Introduction

(Lee and Lee, 2022) proved asymptotic properties for Bayesian Neural Network in Besov space. Theorem 1 proves asymptotic properties of Bayesian ReLU networks with spike-and-slab prior. Since computational cost of the prior is unsuitable, shrinkage prior is introduced and the same property with the prior is proven in Theorem 3. This report includes details of the model setting, statement and proof of Theorem 3.

Main contents

Model

Suppose that n input-output observations $\mathbb{D}_n = (X_i,y_i)_{i=1}^n \subset [0,1]^d \times \mathbb{R}$ are independent random sample from a regression model

$$y_i = f_0(X_i) + \xi_i \ (i = 1,2,\cdots,n), \tag{1}$$

where $(\xi_i)_{i=1}^n$ is an i.i.d sequence of Gaussian noises $\mathcal{N}(0,\sigma^2)$ with known variance $\sigma^2 > 0$ and f_0 is the true regression function belonging to the space $\mathcal{F}$. We consider neural network space $\Phi(L,W,S,B) = \Phi(\Theta(L,W,S,B))$ generated by a parameter space $\Theta(L,W,S,B)$ defined in (Lee and Lee, 2022), and a prior defined as following conditions. The prior will be specified to shrinkage prior later.

$$\pi(L = L_n) = \pi(W = W_n) = \pi(S = S_n) = \pi(B = B_n) = 1, \tag{2}$$

$$\pi(\theta|L,W,S,B) = \prod_{j=1}^{T} g(\theta_j|L,W,S,B) \tag{3}$$

where $T = |\Theta(L,W,S,B)|$, and $g(t) := g(t|L,W,S,B)$ is a symmetric density function decreasing on $t > 0$.
For function spaces, we consider Besov space $B_{p,q}^s(\Omega) = \{f : \|f\|_{B_{p,q}^s} < \infty\}$ defined in (Lee and Lee, 2022). Also, suppose that $0 < F < \infty$, $0 < p,q \leq \infty$, $w := d(1/p - 1/2)_+ < s < \infty$ and set $\nu = (s-w)/(2w)$. Assume that $m \in \mathbb{N}$ satisfies $0 < s < \min\{m,m-1+1/p\}$. Let $N_n = \lceil n^{d/(2s+d)} \rceil$, $W_0 = 6dm(m+2) + 2d$ and

$$\begin{aligned} L_n = L(N_n), W_n = N_n W_0, \\ S_n = (L_n - 1)W_0^2 N_n + N_n, B_n = O(N_n^{\Xi}) \end{aligned} \tag{4}$$

where $c_{(d,m)} = (1 + 2de\frac{(2e)^m}{\sqrt{m}})^{-1}$, $L(N_n) = 3 + 2\lceil \log_2(\frac{3^{d\vee m}}{\tau_n c_{(d,m)}}) + 5 \rceil \lceil \log_2(d \vee m) \rceil$, $\tau_n = N_n^{-s/d-(v^{-1}+d^{-1})(d/p-s)_+}(\log N_n)^{-1}$ and $\Xi = (\nu^{-1} + d^{-1})(1 \vee (d/p - s)_+)$. Then we can show the following equations.

$$L_n = O(\log n), W_n = O(N_n), S_n = O(N_n \log n) \tag{5}$$

Lemmas

Lemma 3, 5, 6 from (Lee and Lee, 2022), and the rest of lemmas are directly used in the proof of theorem 3.

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

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

, where $\|\cdot\|_n$ denotes the empirical L^2 norm,

$$\|f\|_n = (\frac{1}{n}\sum_{i=1}^{n}(f(X_i))^2)^{1/2}$$

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

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \leq n\epsilon_n^2, \tag{6}$$

$$\Pi(A_{\epsilon_n,1}) \geq e^{-Cn\epsilon_n^2}, \tag{7}$$

$$\Pi(\mathcal{F} - \mathcal{F}_n) = o(e^{-(C\sigma^2+2)n\epsilon_n^2}) \tag{8}$$

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

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

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

Lemma 5 *For $L, W, S \in \mathbb{N}$ and $B, a > 0$, define a function space*

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

where

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

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

$$\log N(\epsilon, \Phi(L,W,S,B,a), ||\cdot||_{L^\infty} \leq (S+1)\log(2\epsilon^{-1}L(B \vee 1)^L(W+1)^{2L}) \tag{9}$$

Lemma 6 *Suppose that $0 < p,q,r \leq \infty, w := d(1/p - 1/r)_+ < s < \infty$ and $\nu = (s-w)/(2w)$. 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,*

$$\sup_{f_0 \in U(B^s_{p,q}([0,1]^d))} \inf_{f \in \Pi(L,W,S,B)} \|f_0 - f\|_{L^r} \lesssim N^{-s/d} \tag{10}$$

for

$$L = 3 + 2\left\lceil \log_2\left(\frac{3^{d\vee m}}{\tau(N)c_{(d,m)}}\right) + 5\right\rceil \lceil \log_2(d \vee m)\rceil, W = NW_0, \tag{11}$$

$$S = (L-1)W_0^2N + N, B = O(N^{(v^{-1}+d^{-1})(1\vee(d/p-s)_+)}), \tag{12}$$

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 $\tau(N) = N^{-s/d-(v^{-1}+d^{-1})(d/p-s)_+}(\log N)^{-1}$.

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

$$|f_\theta(x) - f_{\theta^*}(x)| \leq \epsilon L(B \vee 1)^{L-1}(W+1)^L \tag{13}$$

This lemma is proved from the proof of lemma 4 in (Lee and Lee, 2022).

Lemma *(tailbound of binomial distribution, Arratia and Gordon, 1989) Let $S_n \sim B(n,p)$, and $H(a,p)$ be the relative entropy between p, a, i.e.*

$$H = H(a,p) = (a)\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p} \tag{14}$$

. For $0 \leq p < a < 1$, and for $n \in \mathbb{N}$, with $H = H(a,p)$,

$$P(S_n \geq an) \leq e^{-nH} \tag{15}$$

Statement

Assume model (1), prior distribution (2) and (3). Suppose that $0 < F < \infty$, $0 < p, q \leq \infty$ and $d(1/p - 1/2)_+ < s < \min\{m, m-1+1/p\}$. Let $\epsilon_n = n^{-s/(2s+d)}(\log n)^{3/2}$ and $g(t)$ be a function such that

$$\begin{aligned} a_n &\leq \frac{\epsilon_n}{72L_n(B_n \vee 1)^{L_n-1}(W_n+1)^{L_n}} \\ u_n &= \int_{[-a_n,a_n]} g(t|L_n,W_n,S_n,B_n)dt \end{aligned} \tag{16}$$

satisfies

$$\frac{S_n}{T_n} > 1 - u_n \geq \frac{S_n}{T_n}\eta_n, \tag{17}$$

$$-\log g(B_n|L_n,W_n,S_n,B_n) \lesssim (\log n)^2, \tag{18}$$

continuous on $[-B_n, B_n]$ and

$$v_n = \int_{[-B_n,B_n]^c} g(t|L_n,W_n,S_n,B_n)dt = o\left(e^{-K_0 n\epsilon_n^2}\right), \tag{19}$$

where $\eta_n = \exp(-Kn\epsilon_n^2/S_n)$ for some K, $K_0 > 4$. The posterior distribution concentrates at the true function with a rate ϵ_n. That is,

$$\Pi(f_\theta \in \Phi \cap \mathcal{UB}(F) : ||f_\theta - f_0||_n > M_n\epsilon_n|\mathbb{D}_n) \to 0$$

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

Proof

We mainly use lemma 3 to show contraction rate of the posterior distribution. Thus we let $\mathcal{F} = \Phi \cap \mathcal{UB}(F)$, and it is enough to show that there exists a constant $C'' > 2/\sigma^2$ and $\mathcal{F}_n \in \mathcal{F}$ which satisfies

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \leq n\epsilon_n^2 \tag{20}$$

$$-\log \Pi(A_{\epsilon_n,1}) \leq C'' n\epsilon_n^2 \tag{21}$$

$$\Pi(\mathcal{F} - \mathcal{F}_n) = o(e^{-(C''+\sigma^2+2)n\epsilon_n^2}) \tag{22}$$

for sufficiently large n. Let $\mathcal{F}_n = \Phi(L_n,W_n,S_n,B_n,a_n) \cap \mathcal{UB}(F)$, Φ defined as in lemma 5. We check the three condition (20), (21), (22) in (a), (b), (c) respectively.

(a)

$$\begin{aligned}
&\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \\
\leq &\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_{L^\infty}) \\
\leq &\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, \mathcal{F}_n, \|\cdot\|_{L^\infty}) \\
\leq &\log N(\epsilon_n/36, \mathcal{F}_n, \|\cdot\|_{L^\infty}) \\
\leq &(S_n+1)\log L_n + L_n \log((B_n \vee 1)(W_n+1)^2) - \log\frac{\epsilon_n}{72} \\
\lesssim &N_n(\log n)^3 \\
= &n\epsilon_n^2
\end{aligned} \tag{23}$$

for sufficiently large n. the fourth inequality satisfies by lemma 5, and the last inequality satisfies from the previous condtion of parameters (5). Thus, (20) satisfies.

(b) By lemma 6, there is a constant $C > 0$ and $\hat{f}_n = f_{\hat{\theta}} \in \mathcal{F}_n$ such that

$$\|\hat{f}_n - f_0\|_{L^2} \leq C\|f_0\|_{B^s_{p,q}([0,1]^d)} N_n^{-s/d} \leq \epsilon_n/4 \tag{24}$$

$$\|f - f_0\|_n^2 \leq 4\|f - f_0\|_{L^2}^2 \tag{25}$$

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

$$\tilde{\Theta}(L,W,S,B,a) = \{\tilde{\theta} : \theta \in \Theta(L,W,S,B,a)\} \tag{26}$$

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 $\tilde{\Theta}(\hat{\gamma}; L_n, W_n, S_n, B_n, a_n)$. Then,

$$\begin{aligned}
&\Pi(A_{\epsilon_n,1} = \Pi(f \in \mathcal{F} : \|f - f_0\|_n \leq \epsilon_n) \\
\geq &\Pi(A_{\epsilon_n,1} = \Pi(f \in \mathcal{F} : \|f - f_0\|_{L^2} \leq \epsilon_n/2) \\
\geq &\Pi(A_{\epsilon_n,1} = \Pi(f \in \mathcal{F} : \|f - \hat{f}_n\|_{L^2} + \|\hat{f}_n - f_0\|_{L^2} \leq \epsilon_n/2) \\
\geq &\Pi(A_{\epsilon_n,1} = \Pi(f \in \mathcal{F} : \|f - \hat{f}_n\|_{L^2} \leq \epsilon_n/4, \|\hat{f}_n - f_0\|_{L^2} \leq \epsilon_n/4) \\
= &\Pi(f \in \mathcal{F} : \|f - \hat{f}_n\|_{L^2} \leq \epsilon_n/4) \\
\geq &\Pi(f \in \mathcal{F} : \|f - \hat{f}_n\|_{L^\infty} \leq \epsilon_n/4) \\
\geq &\Pi(f \in \mathcal{F}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_{L^\infty} \leq \epsilon_n/4)
\end{aligned} \tag{27}$$

We use (25) for the first inequality, and triangular inequality for the second inequality. The first equality comes from (24). The following inequalities is satisfied by previous lemma (13), and prior constraint (16), (17), (18), (19).

$$\begin{aligned}
&\Pi(f \in \mathcal{F}_n(\hat{\gamma}) : \|f - \hat{f}_n\|_{L^\infty} \leq \epsilon_n/4) \\
\geq &\Pi(\theta \in \mathbb{R}^{T_n} : \theta_{\hat{\gamma}^c} \in [-a_n, a_n]^{T_n - S_n}, \|\theta_{\hat{\gamma}}\|_\infty \leq B_n, \\
&\|\hat{\theta}_{\hat{\gamma}} - \theta_{\hat{\gamma}}\|_\infty \leq \frac{\epsilon_n}{4(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n - 1}}) \\
\geq &u_n^{T_n - S_n} (\int_{B_n - t_n}^{B_n} g(t)dt)^{S_n}
\end{aligned} \tag{28}$$

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

$$y_n = \int_{B_n - tn}^{B_n} g(t)dt \geq t_n g(B_n) \tag{29}$$

, we can induce following inequalities. (29) is true since g decreases on $[0,\infty)$.

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n,1}) \leq &- S_n \log y_n - (T_n - S_n) \log u_n \\
\leq &- S_n \log(t_n g(B_n)) - T_n(1 - S_n/T_n) \log(1 - S_n/T_n) \\
= &- S_n \log t_n - S_n \log g(B_n) + T_n(1 - S_n/T_n)(S_n/T_n + o(S_n/T_n)) \\
\lesssim &S_n (\log n)^2 + S_n + o(S_n) \\
\lesssim &n\epsilon_n^2
\end{aligned} \tag{30}$$

. The equality satisfies by Taylor expansion $-\log(1-x) = x + o(x)$, for $x = S_n/T_n$. Last inequality comes from (5). Thus, there exists a constant C_1 such that $-\log \Pi(A_{\epsilon_n,1}) \leq C_1 n\epsilon_n^2$.

(c) since $\mathcal{F}_n$ has extra constraint of a_n (16), (17), (18), (19) generated by the shrinkage prior, a function f in $\mathcal{F} - \mathcal{F}_n$ should either have a parameter θ_i that $|\theta_i| > B_n$,

or have more than S_n parameters that the absolute value is bigger than a_n. Thus,

$$
\begin{aligned}
&\Pi(\mathcal{F}-\mathcal{F}_n)\\
\leq&\pi(\exists|\theta_i|>B_n \mid L_n,W_n,S_n,B_n)+\pi(\sum_{i=1}^{T_n} I(|\theta_i|>a_n)>S_n|L_n,W_n,S_n,B_n)\\
=&(1-(1-v_n)^{T_n})+P(S>S_n|S\sim B(T_n,1-u_n))\\
\leq&T_n v_n+exp\left(-T_n\left\{(1-S_n/T_n)\log\frac{1-S_n/T_n}{u_n}+\frac{S_n}{T_n}\log\frac{S_n/T_n}{1-u_n}\right\}\right)\\
=&o(e^{-K_0 n\epsilon_n^2+\log T_n})+exp\left(T_n(1-S_n/T_n)\log\frac{u_n}{1-S_n/T_n}-S_n\log\frac{S_n/T_n}{1-u_n}\right)\\
\leq&o(e^{-K_1 n\epsilon_n^2})+exp\left(T_n(1-S_n/T_n)\log\left(\frac{1-\eta_n S_n/T_n}{1-S_n/T_n}+o\left(\frac{(1-\eta_n)S_n/T_n}{1-S_n/T_n}\right)\right)-Kn\epsilon_n^2\right)\\
=&o(e^{-K_1 n\epsilon_n^2})+o(e^{-K_2 n\epsilon_n^2})\\
=&o(e^{-\min\{K_1,K_2\}n\epsilon_n^2})
\end{aligned}
\tag{31}
$$

for some $4<K_1<K_0$ and $4<K_2<K$. The second inequality uses Bernouli inequality and lemma about tailbound of binomial distribution (14). The last inequality satisfies since the scale of T_n is determined by (5), and Taylor expansion. Let $C_2=(\min\{K_1,K_2\}-2)/\sigma^2$.

Let $C''=\min\{C_1,C_2\}$, to cover both the case (b), (c), and we can say the three condition (20), (21), (22) is satisfied.

Thus, theorem 3 is proved.

Conclusion

In theorem 3, shrinkage prior is introduced instead to deal with unsuitable computational cost of implementing spike-and-slab prior, preserving the asymptotic properties within the posterior distribution. We can catch the insight that we can substitute the prior with any prior that has several advantages, though the prior should satisfy strict tail conditions. The key difference to show contraction rate of posterior distribution via new prior was in lemma 3 (or lemma 2) in (Lee and Lee, 2022). Also, determining lower, upper bound of tail distribution led to the conditions ((7),(8) respectively) from the lemma. We can expect to get more decent conditions to prior by inducing tighter bounds for inequalities, though in larger perspective, the limitation of dimension-depended contraction rate of posterior distribution still exists.