_n) : \|f_\theta - f_0\|_n \le \epsilon_n | N_n) \\
&\gtrsim \exp\left(-(C+D)n\epsilon_n^2\right)
\end{aligned} \tag{40}$$

holds for sufficiently large n. (b) holds for $C' = \max\left\{C+D, 1+2/\sigma^2\right\}$. From

$$\Pi(\mathcal{F} - \mathcal{F}_n) \le \pi_N(N > \tilde{N}_n)$$

and *Chernoff bound*, for any positive number $t,\ Z_0 > 0$,

$$P(Z > Z_0) < e^{-t(Z_0+1)} \mathbb{E}\left[e^{tZ}\right] \lesssim e^{-t(Z_0+1)} \left(\exp\left(e^t \lambda_N\right) - 1\right). \tag{41}$$

Letting $t = \log Z_0$, we get

$$P(Z > Z_0) \lesssim e^{-(Z_0+1)\log Z_0} \left(\exp(Z_0 \lambda_N) - 1\right). \tag{42}$$

Thus,

$$\begin{aligned} \pi_N(N > \tilde{N}_n) &\lesssim e^{-\left[(\tilde{N}_n+1)\log \tilde{N}_n + \tilde{N}_n \lambda_N\right](\log n)^2}, \\ (C'\sigma^2+2)n\epsilon_n^2 + \lambda_N \tilde{N}_n (\log n)^2 - (\tilde{N}_n+1)\log \tilde{N}_n (\log n)^2 &\to -\infty \end{aligned} \tag{43}$$

for sufficiently large $C_N > 0$. (c) holds. □

C.6 Proof of Theorem 3

Proof. Let $\mathcal{F} = \Phi \cap \mathcal{UB}(F)$. From Lemma 3, it is enough to show that there exist a constant $C'' > 2/\sigma^2$ and $\mathcal{F}_n \subset \mathcal{F}$ which satisfy

(a) $\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \le n\epsilon_n^2$

(b) $-\log \Pi(A_{\epsilon_n,1}) \le C'' n\epsilon_n^2$

(c) $\Pi(\mathcal{F} - \mathcal{F}_n) = o\left(e^{-(C''\sigma^2+2)n\epsilon_n^2}\right)$

for sufficiently large n. Let $\mathcal{F}_n = \Phi(L_n, W_n, S_n, B_n, a_n) \cap \mathcal{UB}(F)$. It is easy to show that (a) holds from Lemma 5 as in the proof of Theorem 1. By the assumption,

$$\begin{aligned} &\Pi(\mathcal{F} - \mathcal{F}_n) \\ &\le \pi\left(\exists |\theta_i| > B_n \middle| L_n, W_n, S_n, B_n\right) + \pi\left(\sum_{i=1}^{T_n} I(|\theta_i| > a_n) > S_n | L_n, W_n, S_n, B_n\right) \\ &= \left(1 - (1-v_n)^{T_n}\right) + P\left(S > S_n | S \sim B(T_n, 1-u_n)\right) \\ &\le T_n v_n + \exp\left(-T_n \left\{\left(1 - \frac{S_n}{T_n}\right)\log\frac{1 - S_n/T_n}{u_n} + \frac{S_n}{T_n}\log\frac{S_n/T_n}{1-u_n}\right\}\right) \\ &= o\left(e^{-K_0 n\epsilon_n^2 + \log T_n}\right) + \exp\left(T_n\left(1 - \frac{S_n}{T_n}\right)\log\frac{u_n}{1 - S_n/T_n} - S_n \log\frac{S_n/T_n}{1-u_n}\right) \\ &\le o\left(e^{-K_1 n\epsilon_n^2}\right) + \exp\left(T_n\left(1 - \frac{S_n}{T_n}\right)\log\frac{1 - \eta_n S_n/T_n}{1 - S_n/T_n} - S_n \log\frac{S_n/T_n}{1-u_n}\right) \\ &\le o\left(e^{-K_1 n\epsilon_n^2}\right) + \exp\left(T_n\left(1 - \frac{S_n}{T_n}\right)\left(\frac{(1-\eta_n)S_n/T_n}{1 - S_n/T_n} + o\left(\frac{(1-\eta_n)S_n/T_n}{1 - S_n/T_n}\right)\right) - Kn\epsilon_n^2\right) \\ &= o\left(e^{-K_1 n\epsilon_n^2}\right) + \exp\left(S_n(1-\eta_n) + o(S_n(1-\eta_n)) - Kn\epsilon_n^2\right) \\ &= o\left(e^{-K_1 n\epsilon_n^2}\right) + o\left(e^{-K_2 n\epsilon_n^2}\right) \\ &= o\left(e^{-\min\{K_1,K_2\} n\epsilon_n^2}\right). \end{aligned} \tag{44}$$

for some $4 < K_1 < K$ and $4 < K_2 < K_0$. Letting $C'' = (\min\{K_1, K_2\} - 2)/\sigma^2$, (c) holds. We use Bernoulli's inequality and a tail bound for binomial distribution in Arratia and Gordon [1989] for the second inequality. Next, as in the proof of Theorem 1, there is a constant $C_1 > 0$ and $\hat{f}_n = f_{\hat{\theta}} \in \mathcal{F}_n$ such that

$$\begin{aligned} \left\|\hat{f}_n - f_0\right\|_{L^2} \le C_1 \|f_0\|_{B^s_{p,q}([0,1]^d)} N_n^{-s/d} \le \epsilon_n/4, \\ \|f - f_0\|_n^2 \le 4\|f - f_0\|_{L^2}^2 \end{aligned} \tag{45}$$

for sufficiently large n almost surely. Let $\hat{\gamma}$ and $\hat{\theta}_{\hat{\gamma}}$ be index and value of nonzero components of $\hat{\theta}$ respectively. Let

$$\tilde{\Theta}(L, W, S, B, a) = \left\{\tilde{\theta} : \theta \in \Theta(L, W, S, B, a)\right\} = \Theta(L, W, S, B)$$

and $\tilde{\Theta}(\hat{\gamma}; L_n, W_n, S_n, B_n, a_n)$ be a subset of parameter space $\tilde{\Theta}(L_n, W_n, S_n, B_n, a_n)$ consists of parameters which have nonzero components at $\hat{\gamma}$ only and $\mathcal{F}_n(\hat{\gamma}) = \tilde{\Phi}(\hat{\gamma}; L_n, W_n, S_n, B_n, a_n) \cap$

$\mathcal{UB}(F)$ be an uniformly bounded neural network space generated by $\tilde{\Theta}(\hat{\gamma}; L_n, W_n, S_n, B_n, a_n)$. Note

$$\begin{aligned}
\Pi(A_{\epsilon_n,1}) &= \Pi(f \in \mathcal{F} : \|f - f_0\|_n \leq \epsilon_n) \\
&\geq \Pi\left(f \in \mathcal{F} : \left\|f - \hat{f}_n\right\|_{L^2} \leq \epsilon_n/4\right) \\
&\geq \Pi\left(f \in \mathcal{F} : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/4\right) \\
&\geq \Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/4\right)
\end{aligned} \tag{46}$$

for sufficiently large n. As in the proof of Lemma 4,

$$\begin{aligned}
&\Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/4\right) \\
&\geq \Pi\Bigg(\theta \in \mathbb{R}^{T_n} : \theta_{\hat{\gamma}^c} \in [-a_n, a_n]^{T_n - S_n},\ \|\theta_{\hat{\gamma}}\|_\infty \leq B_n, \\
&\qquad \left\|\hat{\theta}_{\hat{\gamma}} - \theta_{\hat{\gamma}}\right\|_\infty \leq \frac{\epsilon_n}{4(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n - 1}}\Bigg) \\
&\geq u_n^{T_n - S_n} \left(\int_{B_n - t_n}^{B_n} g(t)dt\right)^{S_n},
\end{aligned} \tag{47}$$

where $t_n = \frac{\epsilon_n}{4(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n - 1}}$. Letting

$$y_n = \int_{B_n - t_n}^{B_n} g(t)dt \geq t_n g(B_n),$$

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n,1}) &\leq -S_n \log y_n - (T_n - S_n) \log u_n \\
&\leq -S_n \log(t_n g(B_n)) - T_n \left(1 - \frac{S_n}{T_n}\right) \log(1 - S_n/T_n) \\
&= -S_n \log t_n - S_n \log g(B_n) + T_n \left(1 - \frac{S_n}{T_n}\right) (S_n/T_n + o(S_n/T_n)) \\
&\lesssim S_n (\log n)^2 + S_n + o(S_n) \\
&\lesssim n\epsilon_n^2.
\end{aligned} \tag{48}$$

□

D Numerical Examples

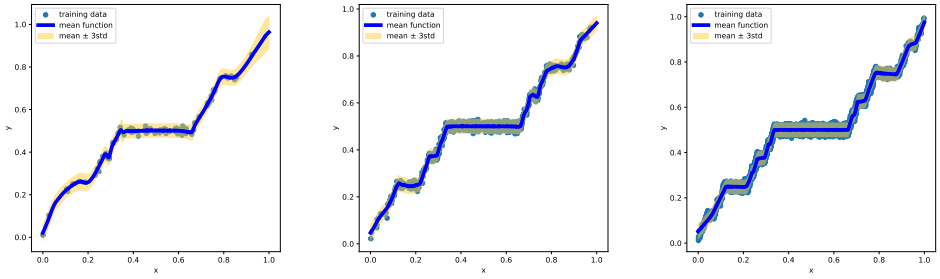
D.1 Gaussian Mixture Prior

Consider a problem of estimating the Besov functions $f_1,\ f_2,\ f_3$ and f_4. Assume the Gaussian mixture prior distribution for each parameter as in Example 2. We sampled n points from $x \sim U(0,1)$ and y from $\mathcal{N}(f_i(x),\ \sigma_i^2)$ for $i = 1, 2, 3$ and 4. We set $\sigma_1^2 = \sigma_2^2 = 0.01^2$ and $\sigma_3^2 = \sigma_4^2 = 0.1^2$.

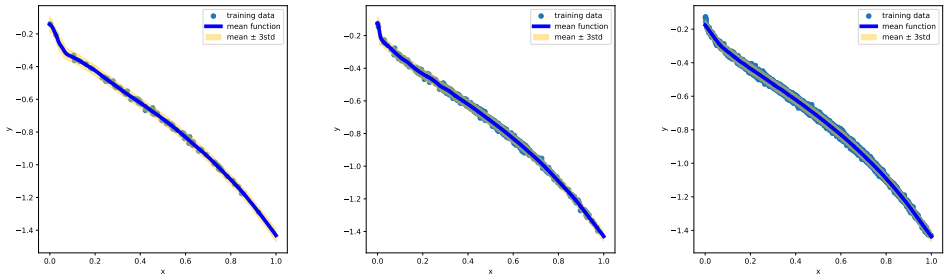
As mentioned in Section 4, we used a smaller model that was weaker than the conditions of the theorem. We fix the depth $L_n = 5$, the width $W_n = 200$ and set $N_n \leftarrow W_n/W_0$ of the model. We consider the prior in Example 2 (Gaussian mixture BNN) and set the smaller variance $\sigma_{1n} \leftarrow \max\{0.001,\ \sigma_{1n}\}$ to avoid numerical precision issues.

We fit the model using the NUTS algorithm [Hoffman et al., 2014] with the python `Pyro` [Bingham et al., 2019] and `PyTorch` [Paszke et al., 2019] packages. Experiments were run on a GPU server with Nvidia GeForce GTX TITAN X and RTX 3090. The code and instructions for the experiment are provided in the supplementary material. Figure 1 shows the results. Overall, as the number of the data n increase, the mean functions get closer to the true regression function.

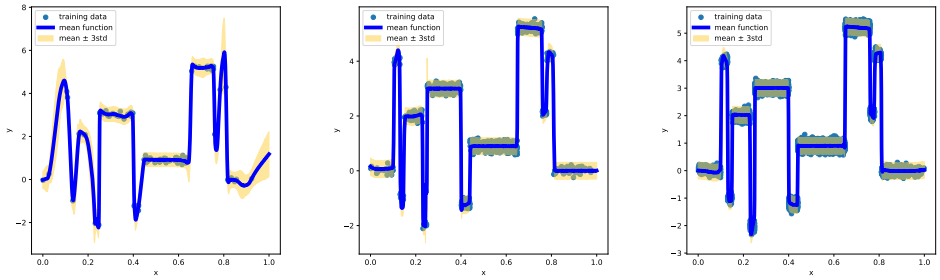
Figure 1: The results of estimating four functions $f_1,\ f_2,\ f_3$ and f_4 using Gaussian mixture BNN model with $n = 100$ (left), $n = 1,000$ (center) and $n = 10,000$ (right) samples. We construct 1,000 functions from the MCMC samples and plot the mean function with training data. The blue lines are the mean functions and the yellow intervals are the prediction intervals.



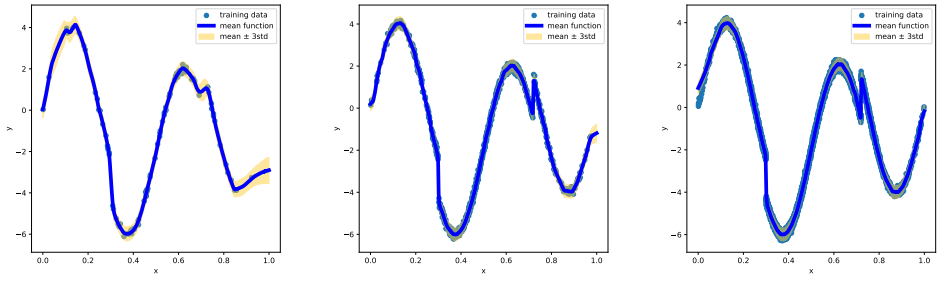
(a) The results of estimating f_1.



(b) The results of estimating f_2.



(c) The results of estimating f_3.



(d) The results of estimating f_4.

D.2 Gaussian Prior

We performed the same experiment as before using Gaussian prior (Gaussian BNN). The standard deviation of the Gaussian prior is the same as σ_{2n} in the Gaussian mixture BNN (see Example 1). Figure 2 shows the results. Overall, as the number of the data n increase, the mean functions get closer to the true regression function.

D.3 Comparison

We compare two model in terms of empirical errors (training error) $\frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|^2$ and L^2 norms (test error) $e_f = \|f - f_0\|_{L^2}$ between true function f_0 and sampled function f. We approximate e_f as $\hat{e}_f$ using Riemann sum,

$$\hat{e}_f = \left(\frac{1}{n^*}\sum_{k=1}^{n^*}(f(x_k^*) - f_0(x_k^*))^2\right)^{1/2}, \quad x_k^* = \frac{k}{n^*},\ 1 \le k \le n^*.$$

As in Table 1 and Table 2, both models give similar results. For the functions f_1 and f_2, Gaussian mixture BNN shows better results for both errors. For the functions f_3 and f_4, Gaussian BNN gives slightly better results for the test error. However, as shown in Figure 1 and Figure 2, the differences are negligible.

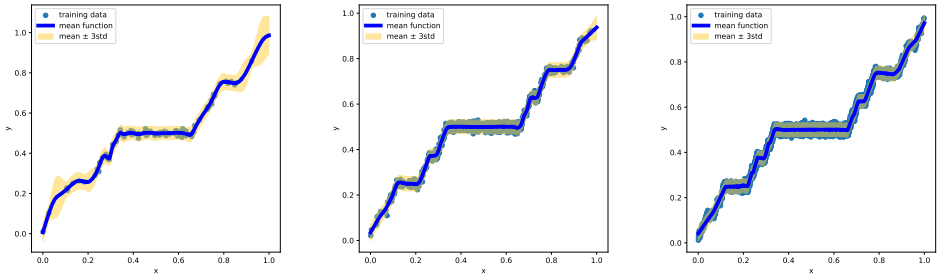
Table 1: Summary statistics of the empirical errors between true functions and sampled functions, where the numbers in the parentheses denote the standard deviations.

f_0	n	Gaussian mixture BNN	Gaussian BNN
f_1	100	1.398×10^{-2} (8.896×10^{-4})	1.386×10^{-2} (8.774×10^{-4})
	1,000	1.451×10^{-2} (2.899×10^{-4})	1.455×10^{-2} (2.840×10^{-4})
	10,000	1.455×10^{-2} (9.233×10^{-5})	1.461×10^{-2} (8.763×10^{-5})
f_2	100	1.336×10^{-2} (8.119×10^{-4})	1.359×10^{-2} (9.023×10^{-4})
	1,000	1.424×10^{-2} (2.781×10^{-4})	1.427×10^{-2} (2.741×10^{-4})
	10,000	1.420×10^{-2} (8.728×10^{-5})	1.419×10^{-2} (8.651×10^{-5})
f_3	100	1.428×10^{-1} (9.301×10^{-3})	1.497×10^{-1} (9.356×10^{-3})
	1,000	2.297×10^{-1} (4.256×10^{-3})	2.289×10^{-1} (4.632×10^{-3})
	10,000	2.200×10^{-1} (3.228×10^{-3})	2.204×10^{-1} (2.668×10^{-3})
f_4	100	1.389×10^{-1} (9.279×10^{-3})	1.390×10^{-1} (9.085×10^{-3})
	1,000	1.430×10^{-1} (2.898×10^{-3})	1.460×10^{-1} (2.921×10^{-3})
	10,000	1.543×10^{-1} (8.986×10^{-4})	1.500×10^{-1} (8.825×10^{-4})

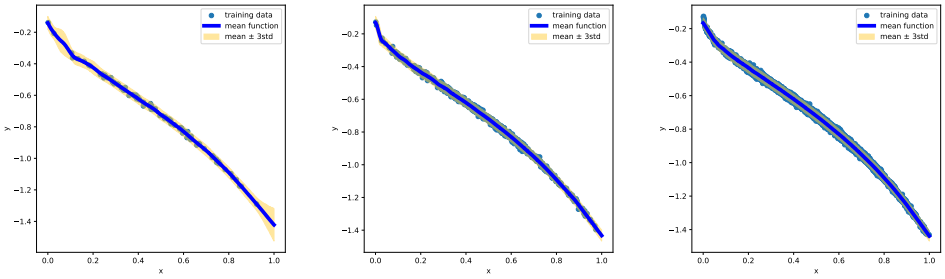
Table 2: Summary statistics of the test errors between true functions and sampled functions, where the numbers in the parentheses denotes the standard deviations.

f_0	n	Gaussian mixture BNN	Gaussian BNN
f_1	100	1.921×10^{-2} (1.441×10^{-3})	2.359×10^{-2} (3.646×10^{-3})
	1,000	1.327×10^{-2} (3.172×10^{-4})	1.335×10^{-2} (5.983×10^{-4})
	10,000	1.212×10^{-2} (2.808×10^{-4})	1.201×10^{-2} (2.643×10^{-4})
f_2	100	1.363×10^{-2} (6.582×10^{-4})	1.652×10^{-2} (2.293×10^{-3})
	1,000	1.151×10^{-2} (3.835×10^{-4})	1.164×10^{-2} (4.203×10^{-4})
	10,000	1.177×10^{-2} (2.626×10^{-4})	1.155×10^{-2} (2.529×10^{-4})
f_3	100	1.120×10^{0} (5.175×10^{-2})	1.041×10^{0} (6.923×10^{-2})
	1,000	3.750×10^{-1} (5.497×10^{-3})	3.699×10^{-1} (6.699×10^{-3})
	10,000	2.478×10^{-1} (4.589×10^{-3})	2.479×10^{-1} (4.036×10^{-3})
f_4	100	4.942×10^{-1} (2.930×10^{-2})	4.493×10^{-1} (4.450×10^{-2})
	1,000	1.715×10^{-1} (8.959×10^{-3})	1.666×10^{-1} (1.029×10^{-2})
	10,000	1.550×10^{-1} (3.242×10^{-3})	1.128×10^{-1} (2.454×10^{-3})

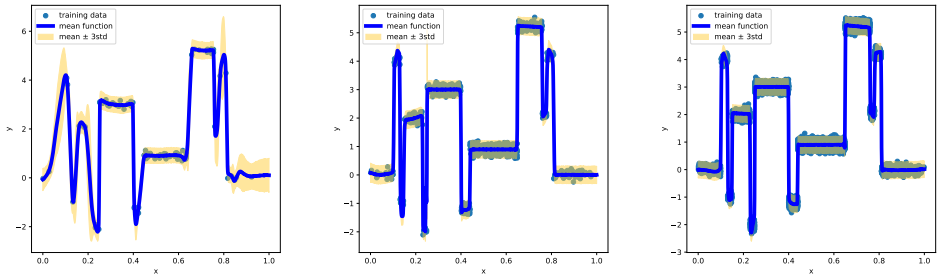
Figure 2: The results of estimating four functions $f_1,\ f_2,\ f_3$ and f_4 using Gaussian BNN models with $n = 100$ (left), $n = 1,000$ (center) and $n = 10,000$ (right) samples. We construct 1,000 functions from the MCMC samples and plot the mean function with training data. The blue lines are the mean functions and the yellow intervals are the prediction intervals.



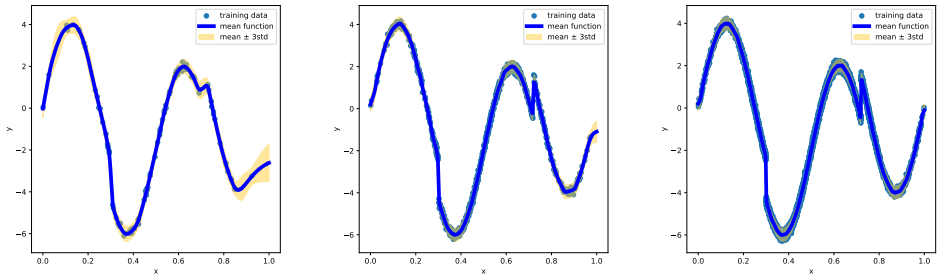
(a) The results of estimating f_1.



(b) The results of estimating f_2.



(c) The results of estimating f_3.



(d) The results of estimating f_4.

References

R. Arratia and L. Gordon. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology*, 51(1):125–131, 1989.

E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019. URL `http://jmlr.org/papers/v20/18-403.html`.

S. Ghosal and A. Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

M. D. Hoffman, A. Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

N. G. Polson and V. Ročková. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

T. Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.