Asymptotic Properties for Bayesian Neural Network in Besov Space

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural networks have shown great predictive power when dealing with various unstructured data such as images and natural languages. The Bayesian neural network captures the uncertainty of prediction by putting a prior distribution for the parameter of the model and computing the posterior distribution. In this paper, we show that the Bayesian neural network using spike-and-slab prior has consistency with nearly minimax convergence rate when the true regression function is in the Besov space. Even when the smoothness of the regression function is unknown the same posterior convergence rate holds and thus the spike and slab prior is adaptive to the smoothness of the regression function. We also consider the shrinkage prior and show that it has the same convergence rate. In other words, we propose a practical Bayesian neural network with guaranteed asymptotic properties.

1 Introduction

The neural network is a machine learning technique which can handle complex structures in data by combining a linear transformation and non-linear transformations [Goodfellow et al., 2016]. The neural network can show great predictive power when dealing with unstructured data ranging from image data to natural language processing. Moreover, the end-to-end learning is possible by grasping the inherent structure of data only with inputs and outputs without explicit intermediate steps. Such properties have played major roles in the rapid development of numerous artificial intelligence applications.

The Bayesian neural network (BNN) provides the uncertainty of the prediction from the probability distribution of the model parameters. Before observing the data, A prior distribution that reflects prior beliefs about the model is put over the parameter space of the neural network model. As observations are added, a posterior distribution is computed by updating the prior distribution via Bayes' rule. Neal [1996], Williams [1996] shows that a shallow BNN which has only one infinite-width hidden layer converges to a Gaussian process regression model. This means that the shallow BNN is an approximation of the nonparametric Bayesian regression model. Lee et al. [2017], Matthews et al. [2018] extended this result to the deep BNN model.

However, it is difficult to compute the exact posterior distribution of the BNN since the normalizing constant of the posterior distribution is intractable. Therefore, various methods have been proposed to infer BNN through approximate Bayesian computation algorithms. The methods based on the Markov chain Monte Carlo (MCMC) algorithm such as Gibbs sampler [Geman and Geman, 1984, Casella and George, 1992] and Hamiltonian Monte Carlo (HMC) [Neal et al., 2011] is effective to approximate true posterior. By constructing a Markov chain that has the desired distribution as its stationary distribution, one can obtain a sample of the desired distribution. Recently, scalable MCMC algorithms such as stochastic gradient Langevin dynamics [Welling and Teh, 2011] and stochastic gradient HMC [Chen et al., 2014] are proposed. The variational inference (VI) is an approximation

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

method that finds the closest variational distribution from the posterior distribution. VI method changes the Bayesian inference problem to an optimization problem. Graves [2011] proposed to use a Gaussian distribution as variational distribution for BNN and Kingma and Welling [2013] suggested the *reparameterization trick* for Gaussian variational distribution. Practically, VI methods for BNN are faster and more stable than MCMC methods. Blundell et al. [2015] suggested *Bayes by Backprop* algorithm for BNN via variational inference. There are also studies that approximate BNN with the dropout [Gal and Ghahramani, 2016] or the Gaussian process [Lee et al., 2017].

1.1 Related works

The performance of the neural network is justified by the *universal approximation* capability of the neural network. It is known that even a shallow neural network can approximate arbitrary continuous functions sufficiently accurately in L^p-sense [Cybenko, 1989, Hornik, 1991], and as the network deepens, the expressive power of the neural network gets stronger [Delalleau and Bengio, 2011, Bianchini and Scarselli, 2014, Telgarsky, 2015, 2016]. Lee III [1999] showed that the BNN has the universal approximation property.

As a measure of the expressive power of a model, consider the range of functional spaces which a model can express. The choice of the activation function can affect the expressive power of a model, and in this paper, we consider the Rectified Linear Unit (ReLU) [Nair and Hinton, 2010], which is the most popular activation function in practice. Yarotsky [2017] evaluated the approximation error of the neural network using the ReLU activation function (ReLU network) for functions in Hölder space. Schmidt-Hieber [2020] showed that the regularized least squared estimator performed by ReLU network in nonparametric regression problem converges to the true regression function with nearly minimax rate up to logarithmic factors on Hölder space. Suzuki [2018] extended the results in Schmidt-Hieber [2020] from the Hölder space to the Besov space which contains Hölder space and Sobolev space. Polson and Ročková [2018] proved a Bayesian version of the results in Schmidt-Hieber [2020]. They proved the Bayesian ReLU network has posterior consistency when the priors on the parameters of the ReLU network are spike-and-slab priors. Chérief-Abdellatif [2020] shows that the results in Polson and Ročková [2018] are still valid when variational methods are applied.

All of the above results assume a sparse structure in the neural networks. Sparsity is often assumed in high-dimensional models for asymptotic results. Castillo et al. [2015] showed theoretical properties including posterior consistency and rates of consistency of the high-dimensional linear regression model under the spike-and-slab prior. Song and Liang [2017] extended the results in Castillo et al. [2015] to general shrinkage priors which are more feasible than spike-and-slab prior.

1.2 Main contribution

We show that the posterior of the Bayesian ReLU network with a spike-and-slab prior with appropriate width, depth, and sparsity level is consistent with a nearly minimax rate. We extend the result of Suzuki [2018] to the Bayesian framework and the results of Polson and Ročková [2018] to the Besov space. As in Polson and Ročková [2018], we show that the same convergence rate is achieved even when the smoothness parameter of the true regression function is not known. That is, the ReLU network has rich expressive power, and this ability is not constrained even without the smoothness information of the true function. We get similar results for the shrinkage prior.

We try to narrow the gap between the study of the theoretical properties and the practical application of the Bayesian ReLU networks. For instance, Gaussian distribution is frequently considered as a prior distribution in practical Bayesian ReLU networks. In this case, the posterior distribution may not be consistent since the posterior becomes heavy tail distribution over the parameter space [Vladimirova et al., 2019]. Approaches to interpreting Bayesian neural networks as GP [Lee et al., 2017] avoid this problem by setting the variance of the prior distribution to be proportional to the width of the network. We suggest similar sufficient conditions for the posterior consistency of the Bayesian ReLU networks. Compared to precedent studys, our setting is closer to practical use. The results in Suzuki [2018] and Polson and Ročková [2018] require to solve optimization problem under the l_0 constraint or compute posterior from spike-and-slab prior respectively. We suggest BNN model, which is easy to infer and consistent with (nearly) optimal convergence rate.

The rest of the paper is organized as follows. We set up the problem and introduce the necessary concepts in Section 2. We present the obtained theoretical results in Section 3 and numerical examples in Section 4. A summary of the paper and a discussion are in Section 5.

2 Preliminaries

2.1 Notation

The floor function $\lfloor x \rfloor$ denotes the largest integer less than or equal to $x \in \mathbb{R}$, and $\|a\|_p$ is an l_p-norm of a real vector a. The u-th derivative of a d-dimensional function f is denoted by $D^u f$ with $u \in \{0, 1, \cdots\}^d$. The covering number $N(\epsilon, A, d)$ is the minimal number of d-balls of radius $\epsilon > 0$ necessary to cover a set A. We denote the dirac-delta function at $a \in \mathbb{R}$ by δ_a. We use $A \lesssim B$ as shorthand for the inequality $A \leq CB$ for some constant $C > 0$ and $A \vee B$ as shorthand for the $\max\{A, B\}$.

2.2 Model

Suppose that n input-output observations $D_n = (X_i, y_i)_{i=1}^n \subset [0,1]^d \times \mathbb{R}$ are i.i.d. 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 $N(0, \sigma^2)$ with known variance $\sigma^2 > 0$ and f_0 is the true regression function belonging to the space $\mathscr{F}$.

In frequentist statistics, the maximum likelihood estimation is used to estimate the function f_0. From the machine learning viewpoint, it is equivalent to finding $\hat{f}$ that minimizes the square loss. Suzuki [2018] showed that f_0 can be estimated with a nearly optimal rate in this setting.

In practice, a sparse regularization term $r(f)$ such as dropout and L_1-regularizer is also considered together with a loss function. This is equivalent to estimating the MAP (maximum a posteriori) in Bayesian modeling. The posterior distribution gives not only predictive values but also the uncertainty of the prediction. Thus, in Bayesian statistics, it is important to find sufficient conditions for a prior distribution or a regularizer which guarantees the consistency or optimal convergence rate of the posterior.

2.3 Functional spaces

In this section, we introduce functional spaces that statisticians have studied for *smoothness* of functions. Let $\Omega = [0,1]^d \subset \mathbb{R}^d$ be the domain of functions we consider.

The L^p-norm of a function is defined as follows:

$$\|f\|_{L^p} := \begin{cases} \left(\int_\Omega |f(x)|^p dx\right)^{1/p} & 0 < p < \infty, \\ \sup_{x \in \Omega} |f(x)| & p = \infty \end{cases}$$

where f is a real-valued function defined on Ω. The L^p space is the functional space consisting of functions with bounded L^p-norm and is denoted by $L^p(\Omega) = \{f : \|f\|_{L^p} < \infty\}$.

Let $s > 0$ be the smoothness parameter and $m = \lfloor s \rfloor$. The s-Hölder norm of function f is defined by

$$\|f\|_{C^s} := \sup_{\|u\|_1 \leq m} \|D^u f\|_{L^\infty} + \sup_{\|u\|_1 = m} \sup_{x, y \in \Omega} \frac{|\partial^u f(x) - \partial^u f(y)|}{|x - y|^{s-m}}.$$

The s-Hölder space $C^s(\Omega)$ is defined as a set of functions with bounded s-Hölder norm. We call s the smoothness parameter of the s-Hölder space. Let $k \in \mathbb{N}$ and $1 \leq p \leq \infty$. The Sobolev norm of a function f is defined by

$$\|f\|_{W^{k,p}} := \begin{cases} \left(\sum_{\|\alpha\|_1 \leq k} \|D^\alpha f\|_{L^p}^p\right)^{1/p} & 1 \leq p < \infty \\ \sup_{\|\alpha\|_1 \leq k} \|D^\alpha f\|_{L^\infty} & p = \infty. \end{cases}$$

The Sobolev space $W^{k,p}(\Omega)$ is defined by the functional space consisting of functions with bounded Sobolev norm.

Note the notion of the smoothness of a function is related to the differentiability of the function. The Besov space extends the concept of the smoothness. Before defining the Besov space, define the r-th modulus of smoothness of the function f as follows [Gine and Nickl, 2016]:

$$w_{r,p}(f,t) = \sup_{\|h\|_2 \le t} \|\Delta_h^r(f)\|_p, \tag{2}$$

$$\Delta_h^r(f)(x) = \begin{cases} \sum_{j=0}^r \binom{r}{j}(-1)^{r-j} f(x+jh) & x \in \Omega, x+rh \in \Omega, \\ 0 & \text{o.w.} \end{cases} \tag{3}$$

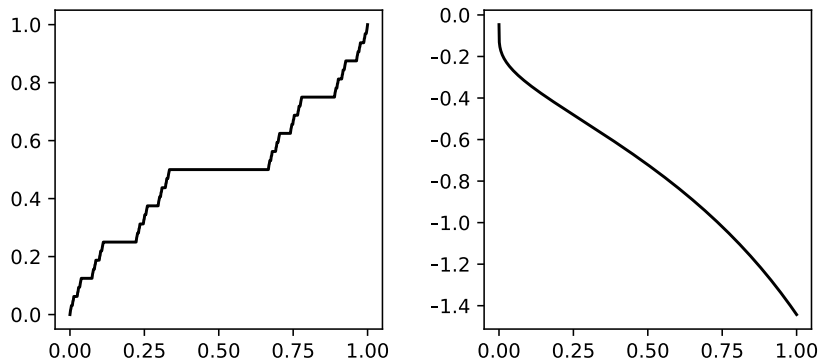
For $0 < p, q \le \infty,\ s > 0,\ r = \lfloor s \rfloor + 1$, define the Besov norm of a function f by

$$\|f\|_{B^s_{p,q}} := \|f\|_p + \begin{cases} \left(\int_0^\infty (t^{-s} w_{r,p}(f,t))^q \frac{dt}{t}\right)^{1/q} & q < \infty, \\ \sup_{t>0}\{t^{-s} w_{r,p}(f,t)\} & q = \infty. \end{cases} \tag{4}$$

The Besov space $B^s_{p,q}(\Omega)$ is defined as the set of functions with finite Besov norms [Gine and Nickl, 2016], i.e.,

$$B^s_{p,q}(\Omega) = \left\{ f : \|f\|_{B^s_{p,q}} < \infty \right\}. \tag{5}$$

Note that Besov spaces can be defined without the differentiability and continuity of functions, and



(a) $f_1(x)$ on $[0,1]$ (b) $f_2(x)$ on $[0,1]$

Figure 1: Example functions in Besov spaces

are more general than the Hölder space and the Sobolev space, which are subspaces of Besov spaces [Suzuki, 2018, Sawano, 2018, Triebel, 1983]. A Besov space may contain complicated functions that cannot be expressed by traditional statistical models. For instance, the Cantor function f_1, Figure 1 (a), does not belong to any Sobolev space since it does not have a weak derivative but belongs to a Besov space [Sawano, 2018]. In fact, $f_1 \in C^{\log 2/\log 3}([0,1]) \subset B^{\log 2/\log 3}_{\infty,\infty}([0,1])$. Also,

$$f_2(x) = \begin{cases} 1/\log(x/2), & 0 < x \le 1 \\ 0, & x = 0 \end{cases} \tag{6}$$

is in $B^{3/2}_{1,1}$ but not in any Hölder space. Figure 1 have plots of f_1 and f_2.

Note the minimax convergence rate of a Besov function f_0 in (1) has a lower bound $n^{-\frac{s}{2s+d}}$ [Gine and Nickl, 2016, Donoho and Johnstone, 1998].

2.4 Sparse neural network

Define (deep) neural network space generated by a parameter space Θ as follows:

$$\Phi(\Theta) = \Big\{ f_\theta(x) = (W^{(L+1)}(\cdot) + b^{(L+1)}) \circ \eta \circ \cdots \circ \eta \circ (W^{(1)}x + b^{(1)}) : \\ \theta = (W^{(1)}, b^{(1)}, \cdots, W^{(L+1)}, b^{(L+1)}) \in \Theta \Big\}. \tag{7}$$

In this paper, we consider the following B-bounded and S-sparse parameter space [Suzuki, 2018]

$$\begin{aligned}\Theta(L,W,S,B) := \Big\{ & W^{(l)} \in \mathbb{R}^{p_{l-1}\times p_l}, b^{(l)} \in \mathbb{R}^{p_l}, \\ & p_l = W, l = 1,2\cdots,L,\ p_0 = d, p_{L+1} = 1, \\ & \|\theta\|_0 = \sum_{l=1}^{L+1}\left(\left\|W^{(l)}\right\|_0 + \left\|b^{(l)}\right\|_0\right) \le S, \\ & \|\theta\|_\infty = \max_l \left\{\left\|W^{(l)}\right\|_\infty \vee \left\|b^{(l)}\right\|_\infty\right\} \le B\Big\}\end{aligned} \tag{8}$$

and denote S-sparse neural network space with B-bounded parameters $\Phi(\Theta(L,W,S,B))$ as $\Phi(L,W,S,B)$.

3 Main results

Consider $\mathscr{F}$ is a Besov space $B^s_{p,q}([0,1]^d)$. For each $n \in \mathbb{N}$, and $\theta \in \Theta$, Let $P^{(n)}_\theta$ be the probability measure of the data $D_n = (X_i, y_i)_{i=1}^n$ with a density function $p^{(n)}_\theta$ with respect to the measure $\mu^{(n)}$ and Π be a prior distribution of θ. The posterior distribution is given by

$$\Pi_n(A|D_n) = \frac{\int_A p^{(n)}_\theta(D_n)d\Pi(\theta)}{\int_\Theta p^{(n)}_\theta(D_n)d\Pi(\theta)}. \tag{9}$$

Let

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

be the norm on empirical measure $\mathbb{P}^X_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

Suppose that $0 < p, q \le \infty,\ \delta := d/p < s < \infty$ and set $\nu = (s-\delta)/(2\delta)$. 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

$$\begin{aligned} L_n = 3 + 2\lceil\log_2\left(\frac{3^{d\vee m}}{\tau_n c_{(d,m)}}\right) + 5\rceil\lceil\log_2(d\vee m)\rceil, \quad W_n = N_n W_0, \\ S_n = (L_n - 1)W_0^2 N_n + N_n, \quad B_n = O(N_n^\xi), \end{aligned} \tag{10}$$

where $c_{(d,m)} = \left(1 + 2de\frac{(2e)^m}{\sqrt{m}}\right)$, $\tau_n = N_n^{-s/d}(\log N_n)^{-1}$ and $\xi = \min\{1, \nu^{-1} + d^{-1}\}$.

Note the following theorems are the similar as the result in Suzuki [2018], a result for a frequentist statistics, and shows that the BNNs also converge in nearly minimax rate. Suzuki [2018] showed that for the least squared estimator $\hat{f} \in \Phi(L_n, W_n, S_n, B_n)$,

$$\mathbb{E}_{D_n}\left[\left\|f_0 - \hat{f}\right\|_n^2\right] \lesssim n^{-\frac{2s}{2s+d}}(\log n)^2, \tag{11}$$

where $\mathbb{E}_{D_n}[A]$ is a expectation of A with respect to the training data D_n.

Spike-and-slab prior

This result is an extension of Theorem 6.1 in Polson and Ročková [2018]. They showed that the Bayesian ReLU networks is consistent with nearly minimax rates on Hölder space. We extend the functional space from the Hölder space to the Besov space which contains Hölder spaces and Sobolev space. Assume the following spike-and-slab prior $\pi(\theta)$

$$\pi(\theta_j|\gamma_j) = \gamma_j\tilde{\pi}(\theta_j) + (1-\gamma_j)\delta_0(\theta_j), \quad \pi(\boldsymbol{\gamma}) = \frac{1}{\binom{T_n}{S_n}}, \tag{12}$$

where $\tilde{\pi} = U([-B_n, B_n])$ and $T_n = |\Theta(L_n, W_n, S_n, B_n)|$.

Theorem 1. *Under the model (1) and prior distribution (12), the posterior distribution concentrates at the rate* $\epsilon_n = n^{-s/(2s+d)}(\log n)^{3/2}$. *That is,*

$$\Pi(f_\theta \in \Phi(L_n, W_n, S_n, B_n) : \|f - f_0\|_n > M_n\epsilon_n | D_n) \to 0$$

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

Proof. See appendix C.3. □

Adaptive estimation

This result is an extension of Theorem 6.2 in Polson and Ročková [2018] to the Besov space. They showed that the Bayesian ReLU networks adapt to smoothness on the Hölder space under the spike-and-slab prior.

Theorem 2. *Assume the model (1), prior distribution (12),*

$$\begin{aligned} \pi(N) &= \frac{\lambda^N}{N!(e^\lambda - 1)} \text{ for } N = 1, 2, \cdots, \\ \pi(L) &= \frac{\rho^L}{L!(e^\rho - 1)} \text{ for } L = 1, 2, \cdots, \\ \pi(B) &= \beta e^{-\beta B} \text{ for } B > 0 \end{aligned} \tag{13}$$

for $\lambda,\ \rho,\ \beta > 0$, *and*

$$\Phi = \bigcup_{N=1}^{\infty} \bigcup_{L=1}^{\infty} \bigcup_{B=0}^{\infty} \Phi(L, NW_1, (L-1)W_1^2 N + N, B), \tag{14}$$

where $W_1 = 6dm(m+2) + 2d$ *for some* $m \in \mathbb{N}$. *Suppose that* $0 < p, q \leq \infty$ *and* $d/p < s < \min\{m, m-1+1/p\}$. *The posterior distribution concentrates at the rate* $\epsilon_n = n^{-s/(2s+d)}(\log n)^{3/2}$. *That is,*

$$\Pi\left(f_\theta \in \Phi : \|f - f_0\|_n > M_n\epsilon_n | D_n\right) \to 0$$

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

Proof. See appendix C.4. □

In Theorem 1, the prior distribution depends on the smoothness parameters p, q, and s. Theorem 2 shows that, as in Polson and Ročková [2018], a similar result can be obtained by considering the appropriate prior distribution even if the values of these parameters are unknown. Note it can be easily shown that BNNs do not overfit as in Polson and Ročková [2018].

Shrinkage prior

Asymptotic properties of the spike-and-slab prior can be obtained using Theorem 1, but its practical use for large neural network models is prohibited due to computation time. The following prior distributions are more practical than a spike-and-slab prior in terms of computation.

The result to be introduced in this section shows that all the aforementioned results still hold even under shrinkage prior. Assume the following prior

$$\pi(\theta) = \prod_{j=1}^{T_n} g(\theta_j), \tag{15}$$

where $T_n = |\Theta(L_n, W_n, S_n, B_n)|$ and $g(t)$ is acontinuous, symmetric density function and decreasing on $t > 0$.

Theorem 3. *Let $\epsilon_n = n^{-s/(2s+d)}(\log n)^{3/2}$. Under the model (1) and prior distribution (15), assume that*

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

satisfies

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

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

and

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

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

$$\Pi(f_\theta \in \Phi(L_n, W_n, T_n) : \|f - f_0\|_n > M_n\epsilon_n | D_n) \to 0$$

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

Proof. See appendix C.5. □

Note that condition (17) is a continuous approximation for the spike part of the (12). The next condition (18) means that the prior should have enough thick tail to sample true function. The last condition (19) restricts the thickness of the tail to prevent function divergent.

Unlike the other priors, the shrinkage prior is relatively straightforward to implement. For example, to implement the model in Theorem 1, one can consider the MCMC algorithm [Sun et al., 2022] which combines Gibbs sampler, the Metropolis-Hastings algorithm, and the SGHMC algorithm. Shrinkage prior avoids the posterior computation with varying dimensions, and enables feasible computation.

Example 1 (Gaussian mixture prior)**.** *Let $\epsilon_n = n^{-s/(2s+d)}(\log n)^{3/2}$, ψ and Ψ be density and inverse survival functions of the standard Gaussian distribution respectively. When $1/2 < \xi \leq 1$ and $B_n \asymp N_n^\xi$,*

$$g(t) = \pi_{1n}\frac{1}{\sigma_{1n}}\psi\left(\frac{t}{\sigma_{1n}}\right) + \pi_{2n}\frac{1}{\sigma_{2n}}\psi\left(\frac{t}{\sigma_{2n}}\right)$$

with

$$\pi_{2n} = \frac{S_n}{T_n},\ \pi_{1n} = 1 - \pi_{2n},$$
$$a_n \leq \frac{\epsilon_n}{72L_n(B_n \vee 1)^{L_n-1}(W_n+1)^{L_n}},\ \sigma_{2n}^2 = \frac{B_n^2}{2K_0 n\epsilon_n^2}$$

and

$$a_n/\Psi\left(\frac{\pi_{2n}}{\pi_{1n}}\left[\frac{1}{2}\eta_n - \Psi^{-1}\left(\frac{a_n}{\sigma_{2n}}\right)\right]\right) \leq \sigma_{1n} < a_n/\Psi\left(\frac{\pi_{2n}}{\pi_{1n}}\left[\frac{1}{2} - \Psi^{-1}\left(\frac{a_n}{\sigma_{2n}}\right)\right]\right)$$

satisfies condition of Theorem 3.

4 Numerical examples

We consider a problem of estimating the Besov functions f_1 and f_2. We sampled n points from $x \sim U(0,1)$ and y from $\sim N(f_1(x), \sigma_0^2)$ and $N(f_2(x), \sigma_0^2)$ respectively, where $\sigma_0^2 = 0.01$. We assume the Gaussian mixture prior distribution for each parameter as Example 1. Specifically, we set

$$B_n = 10N_n^\xi,\ K_0 = 5,\ \eta_n = \exp(-K_0 n\epsilon_n^2/S_n)$$
$$a_n = \frac{\epsilon_n}{72L_n(B_n \vee 1)^{L_n-1}(W_n+1)^{L_n}}$$
$$\sigma_{2n}^2 = \frac{B_n^2}{2(K_0+1)n\epsilon_n^2},\ \sigma_{1n} = a_n/\Psi\left(\frac{\pi_{2n}}{\pi_{1n}}\left[\frac{1}{2}\eta_n - \Psi^{-1}\left(\frac{a_n}{\sigma_{2n}}\right)\right]\right).$$

In Table 1 and Table 2, the parameters to estimate each function were presented. As the number of samples increases, the complexity of the neural network model increases. Its speed is affected by the complexity of the true regression function.

n	L_n	W_n	σ_{1n}	σ_{2n}	π_{1n}	π_{2n}
100	13	400	7.5103×10^{-61}	0.8443	0.8641	0.1359
1,000	15	1100	1.1191×10^{-82}	0.7597	0.9511	0.0489

Table 1: Model parameters of BNN for estimating $f_1 \in B_{\infty,\infty}^{\log 2/\log 3}([0,1])$.

n	L_n	W_n	σ_{1n}	σ_{2n}	π_{1n}	π_{2n}
100	13	200	1.5879×10^{-53}	0.6571	0.7290	0.2710
1,000	17	300	3.5489×10^{-75}	0.4023	0.8229	0.1771

Table 2: Model parameters of BNN for estimating $f_2 \in B_{1,1}^{3/2}([0,1])$.

We fit the model using the Bayes by Backprop algorithm with the python `PyTorch` package [Paszke et al., 2019]. Experiments were run on a GPU server with one Nvidia RTX 3090. The code and instruction related to the experiment were provided in the supplementary material. Figure 2 and Figure 3 show the results. For both problems, as the number of the data increase, mean functions get closer to the true regression function and empirical errors decrease.

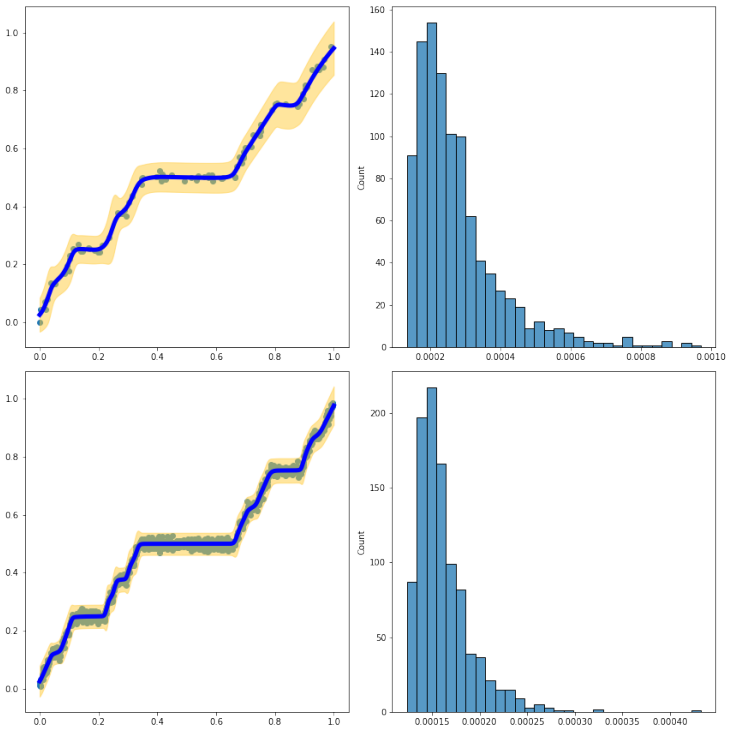


Figure 2: The results of estimating f_1 for $n = 100$ (top) and $1,000$ (bottom) samples using BNN model with parameters in Table 1. We draw 1,000 functions from the variational distribution and plot mean function with training data (left) and histogram of the empirical error $\|f - f_0\|_n$ (right). The blue lines are mean function and the yellow intervals are the prediction interval.

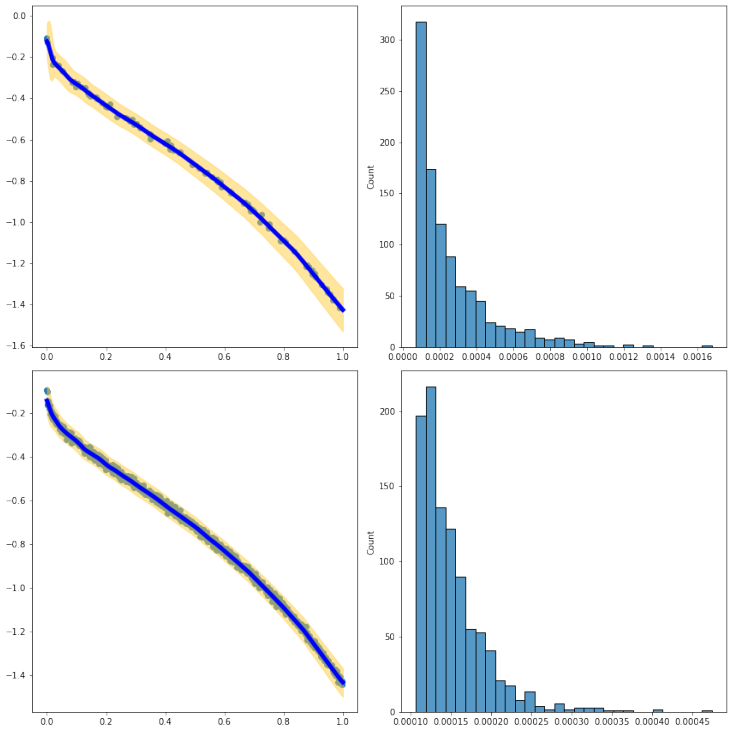


Figure 3: The results of estimating f_2 with $n = 100$ (top) and $1,000$ (bottom) samples using BNN model with parameters in Table 2. We draw 1,000 functions from the variational distribution and plot mean function with training data (left) and histogram of the empirical error $\|f - f_0\|_n$ (right). The blue lines are mean function and the yellow intervals are the prediction interval.

5 Conclusions

In this paper, we justify the theoretical properties of neural networks that are currently widely used. Specifically, we focused on the asymptotic properties of BNNs and extend results in Polson and Ročková [2018] to the Besov space from the Hölder space. Theorem 1 shows that the BNN with a shrinkage prior has consistency posterior with a nearly minimax rate at Besov functions with known smoothness. Theorem 3 extends the result to the shrinkage prior respectively to closing the gap. According to these theorems, BNNs have posterior consistency and (nearly) optimal rates by assuming proper prior such as Gaussian mixture prior. We found there are three sufficient conditions to achieve posterior consistency, (1) enough mass around zero, (2) thick tail to sample true function (3) while not too much. Theorem 2 shows that even if the smoothness of the function is not known, the same rate can be obtained by considering a prior on the architecture of the neural networks.

The results in this paper are an extension of the results of Suzuki [2018] to BNN also. We suggest a practical method that can achieve the optimal rate while providing the uncertainty of the prediction. Note they showed the theoretical properties of the neural networks in the Besov space via nonparametric regression based on B-spline. In other words, similar results can be obtained using the Bayesian LARK B-spline model [Park et al., 2021]. Experiments on performance comparison with other models will also be of great help in establishing the properties of BNN.

We obtained results under the limited assumption, setting the networks fully connected and using the ReLU activation function. The results of this study are at a basic level compared to the needs for BNNs in the industrial field, such as visual data analysis and natural language processing. It is important to research the theoretical properties of BNNs used in the practical area through ongoing studies.

References

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

M. Bianchini and F. Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE transactions on neural networks and learning systems*, 25 (8):1553–1565, 2014.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

I. Castillo, J. Schmidt-Hieber, and A. Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

T. Chen, E. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *International conference on machine learning*, pages 1683–1691. PMLR, 2014.

B.-E. Chérief-Abdellatif. Convergence rates of variational inference in sparse deep learning. In *International Conference on Machine Learning*, pages 1831–1842. PMLR, 2020.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

O. Delalleau and Y. Bengio. Shallow vs. deep sum-product networks. *Advances in neural information processing systems*, 24:666–674, 2011.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921, 1998.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

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

E. Gine and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016. ISBN 978-1-107-33786-2. doi: 10.1017/CBO9781107337862. URL `http://ebooks.cambridge.org/ref/id/CBO9781107337862`.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

A. Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2): 251–257, 1991.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

H. K. H. Lee III. *Model selection and model averaging for neural networks*. PhD thesis, Carnegie Mellon University, 1999.

A. G. d. G. Matthews, M. Rowland, J. Hron, R. E. Turner, and Z. Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

R. M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.

R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11): 2, 2011.

S. Park, H.-S. Oh, and J. Lee. Lévy adaptive b-spline regression via overcomplete systems. *arXiv preprint arXiv:2101.12179*, 2021.

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.

Y. Sawano. *Theory of Besov Spaces*, volume 56 of *Developments in Mathematics*. Springer Singapore, 2018. ISBN 9789811308352. doi: 10.1007/978-981-13-0836-9. URL `http://link.springer.com/10.1007/978-981-13-0836-9`.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Q. Song and F. Liang. Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*, 2017.

Y. Sun, Q. Song, and F. Liang. Learning sparse deep neural networks with a spike-and-slab prior. *Statistics & Probability Letters*, 180:109246, 2022.

T. Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

M. Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

M. Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

H. Triebel. Theory of function spaces, vol. 78 of. *Monographs in mathematics*, 1983.

M. Vladimirova, J. Verbeek, P. Mesejo, and J. Arbel. Understanding priors in bayesian neural networks at the unit level. In *International Conference on Machine Learning*, pages 6458–6467. PMLR, 2019.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

C. Williams. Computing with infinite networks. *Advances in neural information processing systems*, 9, 1996.

D. Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94: 103–114, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 3.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See supplementary material.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

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*

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

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

$$K(f,g) = \mathbb{E}^f\left[\log\frac{f(X)}{g(X)}\right] = \int f\log\frac{f}{g}d\mu, \tag{20}$$

$$V_k(f,g) = \mathbb{E}^f\left[\left|\log\frac{f(X)}{g(X)}\right|^k\right], \tag{21}$$

$$V_{k,0}(f,g) = \mathbb{E}^f\left[\left|\log\frac{f(X)}{g(X)} - K(f,g)\right|^k\right] \tag{22}$$

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

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

for $k \ge 2$. The existence of C_k is guaranteed by Marcinkiewiz-Zygmund inequality. Then the followings hold

Lemma 2 (Theorem 4 of Ghosal and Van Der Vaart [2007])**.** *Let $P_\theta^{(n)}$ be product measures and $d_n = \frac{1}{n}\sum_{i=1}^n\int(\sqrt{p_{\theta,i}} - \sqrt{p_{\theta',i}})^2 d\mu_i$. 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 $\Theta_n \subset \Theta$, the following conditions hold:*

$$\sup_{\epsilon>\epsilon_n}\log N(\epsilon/36, \{\theta\in\Theta_n : d_n(\theta,\theta_0) < \epsilon\}, d_n) \le n\epsilon_n^2, \tag{23}$$

$$\frac{\Pi_n(\Theta\setminus\Theta_n)}{\Pi_n(B_n^*(\theta_0,\epsilon_n;k))} = o\left(e^{-2n\epsilon_n^2}\right), \tag{24}$$

$$\frac{\Pi_n(\theta\in\Theta_n : j\epsilon_n < d_n(\theta,\theta_0) \le 2j\epsilon_n)}{\Pi_n(B_n^*(\theta_0,\epsilon_n;k))} \le e^{n\epsilon_n^2 j^2/4} \tag{25}$$

Then $P_{\theta_0}^{(n)}\left(\Pi_n(\theta : d_n(\theta,\theta_0) \ge M_n\epsilon_n|D_n)\right. \to 0$ for any sequence $M_n \to \infty$.

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

The following is an adapted version of the Lemma 2, which introduced in Polson and Ročková [2018] (white noise model and $k = 2$).

Assume the model (1). Let

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

If there exist $d > 2$ and $\mathcal{F}_n \subset \mathcal{F}$ such that

$$\sup_{\epsilon > \epsilon_n} \log N(\epsilon/36, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n) \leq n\epsilon_n^2, \tag{26}$$

$$\Pi(A_{\epsilon_n,1}) \geq e^{-dn\epsilon_n^2}, \tag{27}$$

$$\Pi(\mathcal{F} - \mathcal{F}_n) = o\left(e^{-(d+2)n\epsilon_n^2}\right) \tag{28}$$

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

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

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

B Consistency of neural network

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

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

$$\log N(\delta, \Phi(L, W, S, B), \|\cdot\|_{L^\infty}) \leq (S+1) \log\left(2\delta^{-1} L (B \vee 1)^L (W+1)^{2L}\right) \tag{29}$$

Proof. See appendix C.1 □

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

$$\Phi(L, W, S, B, a) = \{f_\theta : \theta \in \Theta(L, W, S, B, a)\}, \tag{30}$$

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

$$\log N(\delta, \Phi(L, W, S, B, a), \|\cdot\|_{L^\infty}) \leq (S+1) \log\left(2\delta^{-1} L (B \vee 1)^L (W+1)^{2L}\right) \tag{31}$$

Proof. See appendix C.2 □

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

Lemma 5 (Proposition 1 in Suzuki [2018] with $r = \infty$)**.** *Suppose that $0 < p, q \leq \infty$, $\delta := d/p < s < \infty$ and $\nu = (s - \delta)/(2\delta)$. 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,*

$$\sup_{f_0 \in B^s_{p,q}([0,1]^d)} \inf_{f \in \Phi(L,W,S,B)} \|f_0 - f\|_{L^\infty} \lesssim N^{-s/d} \tag{32}$$

for

$$\begin{aligned} L &= 3 + 2\lceil \log_2\left(\frac{3^{d \vee m}}{\tau_n c_{(d,m)}}\right) + 5\rceil \lceil \log_2(d \vee m)\rceil, \quad W = NW_0, \\ S &= (L-1)W_0^2 N + N, \quad B = O\left(N^{\nu^{-1} + d^{-1}}\right), \end{aligned} \tag{33}$$

where $c_{(d,m)} = \left(1 + 2de\frac{(2e)^m}{\sqrt{m}}\right)$ and $\tau_n = N_n^{-s/d}(\log N_n)^{-1}$.

Proof. See Suzuki [2018]. □

C Proof of theorems

C.1 Proof of lemma 3

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

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

$$\begin{aligned} \mathcal{A}_k^+(f)(x) &= \eta \circ (W^{(k-1)}\eta(\cdot) + b^{(k-1)}) \circ \cdots \circ (W^{(1)}x + b^{(1)}), \\ \mathcal{A}_k^-(f)(x) &= (W^{(L)}\eta(\cdot) + b^{(L)}) \circ \cdots \circ (W^{(k)}x + b^{(k)}) \end{aligned} \tag{34}$$

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

$$\begin{aligned} \left\|\mathcal{A}_k^+(f)\right\|_{L^\infty} &\le \max_j \left\|W_{j,:}^{(k-1)}\right\|_1 \left\|\mathcal{A}_{k-1}^+(f)\right\|_{L^\infty} + \left\|b^{(k-1)}\right\|_{L^\infty} \\ &\le WB\left\|\mathcal{A}_{k-1}^+(f)\right\|_{L^\infty} + B \\ &\le (W+1)(B \vee 1)\left\|\mathcal{A}_{k-1}^+(f)\right\|_{L^\infty} \\ &\le (W+1)^{k-1}(B \vee 1)^{k-1}, \end{aligned} \tag{35}$$

and similarly

$$\left|\mathcal{A}_k^-(f)(x_1) - \mathcal{A}_k^-(f)(x_2)\right| \le (BW)^{L-k+1}\|x_1 - x_2\|_\infty. \tag{36}$$

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

$$\begin{aligned} &|f_\theta(x) - f_{\theta^*}(x)| \\ &= \left|\sum_{k=1}^L \mathcal{A}_{k+1}^-(f_{\theta^*}) \circ (W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^+(f_\theta)(x) - \mathcal{A}_{k+1}^-(f_{\theta^*}) \circ (W^{(k)^*} \cdot + b^{(k)^*}) \circ \mathcal{A}_k^+(f_\theta)(x)\right| \\ &\le \sum_{k=1}^L (BW)^{L-k}\left\|(W^{(k)} \cdot + b^{(k)}) \circ \mathcal{A}_k^+(f_\theta)(x) - (W^{(k)^*} \cdot + b^{(k)^*}) \circ \mathcal{A}_k^+(f_\theta)(x)\right\|_{L^\infty} \\ &\le \sum_{k=1}^L (BW)^{L-k}\epsilon\left[W(B \vee 1)^{k-1}(W+1)^{k-1} + 1\right] \\ &\le \sum_{k=1}^L (BW)^{L-k}\epsilon(B \vee 1)^{k-1}(W+1)^k \\ &\le \epsilon L(B \vee 1)^{L-1}(W+1)^L. \end{aligned} \tag{37}$$

Let s be the number of nonzero component 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 $\theta_1, \cdots, \theta_M$ form each grid divided with length $\epsilon^* = \frac{\delta}{L(B\vee 1)^{L-1}(W+1)^L}$ over $\Theta_\theta(L, W, S, B)$. Then

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

Note

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

and the number of cases of choose s nonzero components is

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

Thus,

$$
\begin{aligned}
N(\delta, \Phi(L,W,S,B), \|\cdot\|_{L^\infty}) &\leq \sum_{s^* \leq S} \binom{T}{s^*} (2B\delta^{-1}(L(B\vee 1)^{L-1}(W+1)^L)^{s^*} \\
&\lesssim \sum_{s^* \leq S} (2\delta^{-1}L(B\vee 1)^L(W+1)^{2L})^{s^*} \\
&\leq (2\delta^{-1}L(B\vee 1)^L(W+1)^{2L})^{S+1}
\end{aligned} \tag{38}
$$

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

C.2 Proof of lemma 4

Proof. Let $\tilde{\theta}(a) = (\theta_i I(|\theta_i| > a))$. Then $\left\|\theta - \tilde{\theta}(a)\right\|_\infty \leq a$ 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 3,

$$
\left\|f_\theta - f_{\tilde{\theta}(a)}\right\|_{L^\infty} \leq aL(B\vee 1)^{L-1}(W+1)^L \leq \delta/2 \tag{39}
$$

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

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

from triangular inequality. We get desired results in the similar way as in the proof of the Lemma 3. □

C.3 Proof of theorem 1

Proof. Let

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

It is enough to show that there exists a constant $C > 2$ which satisfies

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

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

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

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

$$
\mathcal{F} - \mathcal{F}_n = \emptyset
$$

.

Since

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

and Lemma 3,

$$\begin{aligned}
\sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1}\cap\mathcal{F}_n, \|\cdot\|_n\right) &\le \sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1}\cap\mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\le \sup_{\epsilon>\epsilon_n} \log N\left(\frac{\epsilon}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\le \log N\left(\frac{\epsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\
&\lesssim (S_n+1)\left[L_n \log((B_n\vee 1)(W_n+1)) - \log\epsilon_n\right] \\
&\lesssim N_n(\log n)^3 \\
&= n\epsilon_n^2
\end{aligned} \tag{40}$$

(a) holds. Here, the last inequality holds from

$$L_n = O(\log n),\ W_n = O(N_n),\ S_n = O(N_n \log n) \tag{41}$$

Next, from Lemma 5, there is a constant $C_1 > 0$ and $\hat{f}_n = f_{\hat{\theta}} \in \mathcal{F}$ such that

$$\left\|\hat{f}_n - f_0\right\|_{L^\infty} \le C_1 N^{-s/d} \lesssim \epsilon_n/2. \tag{42}$$

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 $\mathcal{F}(\hat{\gamma}) = \Phi(\hat{\gamma}; L_n, W_n, S_n, B_n)$ be the neural network space generated by $\Theta(\hat{\gamma}; L_n, W_n, S_n, B_n)$. Note

$$\begin{aligned}
\Pi(A_{\epsilon_n,1}) &= \Pi(f\in\mathcal{F} : \|f-f_0\|_n \le \epsilon_n) \\
&\ge \Pi(f\in\mathcal{F} : \|f-f_0\|_{L^\infty} \le \epsilon_n) \\
&\gtrsim \Pi\left(f\in\mathcal{F} : \left\|f-\hat{f}_n\right\|_{L^\infty} \le \epsilon_n/2\right) \\
&\ge \Pi\left(f\in\mathcal{F}(\hat{\gamma}) : \left\|f-\hat{f}_n\right\|_{L^\infty} \le \epsilon_n/2\right)
\end{aligned} \tag{43}$$

for sufficiently large n. As in the proof of Lemma 3,

$$\begin{aligned}
&\Pi\left(f\in\mathcal{F}_n(\hat{\gamma}) : \left\|f-\hat{f}_n\right\|_{L^\infty} \le \epsilon_n/2\right) \\
&\ge \Pi\left(\theta\in\mathbb{R}^{T_n} : \theta_{\hat{\gamma}^c} = 0,\ \|\theta\|_\infty \le B_n,\ \left\|\hat{\theta}-\theta\right\|_\infty \le \frac{\epsilon_n}{2(W_n+1)^{L_n}L_n(B_n\vee 1)^{L_n-1}}\right) \\
&\ge \left(\frac{\epsilon_n}{2B_n(W_n+1)^{L_n}L_n(B_n\vee 1)^{L_n-1}}\right)^{S_n} \\
&= \exp\left(-S_n \log\left(\frac{2B_n(W_n+1)^{L_n}L_n(B_n\vee 1)^{L_n-1}}{\epsilon_n}\right)\right)
\end{aligned} \tag{44}$$

Thus,

$$\begin{aligned}
-\log\Pi(A_{\epsilon_n,1}) &\le S_n \log\left(\frac{2B_n(W_n+1)^{L_n}L_n(B_n\vee 1)^{L_n-1}}{\epsilon_n}\right) \\
&\le S_n\left[L_n\log((W_n+1)(B_n\vee 1)) + \log 2L_n - \log\epsilon_n\right] \\
&\lesssim N_n(\log n)^3 \\
&= n\epsilon_n^2
\end{aligned} \tag{45}$$

for sufficiently large n. □

C.4 Proof of theorem 2

Proof. Let $\mathcal{F} = \Phi$. It is enough to show that there exists a constant $C' > 2$ which satisfies

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

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

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

for sufficiently large n. For sufficiently large constants $C_N,\ C_B > 0$, let $\tilde{N}_n = C_N N_n$, $\tilde{L}_n = L(\tilde{N}_n)$, $\tilde{B}_n = C_B n\epsilon_n^2$ and

$$\mathcal{F}_n = \bigcup_{N=1}^{\tilde{N}_n} \bigcup_{L=1}^{\tilde{L}_n} \bigcup_{B=0}^{\tilde{B}_n} \Phi(L, NW_1, (L-1)W_1^2 N + N, (L-1)W_1^2 N + N, B), \tag{46}$$

First, show that (a) holds. From Lemma 3,

$$\begin{aligned} N\left(\frac{\epsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) &\le \sum_{N=1}^{\tilde{N}_n} \sum_{L=1}^{\tilde{L}_n} \sum_{B=0}^{\tilde{B}_n} \left(\frac{36}{\epsilon_n} L(B\vee 1)^{L-1}(NW_1+1)^{2L}\right)^{(L-1)W_1^2 N + N + 1} \\ &\lesssim \tilde{N}_n \tilde{L}_n \tilde{B}_n \left(\frac{36}{\epsilon_n} \tilde{L}_n (\tilde{B}_n \vee 1)^{\tilde{L}_n - 1}(\tilde{N}_n W_1 + 1)^{2\tilde{L}_n}\right)^{(\tilde{L}_n-1)W_1^2 \tilde{N}_n + \tilde{N}_n + 1} \end{aligned} \tag{47}$$

and

$$\begin{aligned} &\sup_{\epsilon > \epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_n\right) \\ &\le \sup_{\epsilon > \epsilon_n} \log N\left(\frac{\epsilon}{36}, A_{\epsilon,1} \cap \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\ &\le \sup_{\epsilon > \epsilon_n} \log N\left(\frac{\epsilon}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\ &\le \log N\left(\frac{\epsilon_n}{36}, \mathcal{F}_n, \|\cdot\|_{L^\infty}\right) \\ &\lesssim \log \tilde{N}_n + \log \tilde{L}_n + \log \tilde{B}_n \\ &\quad + \left[(\tilde{L}_n - 1)W_1^2 \tilde{N}_n + \tilde{N}_n + 1\right] \log\left(\frac{36}{\epsilon_n} \tilde{L}_n (\tilde{B}_n \vee 1)^{\tilde{L}_n}(\tilde{N}_n W_1 + 1)^{2\tilde{L}_n}\right) \\ &\lesssim \tilde{N}_n (\log n)^3 \\ &\lesssim n\epsilon_n^2. \end{aligned} \tag{48}$$

Next, show that (b) holds. For N_n, L_n in Theorem 1, $N_n < \tilde{N}_n,\ L_n < \tilde{L}_n$. Moreover, since $B_n \le N_n \lesssim n\epsilon_n^2$, there exists a constant $D > 0$ such that

$$\begin{gathered} \pi(N_n) = \frac{\lambda^{N_n}}{N_n!(e^\lambda - 1)} \gtrsim \exp\left(-N_n \log \frac{N_n}{\lambda}\right) \ge \exp\left(-\tilde{N}_n \log \frac{\tilde{N}_n}{\lambda}\right) \gtrsim \exp\left(-Dn\epsilon_n^2\right), \\ \exp\left(-\tilde{L}_n \log \frac{\tilde{L}_n}{\lambda}\right) \gtrsim \exp\left(-Dn\epsilon_n^2\right), \\ \pi(B_n) \gtrsim \exp(-\beta B_n) \gtrsim \exp\left(-Dn\epsilon_n^2\right). \end{gathered} \tag{49}$$

Then

$$\begin{aligned} &\Pi(f_\theta \in \mathcal{F}_n : \|f - f_0\|_n \le \epsilon_n) \\ &\ge \pi(N_n)\pi(L_n)\pi(B_n)\Pi(f_\theta \in \Phi(L_n, W_n, S_n, B_n) : \|f - f_0\|_n \le \epsilon_n) \\ &\gtrsim \exp\left(-(C+3D)n\epsilon_n^2\right) \end{aligned} \tag{50}$$

holds and (b) holds for $C' = C + 3D$.
From

$$\Pi(\mathcal{F} - \mathcal{F}_n) < \pi(N > \tilde{N}_n) + \pi(L > \tilde{L}_n) + \pi(B > \tilde{B}_n)$$

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

$$\begin{gathered} \pi(N > \tilde{N}_n) < e^{-t(\tilde{N}_n+1)}\mathbb{E}\left[e^{tN}\right] \lesssim e^{-t(\tilde{N}_n+1)}\left(\exp\left(e^t\lambda\right) - 1\right), \\ \pi(L > \tilde{L}_n) \lesssim e^{-t(\tilde{L}_n+1)}\left(\exp\left(e^t\rho\right) - 1\right). \end{gathered} \tag{51}$$

Letting $t = \log \tilde{N}_n,\ \log \tilde{L}_n$ respectively, we get

$$\begin{aligned} \pi(N > \tilde{N}_n) &\lesssim e^{-(\tilde{N}_n+1)\log \tilde{N}_n + \tilde{N}_n \lambda}, \\ \pi(L > \tilde{L}_n) &\lesssim e^{-(\tilde{L}_n+1)\log \tilde{L}_n + \tilde{L}_n \rho}. \end{aligned} \tag{52}$$

Also, since

$$\begin{aligned} (C'+2)n\epsilon_n^2 + \lambda \tilde{N}_n - (\tilde{N}_n+1)\log \tilde{N}_n &\to -\infty, \\ (C'+2)n\epsilon_n^2 + \rho \tilde{L}_n - (\tilde{L}_n+1)\log \tilde{L}_n &\to -\infty, \\ (C'+2)n\epsilon_n^2 - \beta \tilde{B}_n &\to -\infty, \end{aligned} \tag{53}$$

for sufficiently large $C_N,\ C_B > 0$ and $\pi(B > \tilde{B}_n) = \exp\Big(-\beta \tilde{B}_n\Big)$, (c) holds. □

C.5 Proof of theorem 3

Proof. It is enough to show that there exists a constant $C > 2$ which satisfies

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

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

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

for sufficiently large n. Let $\mathcal{F}_n = \Phi(L_n, W_n, S_n, B_n, a_n)$ and $\mathcal{F} = \Phi(L_n, W_n, T_n)$. It is easy to show that (a) holds from Lemma 4 as in the proof of the Theorem 1. By the assumption,

$$\begin{aligned} &\Pi(\mathcal{F} - \mathcal{F}_n) \\ &\le \pi\left(^{\exists}|\theta_i| > B_n\right) + \pi\left(\sum_{i=1}^{T_n} I(|\theta_i| > a_n) > S_n\right) \\ &= \left(1 - (1-v_n)^{T_n}\right) + P\left(S > S_n | S \sim B(T_n, 1-u_n)\right) \\ &\le T_n v_n + o(|v_n|) + \exp\left(-T_n\left\{\left(1 - \frac{S_n}{T_n}\right)\log\frac{1 - S_n/T_n}{u_n} + \frac{S_n}{T_n}\log\frac{S_n/T_n}{1-u_n}\right\}\right) \\ &= o\left(e^{-K_0 n\epsilon_n^2}\right) + \exp\left(T_n\left(1 - \frac{S_n}{T_n}\right)\log\frac{u_n}{1 - S_n/T_n} - S_n \log\frac{S_n/T_n}{1-u_n}\right) \\ &\le o\left(e^{-K_0 n\epsilon_n^2}\right) + \exp\left(T_n\left(1 - \frac{S_n}{T_n}\right)\log\frac{1 - \eta_n S_n/T_n}{1 - S_n/T_n} - S_n \log\frac{S_n/T_n}{1-u_n}\right) \\ &= o\left(e^{-K_0 n\epsilon_n^2}\right) + \exp\left(T_n\left(1 - \frac{S_n}{T_n}\right)\left(\frac{(1-\eta_n)S_n/T_n}{1 - S_n/T_n} + o\left(\frac{(1-\eta_n)S_n/T_n}{1 - S_n/T_n}\right)\right) - Kn\epsilon_n^2\right) \\ &= o\left(e^{-K_0 n\epsilon_n^2}\right) + \exp\left(S_n(1-\eta_n) + o(S_n(1-\eta_n)) - Kn\epsilon_n^2\right) \\ &= o\left(e^{-K_0 n\epsilon_n^2}\right) + o\left(e^{-K_1 n\epsilon_n^2}\right) \\ &= o\left(e^{-\min\{K_0,K_1\} n\epsilon_n^2}\right). \end{aligned} \tag{54}$$

for some $4 < K_1 < K$. Letting $C = \min\{K_0, K_1\} - 2$, (c) holds. We use tail bound for binomial distribution in Arratia and Gordon [1989] for the second inequality. Next, from Lemma 5, there is a constant $C_1 > 0$ and $\hat{f}_n = f_{\hat{\theta}} \in \mathcal{F}_n$ such that

$$\left\|\hat{f}_n - f_0\right\|_{L^\infty} \le C_1 N^{-s/d} \lesssim \epsilon_n/2. \tag{55}$$

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 $\mathcal{F}_n(\hat{\gamma}) = \tilde{\Phi}(\hat{\gamma}; L_n, W_n, S_n, B_n, a_n)$ be the neural network space generated by $\tilde{\Theta}(\hat{\gamma}; L_n, W_n, S_n, B_n, a_n)$. Note

$$\begin{aligned}
\Pi(A_{\epsilon_n,1}) &= \Pi(f \in \mathcal{F} : \|f - f_0\|_n \leq \epsilon_n) \\
&\geq \Pi\left(f \in \mathcal{F} : \left\|f - \hat{f}_n\right\|_n \leq \epsilon_n/2\right) \\
&\geq \Pi\left(f \in \mathcal{F} : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/2\right) \\
&\geq \Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/2\right)
\end{aligned} \tag{56}$$

for sufficiently large n. As in the proof of Lemma 3,

$$\begin{aligned}
&\Pi\left(f \in \mathcal{F}_n(\hat{\gamma}) : \left\|f - \hat{f}_n\right\|_{L^\infty} \leq \epsilon_n/2\right) \\
&\geq \Pi\left(\theta \in \mathbb{R}^{T_n} : \theta_{\hat{\gamma}^c} \in [-a_n, a_n]^{T_n - S_n},\ \|\theta_{\hat{\gamma}}\|_\infty \leq B_n,\ \left\|\hat{\theta}_{\hat{\gamma}} - \theta_{\hat{\gamma}}\right\|_\infty \leq \frac{\epsilon_n}{2(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n - 1}}\right) \\
&\geq u_n^{T_n - S_n} \left(\int_{B_n - t_n}^{B_n} g(t)dt\right)^{S_n},
\end{aligned} \tag{57}$$

where $t_n = \frac{\epsilon_n}{2(W_n+1)^{L_n} L_n (B_n \vee 1)^{L_n - 1}}$. Letting

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

$$\begin{aligned}
-\log \Pi(A_{\epsilon_n,1}) &\leq -S_n \log y_n - (T_n - S_n) \log u_n \\
&\leq -S_n \log(t_n g(B_n)) - T_n \left(1 - \frac{S_n}{T_n}\right) \log(1 - S_n/T_n) \\
&= -S_n \log t_n - S_n \log g(B_n) + T_n \left(1 - \frac{S_n}{T_n}\right) (S_n/T_n + o(S_n/T_n)) \\
&\lesssim S_n (\log n)^2 + S_n + o(S_n) \\
&\lesssim n\epsilon_n^2.
\end{aligned} \tag{58}$$

□