Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

ASYMPTOTIC PROPERTIES OF NEURAL NETWORK SIEVE ESTIMATORS

By Xiaoxi Shen[†,‡] , Chang Jiang[‡] Lyudmila Sakhanenko[†]
and Qing Lu[*,‡]

Michigan State University[†]
University of Florida[‡]

Neural networks have become one of the most popularly used methods in machine learning and artificial intelligence. Due to the universal approximation theorem (Hornik, Stinchcombe and White, 1989), a neural network with one hidden layer can approximate any continuous function on compact support as long as the number of hidden units is sufficiently large. Statistically, a neural network can be classified into a nonlinear regression framework. However, if we consider it parametrically, due to the unidentifiability of the parameters, it is difficult to derive its asymptotic properties. Instead, we consider the estimation problem in a nonparametric regression framework and use the results from sieve estimation to establish the consistency, the rates of convergence and the asymptotic normality of the neural network estimators. We also illustrate the validity of the theories via simulations.

1. Introduction. With the success of machine learning and artificial intelligence in researches and industry, neural networks have become popularly used methods nowadays. Many newly developed machine learning methods are based on deep neural networks and have achieved great classification and prediction accuracy. We refer interested readers to Goodfellow et al. (2016) for more background and details. In classical statistical learning theory, the consistency and the rate of convergence of the empirical risk minimization principle are of great interest. Many upper bounds have been established for the empirical risk and the sample complexity based on the growth function and the Vapnik-Chervonenkis dimension (see for example, Vapnik (1998); Anthony and Bartlett (2009); Devroye, Györfi and Lugosi (2013)). However, few studies have focused on the asymptotic properties for neural networks. As Thomas J. Sargent said, “artificial intelligence is actually statistics, but in a very gorgeous phrase, it is statistics.” So it is natural and worthwhile to explore whether neural networks possess nice asymptotic properties. As if they do, it may be possible to conduct statistical infer-

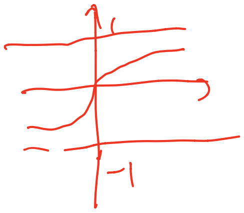
[*] To whom correspondence should be addressed.

Keywords and phrases: Empirical Processes, Entropy Integral

ence based on neural networks. Throughout this paper, we will focus on the asymptotic properties of neural networks with one hidden layer.

In statistics, fitting a neural network with one hidden layer can be viewed as a parametric nonlinear regression problem:

$$y_i = \alpha_0 + \sum_{j=1}^{r} \alpha_j \sigma\left(\boldsymbol{\gamma}_j^T \boldsymbol{x}_i + \gamma_{0,j}\right) + \epsilon_i,$$



where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. random errors with $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] = \sigma^2 < \infty$ and $\sigma(\cdot)$ is an activation function such as $\sigma(z) = 1/(1+e^{-z})$, which is used in this paper. White and Racine (2001) obtained the asymptotic distribution of the resulting estimators under the assumption of the true parameters being unique. In fact, the authors implicitly assumed that the number of hidden units r is known. However, even if we assume that we know the number of hidden units, it is difficult to establish the asymptotic properties for the parameter estimators. In section 6.1, we conducted a simulation based on a single-layer neural network with 2 hidden units. Even for such a simple model, the simulation result suggests that it is unlikely to obtain consistent estimators. Moreover, since the number of hidden units is usually unknown in practice, such an assumption can be easily violated. For example, as pointed out in Fukumizu (1996) and Fukumizu et al. (2003), if the true function is $f_0(x) = \alpha\sigma(\gamma x)$, (i.e., the true number of hidden units is 1), and we fit the model using a neural network with two hidden units, then any parameter $\boldsymbol{\theta} = [\alpha_0, \alpha_1, \ldots, \alpha_r, \gamma_{0,1}, \ldots, \gamma_{0,r}, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_r^T]^T$ in the high-dimensional set

$$\{\boldsymbol{\theta} : \gamma_1 = \gamma, \alpha_1 = \alpha, \gamma_{0,1} = \gamma_{0,2} = \alpha_2 = \alpha_0 = 0\} \cup$$
$$\{\boldsymbol{\theta} : \gamma_1 = \gamma_2 = \gamma, \gamma_{0,1} = \gamma_{0,2} = \alpha_0 = 0, \alpha_1 + \alpha_2 = \alpha\}$$

realizes the true function $f_0(x)$. Therefore, when the number of hidden units is unknown, the parameters in this parametric nonlinear regression problem are unidentifiable. Theorem 1 in Wu (1981) showed that a necessary condition for the weak consistency of nonlinear least square estimators is

$$\sum_{i=1}^{n} [f(\boldsymbol{x}_i, \boldsymbol{\theta}) - f(\boldsymbol{x}_i, \boldsymbol{\theta}')]^2 \to \infty, \text{ as } n \to \infty,$$

for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ in the parameter space as long as the error distribution has finite Fisher information. Such a condition implies that when the parameters are not identifiable, the resulting nonlinear least squares estimators will be inconsistent, which hinders further explorations on the asymptotic properties for the neural network estimators. Liu and Shao (2003) and Zhu and

Zhang (2006) proposed techniques to conduct hypothesis testing under loss of identifiability. However, their theoretical results are not easy to implement in the neural network setting.

Even though a function can have different neural network parametrizations, the function itself can be considered as unique. Moreover, due to the Universal Approximation Theorem (Hornik, Stinchcombe and White, 1989), any continuous function on compact support can be approximated arbitrarily well by a neural network with one hidden layer. So it seems natural to consider it as a nonparametric regression problem and approximate the underlying function class through a class of neural networks with one hidden layer. Specifically, suppose that the true nonparametric regression model is

$$y_i = f_0(\boldsymbol{x}_i) + \epsilon_i,$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. random variables defined on a complete probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{E}[\epsilon] = 0$, $\operatorname{Var}[\epsilon] = \sigma^2 < \infty$; $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathcal{X} \subset \mathbb{R}^d$ are vectors of covariates with $\mathcal{X}$ being a compact set in $\mathbb{R}^d$ and f_0 is an unknown function needed to be estimated. We assume that $f_0 \in \mathcal{F}$, where $\mathcal{F}$ is the class of continuous functions with compact supports. Clearly, f_0 minimizes the population criterion function

$$\begin{aligned} Q_n(f) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 + \sigma^2. \end{aligned}$$

A least squares estimator of the regression function can be obtained by minimizing the empirical squared error loss $\mathbb{Q}_n(f)$:

$$\hat{f}_n = \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{Q}_n(f) = \operatorname{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2.$$ ✓

However, if the class of functions $\mathcal{F}$ is too rich, the resulting least squares estimator may have undesired properties, such as inconsistency (van de Geer, 2000; Shen and Wong, 1994; Shen, 1997). Instead, we can optimize the squared error loss over some less complex function space $\mathcal{F}_n$, which is an approximation of $\mathcal{F}$ while the approximation error tends to 0 as the sample size increases. In the language of Grenander (1981), such a sequence of function classes is known as a sieve. More precisely, we consider a sequence of function classes,

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \cdots \subseteq \mathcal{F},$$

n, $\mathcal{F}_n \subseteq \mathcal{F}$

approximating $\mathcal{F}$ in the sense that $\bigcup_{n=1}^{\infty}\mathcal{F}_n$ is dense in $\mathcal{F}$. In other words, for each $f \in \mathcal{F}$, there exists $\pi_n f \in \mathcal{F}_n$ such that $d(f, \pi_n f) \to 0$ as $n \to \infty$, where $d(\cdot,\cdot)$ is some pseudo-metric defined on $\mathcal{F}$. With some abuse of notation, an approximate sieve estimator $\hat{f}_n$ is defined to be

$\mathbb{Q}_n(\hat{f}_n)$

in $\mathbb{Q}_n(f)+\mathcal{O}_p(\eta_n)$

inf $f \in \mathcal{F}$ $f\in\mathcal{F}_{r_n}$

$$\mathbb{Q}_n(\hat{f}_n) \le \inf_{f\in\mathcal{F}_n} \mathbb{Q}_n(f) + \mathcal{O}_p(\eta_n), \tag{1.1}$$

$f\in\mathcal{F}_{r_n}$ $f\in\mathcal{F}$

where $\eta_n \to 0$ as $n \to \infty$.

$r_n \uparrow$

$\mathcal{F}_{r_n}$

Throughout the rest of the paper, we focus on the sieve of neural networks with one hidden layer and sigmoid activation function. Specifically, we let

width $: r_n$

$$\mathcal{F}_{r_n} = \Bigg\{ \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma\left(\boldsymbol{\gamma}_j^T \boldsymbol{x} + \gamma_{0,j}\right) : \boldsymbol{\gamma}_j \in \mathbb{R}^d, \alpha_j, \gamma_{0,j} \in \mathbb{R},$$

$$\sum_{j=0}^{r_n} |\alpha_j| \le V_n \text{ for some } V_n > 4 \text{ and } \max_{1\le j\le r_n} \sum_{i=0}^{d} |\gamma_{i,j}| \le M_n \text{ for some } M_n > 0 \Bigg\}, \tag{1.2}$$

$n\to\infty$

$\bigcup \mathcal{F}_{r_n} \Rightarrow \mathcal{F}$

$\forall \mathcal{F}_{r_n}$

where $r_n, V_n, M_n \uparrow \infty$ as $n \to \infty$. Such method has been discussed in previous literatures (e.g. White (1989) and White (1990)). In those papers, consistency of the neural network sieve estimators has been established under random designs. However, there are few results on the asymptotic distribution of the neural network sieve estimators, which will be established in this paper. Throughout this paper, we focus on the fixed design. Hornik, Stinchcombe and White (1989) showed that $\bigcup_n \mathcal{F}_{r_n}$ is dense in $\mathcal{F}$ under the sup-norm. But when considering the asymptotic properties of the sieve estimators, we use the pseudo-norm $\|f\|_n^2 = n^{-1}\sum_{i=1}^n f^2(\boldsymbol{x}_i)$ (see Proposition 7.1 in the Appendix) defined on $\mathcal{F}$ and $\mathcal{F}_{r_n}$.

In section 2, we discuss the existence of neural network sieve estimators. The weak consistency and rate of convergence of the neural network sieve estimators will be established in section 3 and section 4, respectively. Section 5 focuses on the asymptotic distribution of the neural network sieve estimators. Simulation results are presented in section 6.

Notations: Throughout the rest of the paper, bold font alphabetic letters and Greek letters are vectors. $C(\mathcal{X})$ is the set of continuous functions defined on $\mathcal{X}$. The symbol $\lesssim$ means "bounded above up to a universal constant" and $a_n \sim b_n$ means $\frac{a_n}{b_n} \to 1$ as $n \to \infty$. For a pseudo-metric space (T, d), $N(\epsilon, T, d)$ is its covering number, which is the minimum number of ϵ-balls needed to cover T. Its natural logarithm is the entropy number and is denoted by $H(\epsilon, T, d)$.

$O_p(1)$

2. Existence. A natural question to ask is whether the sieve estimator based on neural networks exists. Before addressing this question, we first study some properties of $\mathcal{F}_{r_n}$. Proposition 2.1 shows that the sigmoid function is a Lipschitz function with Lipschitz constant $L = 1/4$.

PROPOSITION 2.1. *A sigmoid function $\sigma(z) = e^z/(1+e^z)$ is a Lipschitz function on $\mathbb{R}$ with Lipschitz constant $1/4$.*

PROOF. For all $z_1, z_2 \in \mathbb{R}$, $\sigma(z)$ is continuous on $[z_1, z_2]$ and is differentiable on (z_1, z_2). Note that

$$\sigma'(z) = \sigma(z)(1-\sigma(z)) \leq \frac{1}{4} \quad \forall z \in \mathbb{R}.$$

By using the Mean Value Theorem, we know that

$$\sigma(z_1) - \sigma(z_2) = \sigma'(\lambda z_1 + (1-\lambda)z_2)(z_1 - z_2),$$

for some $\lambda \in [0,1]$. Hence

$$|\sigma(z_1) - \sigma(z_2)| = |\sigma'(\lambda z_1 + (1-\lambda)z_2)||z_1 - z_2| \leq \frac{1}{4}|z_1 - z_2|,$$

which means that $\sigma(z)$ is a Lipschitz function on $\mathbb{R}$ with Lipschitz constant $1/4$. □

The second proposition provides an upper bound for the envelope function $\sup_{f\in\mathcal{F}_{r_n}} |f|$.

PROPOSITION 2.2. *For each fixed n,*

$$\sup_{f\in\mathcal{F}_{r_n}} \|f\|_\infty \leq V_n.$$

PROOF. For any $f \in \mathcal{F}_{r_n}$ with a fixed n and all $\boldsymbol{x} \in \mathcal{X}$, we have

$$\begin{aligned} |f(\boldsymbol{x})| &= \left| \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma\left(\boldsymbol{\gamma}_j^T \boldsymbol{x} + \gamma_{0,j}\right) \right| \\ &\leq |\alpha_0| + \sum_{j=1}^{r_n} |\alpha_j| \sigma\left(\boldsymbol{\gamma}_j^T \boldsymbol{x} + \gamma_{0,j}\right) \leq \sum_{j=0}^{r_n} |\alpha_j| \leq V_n. \end{aligned}$$

Since the right hand side does not depend on $\boldsymbol{x}$ and f, we get

$$\sup_{f\in\mathcal{F}_{r_n}} \|f\|_\infty = \sup_{f\in\mathcal{F}_{r_n}} \sup_{\boldsymbol{x}\in\mathcal{X}} |f(\boldsymbol{x})| \leq V_n.$$

$V_n = V$

□

$V_n \Rightarrow$ [illegible]

Now we quote a general result from White and Wooldridge (1991). The theorem tells us that under some mild conditions, there exists a sieve approximate estimator and such an estimator is also measurable.

① Def 1 is mea

THEOREM 2.1 (Theorem 2.2 in White and Wooldridge (1991)). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a complete probability space and let (Θ, ρ) be a pseudo-metric space. Let $\{\Theta_n\}$ be a sequence of compact subsets of Θ. Let $\mathbb{Q}_n : \Omega \times \Theta_n \to \bar{\mathbb{R}}$ be $\mathcal{A} \otimes \mathcal{B}(\Theta_n)/\mathcal{B}(\bar{\mathbb{R}})$-measurable, and suppose that for each $\omega \in \Omega$, $\mathbb{Q}_n(\omega, \cdot)$ is lower semicontinuous on Θ_n, $n = 1, 2, \ldots$. Then for each $n = 1, 2, \ldots$, there exists $\hat{\theta}_n : \Omega \to \Theta_n$, $\mathcal{A}/\mathcal{B}(\Theta_n)$-measurable such that for each $\omega \in \Omega$, $\mathbb{Q}_n(\omega, \hat{\theta}_n(\omega)) = \inf_{\theta \in \Theta_n} \mathbb{Q}_n(\omega, \theta)$.*

②

③ $F_{\theta n}$ is compact set

Note that

$$
\begin{aligned}
\mathbb{Q}_n(f) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 \quad \cdot\ w(x_i) \qquad \text{bonded support} \\
&= \frac{1}{n}\sum_{i=1}^{n}(f_0(\boldsymbol{x}_i) + \epsilon_i - f(\boldsymbol{x}_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 - 2\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2.
\end{aligned}
$$

Since the randomness only comes from ϵ_i's, it is clear that $\mathbb{Q}_n$ is a measurable function and for a fixed ω, $\mathbb{Q}_n$ is continuous in f. Therefore, to show the existence of the sieve estimator, it suffices to show that $\mathcal{F}_{r_n}$ is compact in $C(\mathcal{X})$, which is proved in the following lemma.

LEMMA 2.1. *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Then for each fixed n, $\mathcal{F}_{r_n}$ is a compact set.*

$F_{rn} \subset F$

PROOF. For each fixed n, let $\boldsymbol{\theta}_n = [\alpha_0, \ldots, \alpha_{r_n}, \boldsymbol{\gamma}_{0,1}, \ldots, \gamma_{0,r_n}, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_{r_n}^T]^T$ belong to $[-V_n, V_n]^{r_n+1} \times [-M_n, M_n]^{r_n(d+1)} := \Theta_n$. If n is fixed, Θ_n is a bounded closed set and hence it is a compact set in $\mathbb{R}^{r_n(d+2)+1}$. Consider a map

$$H : (\Theta_n, \|\cdot\|_2) \to (\mathcal{F}_{r_n}, \|\cdot\|_n)$$

①

$$\boldsymbol{\theta}_n \mapsto H(\boldsymbol{\theta}_n) = \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma\left(\boldsymbol{\gamma}_j^T \boldsymbol{x} + \gamma_{0,j}\right)$$

Note that $\mathcal{F}_{r_n} = H(\Theta_n)$. Therefore, to show that $\mathcal{F}_{r_n}$ is a compact set, it suffices to show that H is a continuous map due to the compactness of Θ_n.

$\mathcal{F}_{\theta_n}$ is compact.

$\|H(\theta_{1,n}) - H(\theta_{2,n})\|_n^2 < C\|\theta_{1,n} - \theta_{2,n}\|_2$

Let $\boldsymbol{\theta}_{1,n}, \boldsymbol{\theta}_{2,n} \in \Theta_n$, then

$$
\begin{aligned}
&\|H(\boldsymbol{\theta}_{1,n}) - H(\boldsymbol{\theta}_{2,n})\|_n^2 \\
=&\frac{1}{n}\sum_{i=1}^{n}\left[\alpha_0^{(1)} + \sum_{j=1}^{r_n}\alpha_j^{(1)}\sigma\left(\boldsymbol{\gamma}_j^{(1)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(1)}\right) - \alpha_0^{(2)} - \sum_{j=1}^{r_n}\alpha_j^{(2)}\sigma\left(\boldsymbol{\gamma}_j^{(2)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(2)}\right)\right]^2 \\
\le&\frac{1}{n}\sum_{i=1}^{n}\left[\left|\alpha_0^{(1)} - \alpha_0^{(2)}\right| + \sum_{j=1}^{r_n}\left|\alpha_j^{(1)}\sigma\left(\boldsymbol{\gamma}_j^{(1)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(1)}\right) - \alpha_j^{(2)}\sigma\left(\boldsymbol{\gamma}_j^{(2)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(2)}\right)\right|\right]^2 \\
=&\frac{1}{n}\sum_{i=1}^{n}\left[\left|\alpha_0^{(1)} - \alpha_0^{(2)}\right| + \sum_{j=1}^{r_n}|\alpha_j^{(1)}|\left|\sigma\left(\boldsymbol{\gamma}_j^{(1)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(1)}\right) - \sigma\left(\boldsymbol{\gamma}_j^{(2)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(2)}\right)\right| + \right. \\
&\left.\qquad |\alpha_j^{(1)} - \alpha_j^{(2)}|\sigma\left(\boldsymbol{\gamma}_j^{(2)^T}\boldsymbol{x}_i + \gamma_{0,j}^{(2)}\right)\right]^2 \\
\le&\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=0}^{r_n}|\alpha_j^{(1)} - \alpha_j^{(2)}| + \frac{V_n}{4}\sum_{j=1}^{r_n}\left|\left(\boldsymbol{\gamma}_j^{(1)} - \boldsymbol{\gamma}_j^{(2)}\right)^T\boldsymbol{x}_i\right| + \left|\gamma_{0,j}^{(1)} - \gamma_{0,j}^{(2)}\right|\right]^2 \\
\le&\left[\sum_{j=0}^{r_n}|\alpha_j^{(1)} - \alpha_j^{(2)}| + \frac{V_n}{4}(1 \vee \|\boldsymbol{x}\|_\infty)\sum_{j=1}^{r_n}\left\|\boldsymbol{\gamma}_j^{(1)} - \boldsymbol{\gamma}_j^{(2)}\right\|_1 + \left|\gamma_{0,j}^{(1)} - \gamma_{0,j}^{(2)}\right|\right]^2 \\
\le&\left(\frac{V_n}{4}(1 \vee \|\boldsymbol{x}\|_\infty)\right)^2[r_n(d+1)]\|\boldsymbol{\theta}_{1,n} - \boldsymbol{\theta}_{2,n}\|_2^2.
\end{aligned}
$$

$\sum_{j=1}^{r_n}|\alpha_j^{(1)}| \le V_n$

$(0,1)$

$a\ f(x)$

$C\sqrt{d}$

$\sup_x f(x) \le V_n.$

Hence, for any $\epsilon > 0$, we choose $\delta = \epsilon/\left(\frac{V_n}{4}(1 \vee \|\boldsymbol{x}\|_\infty)\sqrt{r_n(d+1)}\right)$,. When $\|\boldsymbol{\theta}_{1,n} - \boldsymbol{\theta}_{2,n}\|_2 < \delta$, we have

$$\|H(\boldsymbol{\theta}_{1,n}) - H(\boldsymbol{\theta}_{2,n})\|_n < \epsilon,$$

which implies that H is a continuous map and hence $\mathcal{F}_{r_n}$ is a compact set for each fixed n. □

As a corollary of Lemma 2.1 and Theorem 2.1, we can easily obtain the existence of sieve estimator.

Corollary 2.1. *Based on the notations above, for each $n = 1, 2, \ldots$, there exists $\hat{f}_n : \Omega \to \mathcal{F}_{r_n}$, $\mathcal{A}/\mathcal{B}(\mathcal{F}_{r_n})$-measurable such that $\mathbb{Q}_n(\hat{f}_n(\omega)) = \inf_{f \in \mathcal{F}_{r_n}} \mathbb{Q}_n(f)$.*

3. Consistency.

In this section, we are going to show the consistency of the neural network sieve estimator. The consistency result leans heavily on the following Uniform Law of Large Numbers. We start by considering a simple case with $V_n \equiv V$ for all n. In such a case, $\bigcup_n \mathcal{F}_{r_n}$ is not dense in $\mathcal{F}$ but rather in a subset of $\mathcal{F}$ with functions satisfying a certain smoothness condition.

LEMMA 3.1. *Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. sub-Gaussian random variables with sub-Gaussian parameter σ_0. If $[r_n(d+2)+1]\log[r_n(d+2)+1] = o(n)$, we have*

$$\sup_{f \in \mathcal{F}_{r_n}} |\mathbb{Q}_n(f) - Q_n(f)| \xrightarrow{p^*} 0.$$

$\mathbb{Q}_n f$

$E\,\mathbb{Q}_n(f)$

PROOF. For any $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}} |\mathbb{Q}_n(f) - Q_n(f)| > \delta\right)\\
=&\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2 - 2\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right)\right| > \delta\right)\\
\le&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2\right| > \frac{\delta}{2}\right) + \mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{\delta}{4}\right)\\
:=&(I) + (II).
\end{aligned}$$

For (I), based on the Weak Law of Large Numbers, we know that there exists $N_1 > 0$ such that for all $n \ge N_1$ we have

$$(I) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2\right| > \frac{\delta}{2}\right) < \frac{\delta}{2}.$$

Now, we are going to evaluate (II). From the sub-Gaussianity of $\epsilon_1, \dots, \epsilon_n$, we know that $\epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))$ is also sub-Gaussian with mean 0 and sub-Gaussian parameter $\sigma_0|f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)|$. Hence, by using the Hoeffding inequality,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{\delta}{4}\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{n\delta}{4}\right)\\
&\le 2\exp\left\{-\frac{n^2\delta^2}{32\sigma_0^2\sum_{i=1}^n (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2}\right\}.
\end{aligned}$$

$P(|\epsilon_i|>t) \le$

$2\exp(-\frac{t^2}{\sigma_0^2})$

From Proposition 2.2, we know that $\sup_{f\in\mathcal{F}_{r_n}} \|f\|_n \le V$. Hence, based on Corollary 8.3 in van de Geer (2000), (II) will have an exponential bound if

$(\frac{nt}{4})^2$

$\le 2\exp(- \frac{\quad}{\sigma_0^2 \sum_{i=1}^n (f(x_i) - f_0)^2}$

there exists some constant C, $\delta > 0$ and $\sigma > 0$ satisfying $V > \delta/\sigma$ and

① ∃ exist

$$\sqrt{n}\delta \geq 2C\left(\int_{\delta/(8\sigma)}^{V} H^{1/2}(u, \mathcal{F}_{r_n}, \|\cdot\|_n)du \vee V\right). \tag{3.1}$$

$\epsilon \leq 2b$

Now, we are going to show that (3.1) holds in our case. It follows from Theorem 14.5 in Anthony and Bartlett (2009), which gives an upper bound of the covering number for $\mathcal{F}_{r_n}$,

$d > 2$

$N_\infty(\epsilon, \mathcal{F}_d, m)$ 2

$$N(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_\infty) \leq \left(\frac{4e[r_n(d+2)+1]\left(\frac{1}{4}V\right)^2}{\epsilon\left(\frac{1}{4}V - 1\right)}\right)^{r_n(d+2)+1} := \tilde{A}_{r_n,d,V}\,\epsilon^{-[r_n(d+2)+1]},$$

where $\tilde{A}_{r_n,d,V} = (e[r_n(d+2)+1]V^2/(V-4))^{r_n(d+2)+1}$. By letting

$$\left[\frac{4emb\,W(LV)^L}{\epsilon(LV-1)}\right]^W$$

$$\begin{aligned}
A_{r_n,d,V} &= \log \tilde{A}_{r_n,d,V} - [r_n(d+2)+1] \\
&= [r_n(d+2)+1]\left(\log \frac{e[r_n(d+2)+1]V^2}{V-4} - 1\right) \\
&= [r_n(d+2)+1]\log \frac{[r_n(d+2)+1]V^2}{V-4},
\end{aligned}$$

and noting that $V^2 - eV + 4e \geq 0$ for all V, we have $\log \frac{[r_n(d+2)+1]V^2}{V-4} \geq \log \frac{V^2}{V-4} \geq \log \frac{e(V-4)}{V-4} = 1$. Then,

$$\begin{aligned}
H(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_\infty) &= \log N(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_\infty) \\
&= \log \tilde{A}_{r_n,d,V} + [r_n(d+2)+1]\log \frac{1}{\epsilon} \\
&\leq A_{r_n,d,V} + [r_n(d+2)+1]\frac{1}{\epsilon} \quad (\text{since } \log x \leq x - 1 \text{ for all } x > 0) \\
&\leq A_{r_n,d,V}\left(1 + \frac{1}{\epsilon}\right).
\end{aligned}$$

Note that

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^{n} f^2(\boldsymbol{x}_i) \leq \left(\sup_{\boldsymbol{x}} |f(\boldsymbol{x})|\right)^2 = \|f\|_\infty^2,$$

we have $H(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_n) \leq H(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_\infty)$. Then

$$\begin{aligned}\int_{\delta/(8\sigma)}^{V} H^{1/2}(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_n)\mathrm{d}\epsilon &\leq A_{r_n,d,V}^{1/2} \int_0^V \left(1+\frac{1}{\epsilon}\right)^{1/2} \mathrm{d}\epsilon \\ &= A_{r_n,d,V}^{1/2} \left[\int_0^1 \left(1+\frac{1}{\epsilon}\right)^{1/2} \mathrm{d}\epsilon + \int_1^V \left(1+\frac{1}{\epsilon}\right)^{1/2} \mathrm{d}\epsilon\right] \\ &\leq A_{r_n,d,V}^{1/2} \left[\sqrt{2}\int_0^1 \epsilon^{-\frac{1}{2}}\mathrm{d}\epsilon + \sqrt{2}(V-1)\right] \\ &\leq A_{r_n,d,V}^{1/2} \left[2\sqrt{2} + 2\sqrt{2}(V-1)\right] \\ &= 2\sqrt{2}A_{r_n,d,V}^{1/2} V.\end{aligned}$$

Clearly, $2\sqrt{2}A_{r_n,d,V}^{1/2}V \geq V$. Under the assumption of $[r_n(d+2)+1]\log[r_n(d+2)+1] = o(n)$, for any $\delta > 0$, there exists $N_2 > 0$ such that for all $n \geq N_2$,

$$4\sqrt{2}V\left(\frac{1}{n}A_{r_n,d,V}\right)^{1/2} < \frac{\delta}{4},$$

i.e. (3.1) holds with $C = 1$ and $n \geq N_2$. Hence, based on Corollary 8.3 in van de Geer (2000), for $n \geq N_2$,

$$\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{\delta}{4} \wedge \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \leq \sigma^2\right) \leq \exp\left\{-\frac{n\delta^2}{64V^2}\right\}. \tag{3.2}$$

Since $\int_0^V H^{1/2}(\epsilon, \mathcal{F}_{r_n}, \|\cdot\|_n)\mathrm{d}\epsilon < \infty$, we can take $\sigma \to \infty$ in (3.2) to get

$$\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{\delta}{4}\right) \leq \exp\left\{-\frac{n\delta^2}{64V^2}\right\}.$$

Let $N_3 = \frac{64V^2}{\delta^2}\log\frac{2}{\delta}$, then for $n \geq \max\{N_2, N_3\}$, we have

$$(II) = \mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{\delta}{4}\right) \leq \frac{\delta}{2}.$$

Thus, we conclude that for any $\delta > 0$, by taking $n \geq \max\{N_1, N_2, N_3\}$, we have

$$\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_{r_n}} |\mathbb{Q}_n(f) - Q_n(f)| > \delta\right) < \delta,$$

which proves the desired result. □

Remark 3.1. *Lemma 3.1 shows that if we have a fixed number of features, the desired Uniform Law of Large Numbers holds when the number of hidden units in the neural network sieve does not grow too fast.*

Now, we are going to extend the result to a more general case. In Lemma 3.1, we assume that the errors $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. sub-Gaussian and $V_n \equiv V$. In the following lemma, we are going to relax both restrictions.

Lemma 3.2. *Under the assumption of*

$$[r_n(d+2)+1]V_n^2 \log(V_n[r_n(d+2)+1] = o(n), \text{ as } n \to \infty,$$

we have

$$\sup_{f \in \mathcal{F}_{r_n}} |\mathbb{Q}_n(f) - Q_n(f)| \xrightarrow{p^*} 0, \text{ as } n \to \infty.$$

Proof. As in the proof of Lemma 3.1, it suffices to show that

$$\mathbb{P}^*\left(\sup_{f \in \mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right| > \frac{\delta}{4}\right) \to 0, \text{ as } n \to \infty. \tag{3.3}$$

By using the Markov's inequality, (3.3) holds if we can show

$$\mathbb{E}^*\left[\sup_{f \in \mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] \to 0, \text{ as } n \to \infty.$$

Note that $\mathbb{E}[\epsilon] = 0$ and each $f \in \mathcal{F}_{r_n}$ has its corresponding parametrization $\boldsymbol{\theta}_n$. Since $\boldsymbol{\theta}_n$ is in a compact set, there exists a sequence $\boldsymbol{\theta}_{n,k} \to \boldsymbol{\theta}_n$ as $k \to \infty$ with $\boldsymbol{\theta}_{n,k} \in \mathbb{Q}^{r_n(d+2)+1} \cap ([-V_n, V_n]^{r_n+1} \times [-M_n, M_n]^{r_n(d+1)})$. Each $\boldsymbol{\theta}_{n,k}$ corresponds to a function $f_k \in \mathcal{F}_{r_n}$. Based on continuity, we have $f_k(\boldsymbol{x}) \to f(\boldsymbol{x})$ for each $\boldsymbol{x} \in \mathcal{X}$. From Example 2.3.4 in van der Vaart and Wellner (1996), we know that $\mathcal{F}_{r_n}$ is P-measurable. Based on symmetrization inequality, we have

$$\begin{aligned}
&\mathbb{E}^*\left[\sup_{f \in \mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] \\
\leq &2\mathbb{E}_\epsilon \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \xi_i \epsilon_i \left(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right)\right|\right],
\end{aligned}$$

where $\xi_1, \ldots, \xi_n$ are i.i.d. Rademacher random variables independent of $\epsilon_1, \ldots, \epsilon_n$. Based on the Strong Law of Large Numbers, there exists $N_1 > 0$, such that for all $n \geq N_1$,

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 < \sigma^2 + 1, \text{ a.s.}$$

$\{X_t\}$ - sub-G

for $\delta > 0$

$E \sup |X_s - X_t| \leq$

$K \cdot \int_0^\delta \sqrt{\log D(\varepsilon, d)}\, d\varepsilon$

$E \sup |X_t| \leq E|X_{t_0}| + K\int_0^\infty \sqrt{\quad}\, d\varepsilon$

For fixed $\epsilon_1, \ldots, \epsilon_n$, $\sum_{i=1}^n \xi_i\epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))$ is a sub-Gaussian process indexed by $f \in \mathcal{F}_{r_n}$. Suppose that $(\Xi, \mathcal{C}, \mu)$ is the probability space on which $\xi_1, \ldots, \xi_n$ are defined and let $Y(f, \omega) = \sum_{i=1}^n \xi_i(\omega)\epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))$ with $f \in \mathcal{F}_{r_n}$ and $\omega \in \Xi$. As we have shown above, we have $f_k \to f$ and by continuity, $Y(f_k, \omega) \to Y(f, \omega)$ for any $\omega \in \Xi$. This shows that $\{Y(f, \omega), f \in \mathcal{F}_{r_n}\}$ is a separable sub-Gaussian process. Hence Corollary 2.2.8 in van der Vaart and Wellner (1996) implies that there exists a universal constant K and for any $f_n^* \in \mathcal{F}_{r_n}$ with $n \geq N_1$,

$$\begin{aligned}
&\mathbb{E}_\xi\left[\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{n}\sum_{i=1}^n \xi_i\epsilon_i(f(\boldsymbol{x}_i)) - f_0(\boldsymbol{x}_i))\right|\right] \\
&= \mathbb{E}_\xi\left[\frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i\epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] \\
&\leq \mathbb{E}_\xi\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\epsilon_i(f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] + K\int_0^\infty \sqrt{\frac{\log N\left(\frac{1}{2}\eta, \mathcal{F}_{r_n}, d\right)}{n}}\mathrm{d}\eta \\
&= \mathbb{E}_\xi\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\epsilon_i(f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] + K\int_0^{2V_n} \sqrt{\frac{\log N\left(\frac{1}{2}\eta, \mathcal{F}_{r_n}, d\right)}{n}}\mathrm{d}\eta \\
&\leq \mathbb{E}_\xi\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\epsilon_i(f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right|\right] + K\int_0^{2V_n} \sqrt{\frac{\log N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right)}{n}}\mathrm{d}\eta,
\end{aligned}$$

where the second equality follows from Proposition 2.2. For $f, g \in \mathcal{F}_{r_n}$,

$$\begin{aligned}
d(f,g) &= \left[\sum_{i=1}^n\left(\frac{1}{\sqrt{n}}\epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) - \frac{1}{\sqrt{n}}\epsilon_i(g(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))\right)^2\right]^{1/2} \\
&= \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2(f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i))^2\right)^{1/2}
\end{aligned}$$

so that the last inequality follows by noting that

$$d(f,g) \leq \|f - g\|_\infty \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2\right)^{1/2}.$$

We then evaluate these two terms. For the first term, for $n \geq N_1$, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}_\xi\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\epsilon_i(f_n^*(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i))\right|\right] &\leq \frac{1}{n}\sum_{i=1}^n |\epsilon_i||f_n^*(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)| \\
&\leq \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n (f_n^*(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i))^2\right)^{1/2} \\
&\leq \sqrt{\sigma^2+1}\sup_{\boldsymbol{x}\in\mathcal{X}}|f_n^*(\boldsymbol{x})-f_0(\boldsymbol{x})|, \text{ a.s.}
\end{aligned}$$

By choosing $f_n^* = \pi_{r_n}f_0$ and using the universal approximation theorem introduced by Hornik, Stinchcombe and White (1989), we know that $\sup_{x\in\mathcal{X}}|f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)| \to 0$ as $n \to \infty$. Therefore, for any $\zeta > 0$, there exists $N_2 > 0$, such that for all $n \geq N_2$,

$$\sup_{x\in\mathcal{X}}|f_n^*(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)| < \frac{\zeta}{\sqrt{\sigma^2+1}}.$$

By choosing $n \geq N_1 \vee N_2$, we get

$$\mathbb{E}_\xi\left[\left|\frac{1}{n}\sum_{i=1}^n \xi_i\epsilon_i(f_n^*(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i))\right|\right] < \zeta \text{ a.s.}$$

For the second term, we use the same bound from Theorem 14.5 in Anthony and Bartlett (2009) as we did in the proof of Lemma 3.1:

$$\begin{aligned}
N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right) &\leq \left(\frac{8\sqrt{\sigma^2+1}e[r_n(d+2)+1]\left(\frac{1}{4}V_n\right)^2}{\eta\left(\frac{1}{4}V_n-1\right)}\right)^{r_n(d+2)+1} \\
&:= \tilde{B}_{r_n,d,V_n}\eta^{-[r_n(d+2)+1]},
\end{aligned}$$

where $\tilde{B}_{r_n,d,V_n} = \left(2\sqrt{\sigma^2+1}e[r_n(d+2)+1]V_n^2/(V_n-4)\right)^{r_n(d+2)+1}$. Let

$$\begin{aligned}
B_{r_n,d,V_n} &= \log\tilde{B}_{r_n,d,V_n} - [r_n(d+2)+1] \\
&= [r_n(d+2)+1]\left(\log\frac{2\sqrt{\sigma^2+1}e[r_n(d+2)+1]V_n^2}{V_n-4} - 1\right) \\
&= [r_n(d+2)+1]\left(\log\frac{[r_n(d+2)+1]V_n^2}{V_n-4} + \log(2\sqrt{\sigma^2+1})\right) \\
&\leq 2[r_n(d+2)+1]\log\frac{[r_n(d+2)+1]V_n^2}{V_n-4}, \text{ for all } n \geq N_1 \vee N_3,
\end{aligned}$$

where N_3 is chosen to satisfy $r_n(d+2)+1 \geq 2\sqrt{\sigma^2+1}$. The last inequality then follows by noting that $V_n^2 - V_n + 4 \geq 0$ for all V_n so that $\log \frac{[r_n(d+1)+1]V_n^2}{V_n-4} \geq \log \frac{2\sqrt{\sigma^2+1}(V_n-4)}{V_n-4} = \log(2\sqrt{\sigma^2+1})$. We also have

$$\begin{aligned}
H\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right) &= \log N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right) \\
&= \log \tilde{B}_{r_n,d,V_n} + [r_n(d+2)+1]\log\frac{1}{\eta} \\
&\leq B_{r_n,d,V_n} + [r_n(d+2)+1]\frac{1}{\eta} \\
&\leq B_{r_n,d,V_n}\left(1+\frac{1}{\eta}\right),
\end{aligned}$$

and hence for all $n \geq N_1 \vee N_3$,

$$\begin{aligned}
&\int_0^{2V_n} H^{1/2}\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right) \mathrm{d}\eta \\
&\quad \leq B_{r_n,d,V_n}^{1/2}\int_0^{2V_n}\left(1+\frac{1}{\eta}\right)^{1/2}\mathrm{d}\eta \\
&\quad = B_{r_n,d,V_n}^{1/2}\left[\int_0^1\left(1+\frac{1}{\eta}\right)^{1/2}\mathrm{d}\eta + \int_1^{2V_n}\left(1+\frac{1}{\eta}\right)^{1/2}\mathrm{d}\eta\right] \\
&\quad \leq B_{r_n,d,V_n}^{1/2}\left[\sqrt{2}\int_0^1 \eta^{-1/2}\mathrm{d}\eta + \sqrt{2}(2V_n-1)\right] \\
&\quad \leq 4\sqrt{2}B_{r_n,d,V_n}^{1/2}V_n,
\end{aligned}$$

which implies that

$$\begin{aligned}
\int_0^{2V_n}\sqrt{\frac{H\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty\right)}{n}}\mathrm{d}\eta &\leq 4\sqrt{2}n^{-1/2}B_{r_n,d,V_n}^{1/2}V_n \\
&\sim 8\sqrt{\frac{[r_n(d+2)+1]V_n^2\log(V_n[r_n(d+2)+1])}{n}},
\end{aligned}$$

where the last part follows by noting that $\log\frac{V_n^2}{V_n-4} \sim \log V_n$. Under the assumption given in the Lemma, there exists $N_4 > 0$, such that for all $n \geq N_4$, we have

$$\sqrt{\frac{[r_n(d+2)+1]V_n^2\log(V_n[r_n(d+2)+1])}{n}} < \frac{\zeta}{8}.$$

Therefore, by choosing $n \geq N_1 \vee N_2 \vee N_3 \vee N_4$, we get

$$\mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}_{r_n}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \epsilon_i (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \right| \right] < 2\zeta \text{ a.s.},$$

i.e. $\mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}_{r_n}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \epsilon_i (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \right| \right] \to 0$ a.s.. Moreover, based on what we have shown, for a sufficiently large n, we have

$$\mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}_{r_n}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \epsilon_i (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \right| \right] \leq \sqrt{\sigma^2+1} \|\pi_{r_n} f_0 - f_0\|_\infty + \\ 4\sqrt{2} K B_{r_n,d,V_n}^{1/2} n^{-1/2} V_n \to 0, \text{ a s..}$$

Since $\mathbb{E}_\epsilon \left[\sqrt{\sigma^2+1} \|\pi_{r_n} f_0 - f_0\|_\infty + 4\sqrt{2} K B_{r_n,d,V_n}^{1/2} n^{-1/2} V_n \right] = \sqrt{\sigma^2+1} \|\pi_{r_n} f_0 - f_0\|_\infty + 4\sqrt{2} K B_{r_n,d,V_n}^{1/2} n^{-1/2} V_n \to 0 < \infty$, by using the Generalized Dominated Convergence Theorem, we know that

$$\begin{aligned} &\mathbb{E}^* \left[\sup_{f \in \mathcal{F}_{r_n}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \right| \right] \\ &\leq 2\mathbb{E}_\epsilon \mathbb{E}_\xi \left[\sup_{f \in \mathcal{F}_{r_n}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i \epsilon_i \left(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i) \right) \right| \right] \to 0, \end{aligned}$$

which completes the proof. □

Based on the above lemmas, we are ready to state the theorem on the consistency of neural network sieve estimators.

THEOREM 3.1. *Under the notation given above, if*

$$[r_n(d+2)+1] V_n^2 \log(V_n [r_n(d+2)+1] = o(n), \text{ as } n \to \infty, \tag{3.4}$$

then

$$\|\hat{f}_n - f_0\|_n \xrightarrow{p} 0.$$

PROOF. Since Q is continuous at $f_0 \in \mathcal{F}$ and $Q(f_0) = \sigma^2 < \infty$, for any $\epsilon > 0$, we have

$$\inf_{f: \|f - f_0\|_n \geq \epsilon} Q_n(f) - Q_n(f_0) = \inf_{f: \|f - f_0\|_n \geq \epsilon} \frac{1}{n} \sum_{i=1}^n (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 \geq \epsilon^2 > 0.$$

Hence, based on Lemma 2.1, Lemma 3.2 and Corollary 2.6 in White and Wooldridge (1991), we have

$$\|\hat{f}_n - f_0\|_n \xrightarrow{p} 0.$$

□

Remark 3.2. *We discuss the condition (3.4) in Theorem 3.1 via some simple examples here. If $\alpha_j = \mathcal{O}(1)$ for $j = 1, \dots, r_n$, then $V_n = \mathcal{O}(r_n)$ and*

$$[r_n(d+2)+1]V_n^2 \log(V_n[r_n(d+2)+1]) = \mathcal{O}(r_n^3 \log r_n).$$

Therefore, a possible growth rate for the number of hidden units in a neural network is $r_n = o\left((n/\log n)^{1/3}\right)$. On the other hand, if we have a slow growth rate for the number of hidden units in the neural network, such as $r_n = \log V_n$, then we have

$$[r_n(d+2)+1]V_n^2 \log(V_n[r_n(d+2)+1]) = \mathcal{O}((V_n \log V_n)^2).$$

Hence, a possible growth rate for the upper bound of the weights from the hidden layer to the output layer is $V_n = o\left(n^{1/2}/\log n\right)$.

4. Rate of Convergence.

To obtain the rate of convergence for neural network sieves, we apply Theorem 3.4.1 in van der Vaart and Wellner (1996)van der Vaart and Wellner (1996).

Lemma 4.1. *Let $f_n^* = \pi_{r_n} f_0 \in \mathcal{F}_{r_n}$. Given the above notations, for every n and $\delta > 8\|f_n^* - f_0\|_n$, we have*

$$\sup_{\frac{\delta}{2} < \|f - f_n^*\|_n \le \delta, f \in \mathcal{F}_{r_n}} Q_n(f_n^*) - Q_n(f) \lesssim -\delta^2.$$

Proof. Note that

$$\begin{aligned} Q_n(f_n^*) - Q_n(f) &= \frac{1}{n}\sum_{i=1}^n (f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 + \sigma^2 - \frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 - \sigma^2 \\ &= \|f_n^* - f_0\|_n^2 - \|f - f_0\|_n^2. \end{aligned}$$

In order to show the result, we first provide an upper bound for $Q_n(f_n^*) - Q_n(f)$ in terms of $\|f - f_n^*\|_n$. Due to the fact that $\|\cdot\|_n$ is a pseudo-norm, the triangle inequality gives

$$\begin{aligned} \|f - f_n^*\|_n &\le \|f - f_0\|_n + \|f_n^* - f_0\|_n \\ &= \|f - f_0\|_n - \|f_n^* - f_0\|_n + 2\|f_n^* - f_0\|_n. \end{aligned}$$

Therefore, we have

$$\|f - f_0\|_n - \|f_n^* - f_0\|_n \geq \|f - f_n^*\| - 2\|f_n^* - f_0\|_n,$$

so that for every f ~~satisfying~~ $\|f - f_n^*\|_n^2 \geq 16\|f_n^* - f_0\|_n^2$, i.e., $\|f - f_n^*\|_n \geq 4\|f_n^* - f_0\|_n$, we have

$$\|f - f_0\|_n - \|f_n^* - f_0\|_n \geq \|f - f_n^*\|_n - \frac{1}{2}\|f - f_n^*\|_n = \frac{1}{2}\|f - f_n^*\|_n \geq 0.$$

By squaring both sides, we obtain

$$\begin{aligned}\frac{1}{4}\|f - f_n^*\|_n^2 &\leq \|f - f_0\|_n^2 + \|f_n^* - f_0\|_n^2 - 2\|f - f_0\|_n\|f_n^* - f_0\|_n \\ &\leq \|f - f_0\|_n^2 + \|f_n^* - f_0\|_n^2 - 2\|f_n^* - f_0\|_n^2 \\ &= \|f - f_0\|_n^2 - \|f_n^* - f_0\|_n^2,\end{aligned}$$

and hence

$$\begin{aligned}\sup_{\frac{\delta}{2} < \|f - f_n^*\|_n \leq \delta, f \in \mathcal{F}_{r_n}} Q_n(f_n^*) - Q_n(f) &\leq \sup_{\|f - f_n^*\|_n > \frac{\delta}{2}, f \in \mathcal{F}_{r_n}} \|f_n^* - f_0\|_n^2 - \|f - f_0\|_n^2 \\ &\leq \sup_{\|f - f_n^*\|_n > \frac{\delta}{2}, f \in \mathcal{F}_{r_n}} \left(-\frac{1}{4}\|f - f_n^*\|_n^2\right) \\ &\lesssim -\delta^2.\end{aligned}$$

□

LEMMA 4.2. *For every sufficiently large n and $\delta > 8\|f_n^* - f_0\|_n$, we have*

$$\mathbb{E}^*\left[\sup_{\frac{\delta}{2} < \|f - f_n^*\|_n \leq \delta, f \in \mathcal{F}_{r_n}} \sqrt{n}\left[(\mathbb{Q}_n - Q_n)(f_n^*) - (\mathbb{Q}_n - Q_n)(f)\right]^+\right] \lesssim \int_0^\delta \sqrt{\log N(\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty)}d\eta$$

PROOF. Note that

$$\begin{aligned}(\mathbb{Q}_n - Q_n)(f_n^*) &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2 - \frac{2}{n}\sum_{i=1}^n \epsilon_i(f_n^*(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \\ (\mathbb{Q}_n - Q_n)(f_n^*) &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \sigma^2 - \frac{2}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)),\end{aligned}$$

we have

$$(\mathbb{Q}_n - Q_n)(f_n^*) - (\mathbb{Q}_n - Q_n)(f) = \frac{2}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_n^*(\boldsymbol{x}_i)).$$

Hence, by using the similar arguments as in the proof of Lemma 3.2 and applying Corollary 2.2.8 in van der Vaart and Wellner (1996)van der Vaart and Wellner (1996), we have

$$\begin{aligned}
&\mathbb{E}^*\left[\sup_{\frac{\delta}{2}<\|f-f_n^*\|_n\le\delta, f\in\mathcal{F}_{r_n}}\sqrt{n}\left[(\mathbb{Q}_n-Q_n)(f_n^*)-(\mathbb{Q}_n-Q_n)(f)\right]^+\right]\\
\le&\mathbb{E}^*\left[\sup_{\frac{\delta}{2}<\|f-f_n^*\|_n\le\delta, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\epsilon_i(f(\boldsymbol{x}_i)-f_n^*(\boldsymbol{x}_i))\right|\right]\\
\le&2\mathbb{E}_\epsilon\mathbb{E}_\xi\left[\sup_{\frac{\delta}{2}<\|f-f_n^*\|_n\le\delta, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i\epsilon_i\left(f(\boldsymbol{x}_i)-f_n^*(\boldsymbol{x}_i)\right)\right|\right]\\
\lesssim&\int_0^\delta\sqrt{\log N(\eta,\mathcal{F}_{r_n},\|\cdot\|_\infty)}\mathrm{d}\eta,
\end{aligned}$$

where the last inequality follows since $f_n^*\in\mathcal{F}_{r_n}$ for a large enough n. □

Now we are ready to apply Theorem 3.4.1 in van der Vaart and Wellner (1996) to obtain the rate of convergence for neural network sieve estimators.

THEOREM 4.1. *Based on the above notations, if*

$$\eta_n=\mathcal{O}\left(\min\{\|\pi_{r_n}f_0-f_0\|_n^2, r_n(d+2)\log(r_nV_n(d+2))/n, r_n(d+2)\log n/n\}\right),$$

then

$$\|\hat{f}_n-f_0\|_n=\mathcal{O}_p\left(\max\left\{\|\pi_{r_n}f_0-f_0\|_n,\sqrt{\frac{r_n(d+2)\log[r_nV_n(d+2)]}{n}},\sqrt{\frac{r_n(d+2)\log n}{n}}\right\}\right).$$

PROOF. Use the same bound from Theorem 14.5 in Anthony and Bartlett (2009), we have

$$\begin{aligned}
\log N(\eta,\mathcal{F}_{r_n},\|\cdot\|_n)&\le\log N(\eta,\mathcal{F}_{r_n},\|\cdot\|_\infty)\\
&\le\log\left(\frac{4e[r_n(d+2)+1]\left(\frac{1}{4}V_n\right)^2}{\eta\left(\frac{1}{4}V_n-1\right)}\right)^{r_n(d+2)+1}\\
&=[r_n(d+2)+1]\log\frac{\tilde{C}_{r_n,d,V_n}}{\eta},
\end{aligned}$$

where $\tilde{C}_{r_n,d,V_n} = \frac{e[r_n(d+2)+1]V_n^2}{V_n-4} > e$. Then from Lemma 3.8 in Mendelson (2003), for $\delta < 1$,

$$\begin{aligned}\int_0^\delta \sqrt{\log N(\eta, \mathcal{F}_{r_n}, \|\cdot\|_n)}\mathrm{d}\eta &\leq [r_n(d+2)+1]^{1/2} \int_0^\delta \sqrt{\log \frac{\tilde{C}_{r_n,d,V_n}}{\eta}}\mathrm{d}\eta \\ &\lesssim [r_n(d+2)+1]^{1/2}\delta\sqrt{\log \frac{\tilde{C}_{r_n,d,V_n}}{\delta}} \\ &:= \phi_n(\delta).\end{aligned}$$

Define $h : \delta \mapsto \phi_n(\delta)/\delta^\alpha = [r_n(d+2)+1]^{1/2}\delta^{1-\alpha}\sqrt{\log \frac{\tilde{C}_{r_n,d,V_n}}{\delta}}$. Since for $0 < \delta < 1$ and $1 < \alpha < 2$

$$\begin{aligned}h'(\delta) &= [r_n(d+2)+1]^{1/2}\left((1-\alpha)\delta^{-\alpha}\sqrt{\log \frac{\tilde{C}_{r_n,d,V_n}}{\delta}} - \frac{1}{2}\frac{\delta^2}{\tilde{C}_{r_n,d,V_n}}\frac{\tilde{C}_{r_n,d,V_n}}{\delta^2}\log^{-1/2}\frac{\tilde{C}_{r_n,d,V_n}}{\delta}\right) \\ &= [r_n(d+2)+1]^{1/2}\left((1-\alpha)\delta^{-\alpha}\sqrt{\log \frac{\tilde{C}_{r_n,d,V_n}}{\delta}} - \frac{1}{2}\log^{-1/2}\frac{\tilde{C}_{r_n,d,V_n}}{\delta}\right) \\ &< 0,\end{aligned}$$

$\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing on $(0, \infty)$. Let $\rho_n \lesssim \|\pi_{r_n}f_0 - f_0\|_n^{-1}$. Note that

$$\begin{aligned}\rho_n^2\phi_n\left(\frac{1}{\rho_n}\right) &= \rho_n[r_n(d+2)+1]^{1/2}\log^{1/2}\left(\rho_n\tilde{C}_{r_n,d,V_n}\right) \\ &= [r_n(d+2)+1]^{1/2}\rho_n\sqrt{\log\rho_n + \log\tilde{C}_{r_n,d,V_n}}\end{aligned}$$

and

$$\begin{aligned}\log\tilde{C}_{r_n,d,V_n} &= 1 + \log\frac{[r_n(d+2)+1]V_n^2}{V_n-4} \lesssim \log\frac{[r_n(d+2)+1]V_n^2}{V_n-4} \\ &\sim \log[r_nV_n(d+2)],\end{aligned}$$

we have

$$\rho_n^2\phi_n\left(\frac{1}{\rho_n}\right) \lesssim \sqrt{n} \Leftrightarrow r_n(d+2)\rho_n^2\left(\log\rho_n + \log[r_nV_n(d+2)]\right) \lesssim n.$$

Therefore, for

$$\rho_n \lesssim \min\left\{\left(\frac{n}{r_n(d+2)\log[r_nV_n(d+2)]}\right)^{1/2}, \left(\frac{n}{r_n(d+2)\log n}\right)^{1/2}\right\},$$

we have $\rho_n^2\phi_n\left(\frac{1}{\rho_n}\right) \lesssim \sqrt{n}$. Based on these observation, Lemma 4.1, Lemma 4.2 and Theorem 3.4.1 in van der Vaart and Wellner (1996)van der Vaart and Wellner (1996) imply that

$$\|\hat{f}_n - \pi_{r_n} f_0\|_n = \mathcal{O}_p\left(\max\left\{\|\pi_{r_n} f_0 - f_0\|_n, \sqrt{\frac{r_n(d+2)\log[r_n V_n(d+2)]}{n}}, \sqrt{\frac{r_n(d+2)\log n}{n}}\right\}\right).$$

By using the triangle inequality, we can further get

$$\begin{aligned}\|\hat{f}_n - f_0\|_n &\leq \|\hat{f}_n - \pi_{r_n} f_0\|_n + \|\pi_{r_n} f_0 - f_0\|_n \\ &= \mathcal{O}_p\left(\max\left\{\|\pi_{r_n} f_0 - f_0\|_n, \sqrt{\frac{r_n(d+2)\log[r_n V_n(d+2)]}{n}}, \sqrt{\frac{r_n(d+2)\log n}{n}}\right\}\right).\end{aligned}$$

□

REMARK 4.1. *Recall that a sufficient condition to ensure consistency is $r_n(d+2)V_n^2\log[r_nV_n(d+2)] = o(n)$. Under such a condition, $r_n(d+2)\log[r_nV_n(d+2)] \leq n$, the rate of convergence can be simplified to*

$$\|\hat{f}_n - f_0\|_n = \mathcal{O}_p\left(\max\left\{\|\pi_{r_n} f_0 - f_0\|_n, \sqrt{\frac{r_n(d+2)\log n}{n}}\right\}\right).$$

If we assume $f_0 \in \mathcal{F}$ where $\mathcal{F}$ is the space of functions with finite first absolute moments of the Fourier magnitude distributions, i.e.,

$$\begin{aligned}\mathcal{F} = \Bigg\{ f : \mathbb{R}^d \to \mathbb{R} : f(\boldsymbol{x}) = \int \exp\left\{i\boldsymbol{a}^T\boldsymbol{x}\right\} \mathrm{d}\mu_f(\boldsymbol{a}),\\ \|\mu_f\|_1 := \int \max(\|\boldsymbol{a}\|_1, 1)\mathrm{d}|\mu_f|(\boldsymbol{a}) \leq C\Bigg\},\end{aligned} \tag{4.1}$$

where μ_f is a complex measure on $\mathbb{R}^d$. $|\mu_f|$ denotes the total variation of μ_f, i.e., $|\mu|(A) = \sup\sum_{n=1}^\infty |\mu(A_n)|$ and the supremum is taken over all measurable partitions $\{A_n\}_{n=1}^\infty$ of A. $\|\boldsymbol{a}\|_1 = \sum_{i=1}^d |a_i|$ for $\boldsymbol{a} = [a_1, \ldots, a_d]^T \in \mathbb{R}^d$. Theorem 3 in Makovoz (1996) shows that $\delta_n := \|f_0 - \pi_{r_n} f_0\|_n \lesssim r_n^{-1/2-1/(2d)}$. Therefore, if we let d fixed and $\rho_n = \delta_n^{-1}$ and $V_n \equiv V$ in the proof of Theorem 4.1, δ_n must also satisfy the following inequality:

$$\begin{aligned}& \rho_n^2\phi\left(\frac{1}{\rho_n}\right) \lesssim \rho_n r_n^{1/2}\log^{1/2}\left(\rho_n\tilde{C}_{r_n,d,V_n}\right) \lesssim \sqrt{n} \\ \Rightarrow\quad & \rho_n^2 r_n \log\rho_n + \rho_n^2 r_n\log r_n \lesssim n \\ \Rightarrow\quad & \delta_n^{-2}\left(-r_n\log\delta_n + r_n\log r_n\right) \lesssim n \\ \Rightarrow\quad & r_n^{1+\frac{1}{d}} r_n \log r_n \lesssim n.\end{aligned}$$

One possible choice of r_n to satisfy such condition is $r_n \asymp (n/\log n)^{\frac{d}{2+d}}$. In such a case, we obtain

$$\|\hat{f}_n - f_0\|_n = \mathcal{O}_p\left(\left(\frac{n}{\log n}\right)^{-\frac{1+1/d}{4(1+1/(2d))}}\right),$$

which is the same rate obtained in Chen and Shen (1998). It is interesting to note that in the case where $d = 1$, we have $\|\hat{f}_n - f_0\|_n = \mathcal{O}_p\left((n/\log n)^{-1/3}\right)$. Such rate is close to the $\mathcal{O}_p(n^{-1/3})$, which is the convergence rate in nonparametric least square problems when the class of functions considered has bounded variation in $\mathbb{R}$ (see Example 9.3.3 in van de Geer (2000)). As shown in Proposition 7.3 in the Appendix, $\mathcal{F}_{r_n}$ is a class of functions with bounded variation in $\mathbb{R}$. Therfore, the convergence rate we obtained makes sense.

5. Asymptotic Normality.

To establish the asymptotic normality of sieve estimator for neural network, we follow the idea in Shen (1997) and start by calculating the Gâteaux derivative of the empirical criterion function $\mathbb{Q}_n(f) = n^{-1}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2$,

$$\begin{aligned}
\mathbb{Q}'_{n,f_0}[f - f_0] &= \lim_{t\to 0}\frac{1}{t}\left[\frac{1}{n}\sum_{i=1}^n (y_i - f_0(\boldsymbol{x}_i) - t(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)))^2 - \frac{1}{n}\sum_{i=1}^n (y_i - f_0(\boldsymbol{x}_i))^2\right] \\
&= \lim_{t\to 0}\frac{1}{n}\sum_{i=1}^n \frac{1}{t}\big[(y_i - f_0(\boldsymbol{x}_i))^2 - 2t(y_i - f_0(\boldsymbol{x}_i))(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \\
&\qquad + t^2(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 - (y_i - f_0(\boldsymbol{x}_i))^2\big] \\
&= -\frac{2}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)).
\end{aligned}$$

Then the remainder of first-order functional Taylor series expansion is

$$\begin{aligned}
R_n[f - f_0] &= \mathbb{Q}_n(f) - \mathbb{Q}_n(f_0) - \mathbb{Q}'_{n,f_0}[f - f_0] \\
&= \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 - \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2 + \frac{2}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \\
&= \frac{1}{n}\sum_{i=1}^n (\epsilon_i + f_0(\boldsymbol{x}_i) - f(\boldsymbol{x}_i))^2 - \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + \frac{2}{n}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) \\
&= \frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i))^2 \\
&= \|f - f_0\|_n^2.
\end{aligned}$$

As will be seen in the proof of asymptotic normality, the rate of convergence for the empirical process $\{n^{-1/2}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) : f \in \mathcal{F}_{r_n}\}$ plays an important role. Here we establish a lemma, which will be used to find the desired rate of convergence.

Lemma 5.1. *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim P_i$. Define the empirical process $\{\nu_n(f)\}$ as*

$$\nu_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [f(X_i) - P_i f].$$

Let $\mathcal{F}_n = \{f : \|f\|_\infty \le V_n\}$, $\epsilon > 0$ and $\alpha \ge \sup_{f\in\mathcal{F}_n} n^{-1}\sum_{i=1}^n \mathit{Var}[f(X_i)]$ be arbitrary. Define t_0 by $H(t_0, \mathcal{F}_n, \|\cdot\|_\infty) = \frac{\epsilon}{4}\psi(M,n,\alpha)$, where $\psi(M,n,\alpha) = M^2/\left[2\alpha\left(1 + \frac{MV_n}{2\sqrt{n}\alpha}\right)\right]$. If

$$H(u, \mathcal{F}_n, \|\cdot\|_\infty) \le A_n u^{-r}, \tag{5.1}$$

for some $0 < r < 2$ and $u \in (0, a]$, where a is a small positive number, and there exists a positive constant $K_i = K_i(r, \epsilon)$, $i = 1, 2$ such that

$$M \ge K_1 A_n^{\frac{2}{r+2}} V_n^{\frac{2-r}{r+2}} n^{\frac{r-2}{2(r+2)}} \vee K_2 A_n^{1/2} \alpha^{\frac{2-r}{4}},$$

we have

$$\mathbb{P}^*\left(\sup_{f\in\mathcal{F}_n} |\nu_n(f)| > M\right) \le 5\exp\{-(1-\epsilon)\psi(M,n,\alpha)\}.$$

Proof. The proof of the lemma is similar to the proof of Corollary 2.2 in Alexander (1984) and the proof of Lemma 1 in Shen and Wong (1994). Since $H(u, \mathcal{F}_n, \|\cdot\|_\infty) \le A_n u^{-r}$ for some $0 < r < 2$, we have

$$I(s,t) := \int_s^t H^{1/2}(u, \mathcal{F}_n, \|\cdot\|_\infty)\mathrm{d}u \le 2(2-r)^{-1} A_n^{\frac{1}{2}} t^{1-\frac{r}{2}}.$$

Based on the assumption of

$$A_n t_0^{-r} \ge H(t_0, \mathcal{F}_n, \|\cdot\|_\infty) = \frac{\epsilon}{4}\psi(M,n,\alpha),$$

we have $t_0 \le \left[\frac{4A_n}{\epsilon\psi}\right]^{1/r}$. Note that $\psi(M,n,\alpha) \ge M^2/(4\alpha)$ if $M \le 3\sqrt{n}\alpha/V_n$ and $2(\sqrt{n}\alpha + MV_n/3) \le 4MV_n/3$ if $M \ge 3\sqrt{n}\alpha/V_n$ and hence $\psi(M,n,\alpha) \ge 3\sqrt{n}M/(4V_n)$. In summary,

$$\psi(M,n,\alpha) \ge \begin{cases} M^2/(4\alpha) & \text{if } M < 3\sqrt{n}\alpha/V_n, \\ 3\sqrt{n}M/(4V_n) & \text{if } M \ge 3\sqrt{n}\alpha/V_n \end{cases}.$$

Therefore, if $M \geq 3\sqrt{n}\alpha/V_n$,

$$\begin{aligned}
2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) &\leq 2^9\epsilon^{-3/2}(2-r)^{-1}A_n^{1/2}t_0^{1-\frac{r}{2}} \\
&\leq 2^9\epsilon^{-3/2}(2-r)^{-1}\left(\frac{4}{\epsilon}\right)^{\frac{1}{r}-\frac{1}{2}} A_n^{1/r}\left(\frac{3}{4V_n}\sqrt{n}M\right)^{\frac{1}{2}-\frac{1}{r}} \\
&= \tilde{K}_1 A_n^{1/r} V_n^{\frac{1}{r}-\frac{1}{2}} n^{\frac{1}{4}-\frac{1}{2r}} M^{\frac{1}{2}-\frac{1}{r}},
\end{aligned}$$

where $\tilde{K}_1 = 2^9\epsilon^{-3/2}(2-r)^{-1}\left(\frac{4}{\epsilon}\right)^{\frac{1}{r}-\frac{1}{2}}\left(\frac{3}{4}\right)^{\frac{1}{2}-\frac{1}{r}}$. Hence

$$\begin{aligned}
2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) < M &\Leftrightarrow \tilde{K}_1 A_n^{1/r} V_n^{\frac{1}{r}-\frac{1}{2}} n^{\frac{1}{4}-\frac{1}{2r}} M^{\frac{1}{2}-\frac{1}{r}} < M \\
&\Leftrightarrow \tilde{K}_1 A_n^{1/r} V_n^{\frac{1}{r}-\frac{1}{2}} n^{\frac{r-2}{4r}} < M^{\frac{1}{r}+\frac{1}{2}} \\
&\Leftrightarrow M > K_1 A_n^{\frac{2}{r+2}} V_n^{\frac{2-r}{r+2}} n^{\frac{r-2}{2(r+2)}},
\end{aligned}$$

where $K_1 = \tilde{K}_1^{\frac{2r}{r+2}}$. On the other hand, if $M < 3\sqrt{n}\alpha/V_n$,

$$\begin{aligned}
2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) &\leq 2^9\epsilon^{-3/2}(2-r)^{-1}A_n^{1/2}t_0^{1-\frac{r}{2}} \\
&\leq 2^9\epsilon^{-3/2}(2-r)^{-1}\left(\frac{4}{\epsilon}\right)^{\frac{1}{r}-\frac{1}{2}} A_n^{1/r}\left(\frac{M^2}{4\alpha}\right)^{\frac{1}{2}-\frac{1}{r}} \\
&= \tilde{K}_2 A_n^{1/r} M^{1-\frac{2}{r}} \alpha^{\frac{1}{r}-\frac{1}{2}},
\end{aligned}$$

where $\tilde{K}_2 = 2^9\epsilon^{-3/2}(2-r)^{-1}\left(\frac{4}{\epsilon}\right)^{\frac{1}{r}-\frac{1}{2}}\left(\frac{1}{4}\right)^{\frac{1}{2}-\frac{1}{r}}$. Hence

$$\begin{aligned}
2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) < M &\Leftrightarrow \tilde{K}_2 A_n^{1/r} M^{1-\frac{2}{r}} \alpha^{\frac{1}{r}-\frac{1}{2}} < M \\
&\Leftrightarrow \tilde{K}_2 A_n^{1/r} \alpha^{\frac{2-r}{2r}} < M^{\frac{2}{r}} \\
&\Leftrightarrow M > K_2 A_n^{1/2} \alpha^{\frac{2-r}{4}},
\end{aligned}$$

where $K_2 = \tilde{K}_2^{r/2}$. In conclusion, if $M \geq K_1 A_n^{\frac{2}{r+2}} V_n^{\frac{2-r}{r+2}} n^{\frac{r-2}{2(r+2)}} \vee K_2 A_n^{1/2}\alpha^{\frac{2-r}{4}}$, then $2^8\epsilon^{-3/2} I\left(\frac{\epsilon M}{64\sqrt{n}}, t_0\right) < M$. By Theorem 2.1 in Alexander (1984), we have the desired result. □

As a Corollary to Lemma 5.1, we can show that the supremum of the empirical process $\{n^{-1/2}\sum_{i=1}^n \epsilon_i(f(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) : f \in \mathcal{F}_{r_n}\}$ converges to 0 in probability.

Corollary 5.1. *Let ρ_n satisfy $\rho_n\|\hat{f}_n - f_0\|_n = \mathcal{O}_p(1)$ and $\mathcal{F}_{r_n}$ be the class of neural network sieves as defined in (1.2). Suppose that $\mathbb{E}[|\epsilon|^{2+\lambda}] < \infty$ for some $\lambda > 0$. Then under the conditions*

(C1) $r_n(d+2)V_n\log[r_nV_n(d+2)] = o(n^{1/4})$;
(C2) $n\rho_n^{-2}/V_n^{\lambda} = o(1)$,

we have

$$\sup_{\|f-f_0\|_n\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(f-f_0)(\boldsymbol{x}_i)\right| = o_p(1).$$

Proof. To establish the desired result, we apply the truncation device.

$\|f\|_\infty \le V_n$

$$\begin{aligned}
&\mathbb{P}^*\left(\sup_{\|f-f_0\|_n\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(f-f_0)(\boldsymbol{x}_i)\right| \gtrsim M\right)\\
\le&\mathbb{P}^*\left(\sup_{\|f-f_0\|_n\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\mathbb{I}_{\{|\epsilon_i|\le V_n\}}(f-f_0)(\boldsymbol{x}_i)\right| \gtrsim M\right)\\
&\qquad + \mathbb{P}^*\left(\sup_{\|f-f_0\|_n\le\rho_n^{-1}, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\mathbb{I}_{\{|\epsilon_i|> V_n\}}(f-f_0)(\boldsymbol{x}_i)\right| \gtrsim M\right)\\
:=&(I) + (II)
\end{aligned}$$

For (I), we can apply Lemma 5.1 directly. Note that $|\epsilon\mathbb{I}_{\{|\epsilon|\le V_n\}}(f-f_0)(\boldsymbol{x})| \le V_n(V_n + \|f_0\|_\infty) \lesssim V_n^2$ since $\|f_0\|_\infty < \infty$ and for $0 < \eta < 1$,

$$\begin{aligned}
\log N(\eta, \mathcal{F}_{r_n}, \|\cdot\|_\infty) &\le \log\left(\frac{4e[r_n(d+2)+1]\left(\frac{1}{4}V_n\right)^2}{\eta\left(\frac{1}{4}V_n - 1\right)}\right)^{r_n(d+2)+1}\\
&= [r_n(d+2)+1]\left(\log\tilde{C}_{r_n,d,V_n} + \log\frac{1}{\eta}\right)\\
&\le [r_n(d+2)+1]\left(\log\tilde{C}_{r_n,d,V_n} + \frac{1}{\eta} - 1\right)\\
&= C_{r_n,d,V_n}\left(1 + \frac{1}{\eta}\right)\\
&\le 2C_{r_n,d,V_n}\frac{1}{\eta},
\end{aligned}$$

where $\tilde{C}_{r_n,d,V_n} = \frac{e[r_n(d+2)+1]V_n^2}{V_n-4}$ and

$$\begin{aligned} C_{r_n,d,V_n} &= [r_n(d+2)+1]\log \tilde{C}_{r_n,d,V_n} - [r_n(d+2)+1] \\ &= [r_n(d+2)+1]\log \frac{[r_n(d+2)+1]V_n^2}{V_n-4} \\ &\sim r_n(d+2)\log[r_nV_n(d+2)]. \end{aligned}$$

Therefore, equation (5.1) is satisfied with $r=1$ and $A_n = 2C_{r_n,d,V_n}$. Following from Lemma 5.1, for $M \gtrsim C_{r_n,d,V_n}^{2/3}V_n^{2/3}n^{-1/6} \vee C_{r_n,d,V_n}^{1/2}\alpha^{1/4}$, we have $(I) \leq 5\exp\{-(1-\epsilon)\psi(M,n,\alpha)\}$ and hence

$\alpha \quad 1 \leq \rho^{-1}(C)$

$$\sup_{\|f-f_0\|\leq\rho_n^{-1}, f\in\mathcal{F}_{r_n}} \left|\underbrace{\frac{1}{\sqrt{n}}\epsilon_i\mathbb{I}_{\{|\epsilon_i|\leq V_n\}}(f-f_0)(\boldsymbol{x}_i)}\right| = \mathcal{O}_p\left(\frac{C_{r_n,d,V_n}^{2/3}V_n^{2/3}}{n^{1/6}}\right).$$

$\rho_n^{-1} =$

Since by using (C1),

$$\frac{C_{r_n,d,V_n}^{2/3}V_n^{2/3}}{n^{1/6}} \sim \left(\underbrace{\frac{r_n(d+2)V_n\log[r_nV_n(d+2)]}{n^{1/4}}}\right)^{2/3} = o_p(1).$$

For (II), by using the Cauchy-Schwarz inequality, we have

$$\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\mathbb{I}_{\{|\epsilon_i|>V_n\}}(f-f_0)(\boldsymbol{x}_i)\right| \leq \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2\mathbb{I}_{\{|\epsilon_i|>V_n\}}\right)^{1/2}\|f-f_0\|_n.$$

Then it follows from the Markov inequality that

$$\begin{aligned} \boxed{(II)} &= \mathbb{P}^*\left(\sup_{\|f-f_0\|_n\leq\rho_n^{-1}, f\in\mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\mathbb{I}_{\{|\epsilon_i|>V_n\}}\underline{(f-f_0)}(\boldsymbol{x}_i)\right| \gtrsim Mn^{-1/2}\right) \\ &\leq \mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2\mathbb{I}_{\{|\epsilon_i|>V_n\}}\right)^{1/2}\underline{\rho_n^{-1}} \gtrsim Mn^{-1/2}\right) \\ &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i^2\mathbb{I}_{\{|\epsilon_i|>V_n\}} \gtrsim M^2n^{-1}\rho_n^2\right) \\ &\lesssim M^{-2}n\rho_n^{-2}\mathbb{E}[\epsilon^2\mathbb{I}_{|\epsilon|>V_n}] \\ &\lesssim \boxed{M^{-2}n\rho_n^{-2}}\frac{\mathbb{E}[|\epsilon|^{2+\lambda}]}{V_n^\lambda}. \end{aligned}$$

$\rho_n \ll l_n = o_p(1)$

$\partial.$ $\boxed{o(1)}$ $\underline{\rho_n \ll l_n = o_p(1)}$

Based on condition $\boxed{(C2)}$, we have $(II) \to 0$, and

$$\underbrace{\sup_{\|f-f_0\|_n\leq\rho_n^{-1}, f\in\mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\mathbb{I}_{\{|\epsilon_i|>V_n\}}(f-f_0)(\boldsymbol{x}_i)\right| = o_p(1).}$$

Combining the results we obtained above, we get

$$\sup_{\|f-f_0\|_n\leq\rho_n^{-1}, f\in\mathcal{F}_{r_n}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i(f-f_0)(\boldsymbol{x}_i)\right| = o_p(1)$$

□

Remark 5.1. *Condition (C2) can be further simplified using the results from Theorem 4.1. If*

$$\eta_n = \mathcal{O}\left(\min\{\|\pi_{r_n}f_0-f_0\|_n^2, r_n(d+2)\log(r_nV_n(d+2))/n, r_n(d+2)\log n/n\}\right),$$

then

$$\rho_n^{-1}\asymp\max\left\{\|\pi_{r_n}f_0-f_0\|_n, \sqrt{r_n(d+2)\log[r_nV_n(d+2)]/n}, \sqrt{r_n(d+2)\log n/n}\right\}.$$

It follows from condition (C1) that

$$\rho_n^{-1}\asymp\max\left\{\|\pi_{r_n}f_0-f_0\|_n, \sqrt{r_n(d+2)\log n/n}\right\}.$$

For simplicity, we assume that $\rho_n^{-1}\asymp\sqrt{r_n(d+2)\log n/n}$, which holds for functions having finite first absolute moments of the Fourier magnitude distributions as discussed at the end of section 4.4. Then in this case,

$$n\rho_n^{-2}/V_n^\lambda\asymp r_n(d+2)\log n/V_n^\lambda,$$

so that condition (C2) becomes $r_n(d+2)\log n/V_n^\lambda\to 0$.

Now we are going to establish the asymptotic normality for neural network estimators. For $f\in\{f\in\mathcal{F}_{r_n}:\|f-f_0\|_n\leq\rho_n^{-1}\}$, we consider a local alternative

$$\tilde{f}_n(f) = (1-\delta_n)f+\delta_n(f_0+\iota), \tag{5.2}$$

where $0\leq\delta_n=\eta_n^{1/2}=o(n^{-1/2})$ is chosen such that $\rho_n\delta_n=o(1)$ and $\iota(\boldsymbol{x})\equiv 1$.

Theorem 5.1 (Asymptotic Normality). *Suppose that $0\leq\eta_n=o(n^{-1})$ and conditions (C1) and (C2) in Corollary 5.1 hold. We further assume that the following two conditions hold*

(C3) $\sup_{f\in\mathcal{F}_{r_n}:\|f-f_0\|_n\leq\rho_n^{-1}}\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n=\mathcal{O}_p(\rho_n\delta_n^2)$*;*

(C4) $\sup_{f\in\mathcal{F}_{r_n}:\|f-f_0\|_n\leq\rho_n^{-1}}\frac{1}{n}\sum_{i=1}^n\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\tilde{f}_n(f)(\boldsymbol{x}_i)\right)=\mathcal{O}_p(\delta_n^2)$,

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right]\xrightarrow{d}\mathcal{N}(0,\sigma^2).$$

Before we proceed to the proof of the theorem, let us focus on the conditions given in the theorem. Note that if (C1) holds, we have

$$r_n(d+2)V_n^2\log[r_nV_n(d+2)]\le[r_n(d+2)]^4V_n^4\left(\log[r_nV_n(d+2)]\right)^4=o(n),$$

so it is a sufficient condition to ensure the consistency of the neural network sieve estimator. As in Remark 3.2, we consider some simple scenarios here. If $V_n=\mathcal{O}(r_n)$, then $r_n(d+2)V_n\log[r_nV_n(d+2)]=\mathcal{O}\left(r_n^2\log r_n\right)$ so that a possible growth rate for r_n is $r_n=o\left(n^{1/8}/(\log n)^2\right)$. On the other hand, if $r_n=\log V_n$, then $r_n(d+2)V_n\log[r_nV_n(d+2)]=\mathcal{O}\left(V_n(\log V_n)^2\right)$ and a possible growth rate for V_n is $V_n=o(n^{1/4}/(\log n)^2)$. Thus, in both cases, the growth rate required for the asymptotic normality of neural network sieve estimator is slower than the growth rate required for the consistency as given in Remark 3.2. One explanation is that due to the Universal Approximation Theorem, a neural network with one hidden layer can approximate a continuous function on compact support arbitrarily well if the number of hidden units is sufficiently large. Therefore, if the number of hidden units is too large, the neural network sieve estimator $\hat{f}_n$ may be very close to the best projector of the true function f_0 in $\mathcal{F}_{r_n}$ so that the error $\sum_{i=1}^n\left[\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right]$ could be close to zero, resulting a small variation. By allowing slower growth rate of the number of hidden units can increase the variations of $\sum_{i=1}^n\left[\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right]$, which makes the asymptotic normality more reasonable. On the other hand, condition (C3) and condition (C4) are similar conditions as in Shen (1997), which are known for conditions on approximation error. These conditions indicate that the approximation rate of a single layer neural network cannot be too slow, otherwise it may require a huge number of samples to reach the desired approximation error. Therefore, the conditions in the theorem can be considered as a trade-off between bias and variance.

Proof of Theorem 5.1. The main idea of the proof is to use the functional Taylor series expansion for $\mathbb{Q}_n(f)$ and then carefully bound each term in the expansion. For any $f\in\{f\in\mathcal{F}_{r_n}:\|f-f_0\|_n\le\rho_n^{-1}\}$,

$$\begin{aligned}\mathbb{Q}_n(f)&=\mathbb{Q}_n(f_0)+\mathbb{Q}'_{n,f_0}[f-f_0]+R_n[f-f_0]\\&=\frac{1}{n}\sum_{i=1}^n\epsilon_i^2-\frac{2}{n}\sum_{i=1}^n\epsilon_i(f(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i))+\frac{1}{n}\sum_{i=1}^n(f(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i))^2.\end{aligned}\tag{5.3}$$

Note that

$$\begin{aligned}\|\tilde{f}_n(f) - f_0\|_n &= \|(1-\delta_n)\hat{f}_n + \delta_n(f_0 + \iota) - f_0\|_n \\ &= \|(1-\delta_n)(\hat{f}_n - f_0) + \delta_n \iota\|_n \\ &\leq (1-\delta_n)\|\hat{f}_n - f_0\|_n + \delta_n,\end{aligned}$$

and since $\delta_n = o(n^{-1/2})$, we can know that with probability tending to 1, $\|\tilde{f}_n(f) - f_0\|_n \leq \rho_n^{-1}$. Then replacing f in (5.3) by $\hat{f}_n$ and $\pi_{r_n}\tilde{f}_n(f)$ as defined in (5.2), we get

$$\begin{aligned}\mathbb{Q}_n(\hat{f}_n) &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \frac{2}{n}\sum_{i=1}^n \epsilon_i(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) + \|\hat{f}_n - f_0\|_n^2 \\ \mathbb{Q}_n(\pi_{r_n}\tilde{f}_n(f)) &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \frac{2}{n}\sum_{i=1}^n \epsilon_i(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)) + \|\pi_{r_n}\tilde{f}_n(f) - f_0\|_n^2.\end{aligned}$$

Subtracting these two equations yields

$$\mathbb{Q}_n(\hat{f}_n) = \mathbb{Q}_n(\pi_{r_n}\tilde{f}_n(f)) + \frac{2}{n}\sum_{i=1}^n \epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i) - \hat{f}_n(\boldsymbol{x}_i)\right) + \|\hat{f}_n - f_0\|_n^2 - \|\pi_{r_n}\tilde{f}_n(f) - f_0\|_n^2.$$

Now note that

$$\begin{aligned}\|\pi_{r_n}\tilde{f}_n(f) - f_0\|_n^2 &= \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + \tilde{f}_n(f) - f_0\|_n^2 \\ &= \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + (1-\delta_n)\hat{f}_n + \delta_n(f_0+\iota) - f_0\|_n^2 \\ &= \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + (1-\delta_n)(\hat{f}_n - f_0) + \delta_n\iota\|_n^2 \\ &= \Big\langle \pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + (1-\delta_n)(\hat{f}_n - f_0) + \delta_n\iota, \\ &\qquad\qquad \pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f) + (1-\delta_n)(\hat{f}_n - f_0) + \delta_n\iota\Big\rangle \\ &= \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n^2 + (1-\delta_n)^2\|\hat{f}_n - f_0\|_n^2 + \delta_n^2 \\ &\qquad + 2(1-\delta_n)\left\langle \pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f), \hat{f}_n - f_0\right\rangle \\ &\qquad + 2\delta_n\left\langle \pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f), \iota\right\rangle \\ &\qquad + 2(1-\delta_n)\delta_n\left\langle \hat{f}_n - f_0, \iota\right\rangle \\ &\leq (1-\delta_n)^2\|\hat{f}_n - f_0\|_n^2 + 2(1-\delta_n)\left\langle \hat{f}_n - f_0, \delta_n\iota\right\rangle + \delta_n^2 \\ &\qquad + 2(1-\delta_n)\|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n\|\hat{f}_n - f_0\|_n \\ &\qquad + 2\delta_n\|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n + \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n^2,\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Since

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\hat{f}_n(\boldsymbol{x}_i)\right)&=\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\tilde{f}_n(f)(\boldsymbol{x}_i)+\tilde{f}_n(f)(\boldsymbol{x}_i)-\hat{f}_n(\boldsymbol{x}_i)\right)\\&=\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\tilde{f}_n(f)(\boldsymbol{x}_i)\right)\\&\quad-\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right)-\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i,\end{aligned}$$

by the definition of $\hat{f}_n$, we have

$$\begin{aligned}-\mathcal{O}_p(\delta_n^2)&\le\inf_{f\in\mathcal{F}_{r_n}}\mathbb{Q}_n(f)-\mathbb{Q}_n(\hat{f}_n)\\&\le\mathbb{Q}_n(\pi_{r_n}\tilde{f}_n(f))-\mathbb{Q}_n(\hat{f}_n)\\&=\|\pi_{r_n}\tilde{f}_n(f)-f_0\|_n^2-\|\hat{f}_n-f_0\|_n^2-\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\hat{f}_n(\boldsymbol{x}_i)\right)\\&\le(1-\delta_n)^2\|\hat{f}_n-f_0\|_n^2-\|\hat{f}_n-f_0\|_n^2+2(1-\delta_n)\delta_n\left\langle\hat{f}_n-f_0,\iota\right\rangle\\&\qquad+2(1-\delta_n)\|\hat{f}_n-f_0\|_n\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n\\&\qquad+2\delta_n\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n+\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n^2\\&\qquad-\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\tilde{f}_n(f)(\boldsymbol{x}_i)\right)+\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right)\\&\qquad+\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i+\mathcal{O}_p(\delta_n^2)\\&=(-2\delta_n+\delta_n^2)\|\hat{f}_n-f_0\|_n^2\\&\qquad+2(1-\delta_n)\delta_n\left\langle\hat{f}_n-f_0,\iota\right\rangle+2(1-\delta_n)\|\hat{f}_n-f_0\|_n\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n\\&\qquad+2\delta_n\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n+\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n^2\\&\qquad-\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i)-\tilde{f}_n(f)(\boldsymbol{x}_i)\right)+\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right)\\&\qquad+\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i+\mathcal{O}_p(\delta_n^2)\\&\le\delta_n^2\|\hat{f}_n-f_0\|_n^2+2(1-\delta_n)\delta_n\left\langle\hat{f}_n-f_0,\iota\right\rangle+2(1-\delta_n)\|\hat{f}_n-f_0\|_n\|\pi_{r_n}\tilde{f}_n(f)-\tilde{f}_n(f)\|_n\end{aligned}$$

$$
\begin{aligned}
&+ 2\delta_n \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n + \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n^2 \\
&- \frac{2}{n}\sum_{i=1}^n \epsilon_i \left(\pi_{r_n}\tilde{f}_n(f)(\boldsymbol{x}_i) - \tilde{f}_n(f)(\boldsymbol{x}_i)\right) + \frac{2}{n}\delta_n \sum_{i=1}^n \epsilon_i \left(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right) \\
&+ \frac{2}{n}\delta_n \sum_{i=1}^n \epsilon_i + \mathcal{O}_p(\delta_n^2),
\end{aligned}
\tag{5.4}
$$

where the last inequality follows by noting that $(1-\delta_n)^2 - 1 = -2\delta_n + \delta_n^2 \leq \delta_n^2$. From the condition (C1), we can get

$$
\begin{aligned}
&[r_n(d+2)+1]V_n^2 \log[r_n V_n(d+2)+1] \\
&\qquad \leq ([r_n(d+2)+1]V_n \log[r_n V_n(d+2)+1])^4 = o(n).
\end{aligned}
$$

Combining with Theorem 3.1, we obtain that $\|\hat{f}_n - f_0\|_n = o_p(1)$ and hence $\delta_n^2\|\hat{f}_n - f_0\|_n^2 = o_p(\delta_n^2)$. From condition (C3), we have

$$
\begin{aligned}
2(1-\delta_n)\|\hat{f}_n - f_0\|_n \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n &\leq 2\|\hat{f}_n - f_0\|_n \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n \\
&= \mathcal{O}_p\left(\rho_n^{-1}\rho_n\delta_n^2\right) = \mathcal{O}_p(\delta_n^2).
\end{aligned}
$$

Similarly, since $\rho_n\delta_n = o(1)$, we have

$$
\begin{aligned}
2\delta_n \|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n &= \mathcal{O}_p(\delta_n \cdot \rho_n\delta_n^2) = o_p(\delta_n^2) \\
\|\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\|_n^2 &= \mathcal{O}_p(\rho_n^2\delta_n^4) = o_p(\delta_n^2).
\end{aligned}
$$

Based on condition (C4), we know that

$$
\frac{2}{n}\sum_{i=1}^n \epsilon_i \left(\pi_{r_n}\tilde{f}_n(f) - \tilde{f}_n(f)\right) = \mathcal{O}_p(\delta_n^2),
$$

and from Corollary 5.1, we also have

$$
\frac{2}{n}\delta_n \sum_{i=1}^n \epsilon_i \left(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right) = o_p(\delta_n \cdot n^{-1/2}).
$$

It follows from these observations that

$$
-2(1-\delta_n)\left\langle \hat{f}_n - f_0, \delta_n\iota \right\rangle + \frac{2\delta_n}{n}\sum_{i=1}^n \epsilon_i \leq \mathcal{O}_p(\delta_n^2) + \varnothing_p(\delta_n^2) + o_p(\delta_n \cdot n^{-1/2}),
$$

which implies that

$$
-(1-\delta_n)\left\langle \hat{f}_n - f_0, \iota \right\rangle + \frac{1}{n}\sum_{i=1}^n \epsilon_i \leq \mathcal{O}_p(\delta_n) + o_p(n^{-1/2}) = o_p(n^{-1/2}).
$$

By replacing ι with $-\iota$, we can obtain the same result and hence

$$\begin{aligned}\left|\left\langle \hat{f}_n - f_0, \iota \right\rangle - \frac{1}{n}\sum_{i=1}^n \epsilon_i\right| &\leq \left|(1-\delta_n)\left\langle \hat{f}_n - f_0, \iota \right\rangle - \frac{1}{n}\sum_{i=1}^n \epsilon_i\right| + \delta_n \left|\left\langle \hat{f}_n - f_0, \iota \right\rangle\right| \\ &\leq o_p(n^{-1/2}) + \delta_n \|\hat{f}_n - f_0\|_n \\ &= o_p(n^{-1/2}).\end{aligned}$$

Therefore,

$$\left\langle \hat{f}_n - f_0, \iota \right\rangle = \frac{1}{n}\sum_{i=1}^n \epsilon_i + o_p(n^{-1/2}),$$

and the desired result follows from the classical Central Limit Theorem. □

Theorem 5.1 can be used directly for hypothesis testing of neural network with one hidden layer if we know the variance of the random error σ^2. In practice, this is rarely the case. To perform hypothesis testing when σ^2 is unknown, it is natural to find a good estimator of σ^2 and use a "plug-in" test statistic. A natural estimator for σ^2 is

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(y_i - \hat{f}_n(\boldsymbol{x}_i)\right)^2 = \mathbb{Q}_n\left(\hat{f}_n\right).$$

We then need to establish the asymptotic normality for the statistic $\frac{1}{\hat{\sigma}_n\sqrt{n}}\sum_{i=1}^n \left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$.

Theorem 5.2 (Asymptotic Normality for Plug-in Statistic). *Suppose that $f_0 \in C(\mathcal{X})$, where $\mathcal{X} \subset \mathbb{R}^d$ is a compact set and $0 \leq \eta_n = o\left(n^{-1}\right)$. Then under the conditions as stated in Theorem 5.1, we have*

$$\frac{1}{\hat{\sigma}_n\sqrt{n}}\sum_{i=1}^n \left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right] \xrightarrow{d} \mathcal{N}(0,1).$$

Proof. Note that

$$\begin{aligned}\hat{\sigma}_n^2 = \mathbb{Q}_n(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^n \left(y_i - \hat{f}_n(\boldsymbol{x}_i)\right)^2 &= \frac{1}{n}\sum_{i=1}^n \left(f_0(\boldsymbol{x}_i) + \epsilon_i - \hat{f}_n(\boldsymbol{x}_i)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n \left(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right)^2 - \frac{2}{n}\sum_{i=1}^n \epsilon_i\left(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right) + \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \\ &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - \frac{2}{n}\sum_{i=1}^n \epsilon_i\left(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right) + \|\hat{f}_n - f_0\|_n^2\end{aligned}$$

Based on the rate of convergence of $\hat{f}_n$ we obtained in Theorem 4.1 and condition (C1), we know that

$$\left\|\hat{f}_n - f_0\right\|_n^2 = \mathcal{O}_p^*\left(\max\left\{\|\pi_{r_n} f_0 - f_0\|_n^2, \frac{r_n(d+2)\log n}{n}\right\}\right).$$

Under (C3), $\|\pi_{r_n} f_0 - f_0\|_n^2 = o\left(\rho_n^2 \delta_n^4\right) = o(n^{-1/2})$ and under (C1), we have

$$\begin{aligned}\left(\frac{r_n(d+2)\log n}{n}\right) &\leq o\left(\frac{n^{1/4}\log n}{n}\right)\\ &= o\left(\frac{\log n}{n^{3/4}}\right) = o(n^{-1/2}),\end{aligned}$$

which implies that $\left\|\hat{f}_n - f_0\right\|_n^2 = o_p(n^{-1/2})$. Moreover, by the same arguments as in the proof of Theorem 5.1, we can show that

$$\frac{2}{n}\sum_{i=1}^{n} \epsilon_i \left(\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right) = o_p(n^{-1/2}).$$

Therefore,

$$\mathbb{Q}_n(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2 + o_p(n^{-1/2}).$$

Based on the Weak Law of Large Numbers, we know that $\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 = \sigma^2 + o_p(1)$. Therefore,

$$\hat{\sigma}_n^2 = \mathbb{Q}_n(\hat{f}_n) = \sigma^2 + o_p(1),$$

and it follows from the Slutsky's Theorem and Theorem 5.1, we obtain

$$\frac{1}{\hat{\sigma}_n\sqrt{n}}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right] = \frac{\sigma}{\hat{\sigma}_n}\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right] \xrightarrow{d} \mathcal{N}(0,1).$$

□

6. Simulation Studies.

In this section, simulations were conducted to check the validity of the theoretical results obtained in the previous sections. We first used a simple simulation to show that it is hard for the parameter estimators in a neural network with one hidden layer to reach parametric consistency. Then the consistency of the neural network sieve estimators was examined under various simulation scenarios. Finally, we evaluated the asymptotic normality of the neural network sieve estimators.

6.1. *Parameter Inconsistency.* As mentioned in the introduction, due to the loss of identifiability of the parameters, the parameter estimators obtained from a neural network model are unlikely to be consistent. In this simulation, we use empirical results to confirm such observations. We simulated the response through the following model:

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{6.1}$$

where the total sample size $n = 500$, $x_1, \dots, x_n \sim$ i.i.d.$\mathcal{N}(0, 1)$, $\epsilon_1, \dots, \epsilon_n \sim$ i.i.d.$\mathcal{N}(0, 0.1^2)$. The true model is a single-layer neural network with two hidden units.

$$f_0(x_i) = -1 + \sigma(2x_i + 1) - \sigma(-x_i + 1), \tag{6.2}$$

When we conducted the simulation, we fitted the simulated data with a single-layer neural network to fit the data by setting the learning rate as 0.1 and performing 3e4 iterations for the back propagation. The cost after 3e4 iterations is 0.0106. Table 1 summarizes the estimated values of the parameters in this model.

Table 1: Comparison of the true parameters and the estimated parameters in a single-layer neural network with 2 hidden units.

Estimated Values	Weights				Biases		
	γ_1	γ_2	α_1	α_2	$\gamma_{0,1}$	$\gamma_{0,2}$	α_0
True Value	2.00	-1.00	1.00	-1.00	1.00	1.00	-1.00
Estimated Value	0.82	1.30	-0.34	-0.58	-0.03	-0.03	-1.04

Based on the results in Table 1, it is clear that the estimators for most of the weights and biases (except α_0) are far from reaching consistency. On the other hand, if we look at the curve of the true function and the curve of the fitted function as shown in Figure 1, we can see that most parts are fitted extremely well except for the tail parts. The approximation error $\|\hat{f} - f_0\|_n$ is almost zero as shown in the Figure. This suggests that we should study the asymptotic properties of the estimated function instead of the estimated parameters in the neural network.

6.2. *Consistency for Neural Network Sieve Estimators.* In this simulation, we are going to check the consistency result from Section 3 and the

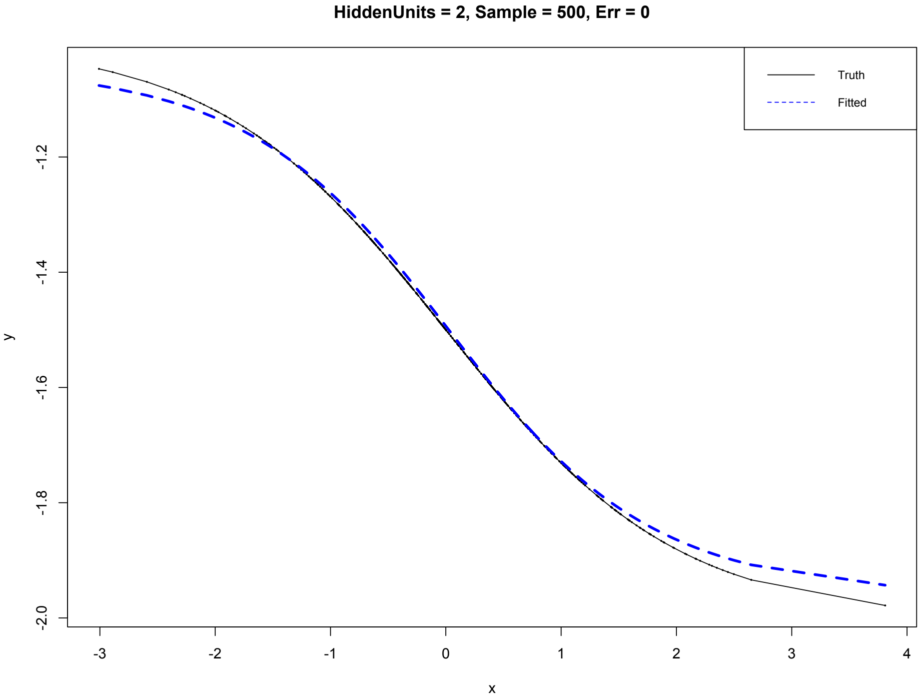


Fig 1: Comparison of the true function and the fitted function under the simulation model (6.1). The black curve is the true function defined in (6.2) and the blue dashed curve is the fitted curve obtained after fitting the neural network model. “Err” stands for the sqaure of empirical ℓ_2 distance between the esimated function and the true function, that is $Err = (1/n)\sum_{i=1}^{n}\left[\hat{f}_n(x_i) - f_0(x_i)\right]^2$.

validity of the assumption made in Theorem 3.1. Based on our construction of the neural network sieve estimators, in each sieve space $\mathcal{F}_{r_n}$, there is a constraint on the ℓ_1 norm for $\boldsymbol{\alpha}$: $\sum_{i=0}^{r_n} |\alpha_i| \leq V_n$. So finding the nearly optimal function in $\mathcal{F}_{r_n}$ for $\mathbb{Q}_n(f)$ is in fact a constrained optimization problem. A classical way to conduct this optimization is through introducing a Lagrange multiplier for each constraint. Nevertheless, it is usually hard to find an explicit connection between the Lagrange multiplier and the upper bound in the inequality constraint. Instead, we use the subgradient method as discussed in section 7 in Boyd and Mutapcic (2008). The basic idea is to update the parameter $\alpha_0, \ldots, \alpha_{r_n}$ through

$$\alpha_i^{(k+1)} = \alpha_i^{(k)} - \delta_k g^{(k)}, \quad i = 0, \ldots, r_n$$

where $\delta_k > 0$ is a step size and δ_k is chosen to be $0.1/\log(e+k)$ throughout the simulation, which is known as a nonsummable diminishing step size rule. $g^{(k)}$ is a subgradient of the objective or the constraint function $\sum_{j=0}^{r_n} |\alpha_j| - V_n$ at $\boldsymbol{\alpha}^{(k)}$. More specifically, we take

$$g^{(k)} \in \begin{cases} \partial_{\boldsymbol{\alpha}^{(k)}} \mathbb{Q}_n(f) & \text{if } \sum_{j=0}^{r_n} |\alpha_j| \leq V_n \\ \partial_{\boldsymbol{\alpha}^{(k)}} \sum_{j=0}^{r_n} |\alpha_j| & \text{if } \sum_{j=0}^{r_n} |\alpha_j| > V_n \end{cases} .$$

The updating equations of $\boldsymbol{\gamma}_1, \ldots \boldsymbol{\gamma}_{r_n}, \gamma_{0,1}, \ldots, \gamma_{0,r_n}$ remains the same as those in the classical gradient descent algorithm.

We still used equation (6.1) to simulate response, but assumed that the random error $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, 0.7^2)$. For the true function $f_0(x)$, we considered the following three functions:

(1) A neural network with a single hidden layer and 2 hidden units, which is the same as in equation (6.2).
(2) A trigonometric function:

$$f_0(x) = \sin\left(\frac{\pi}{3}x\right) + \frac{1}{3}\cos\left(\frac{\pi}{4}x + 1\right) \tag{6.3}$$

(3) A continuous function having a non-differential point

$$f_0(x) = \begin{cases} -2x & \text{if } x \leq 0 \\ \sqrt{x}\left(x - \frac{1}{4}\right) & \text{if } x > 0 \end{cases} \tag{6.4}$$

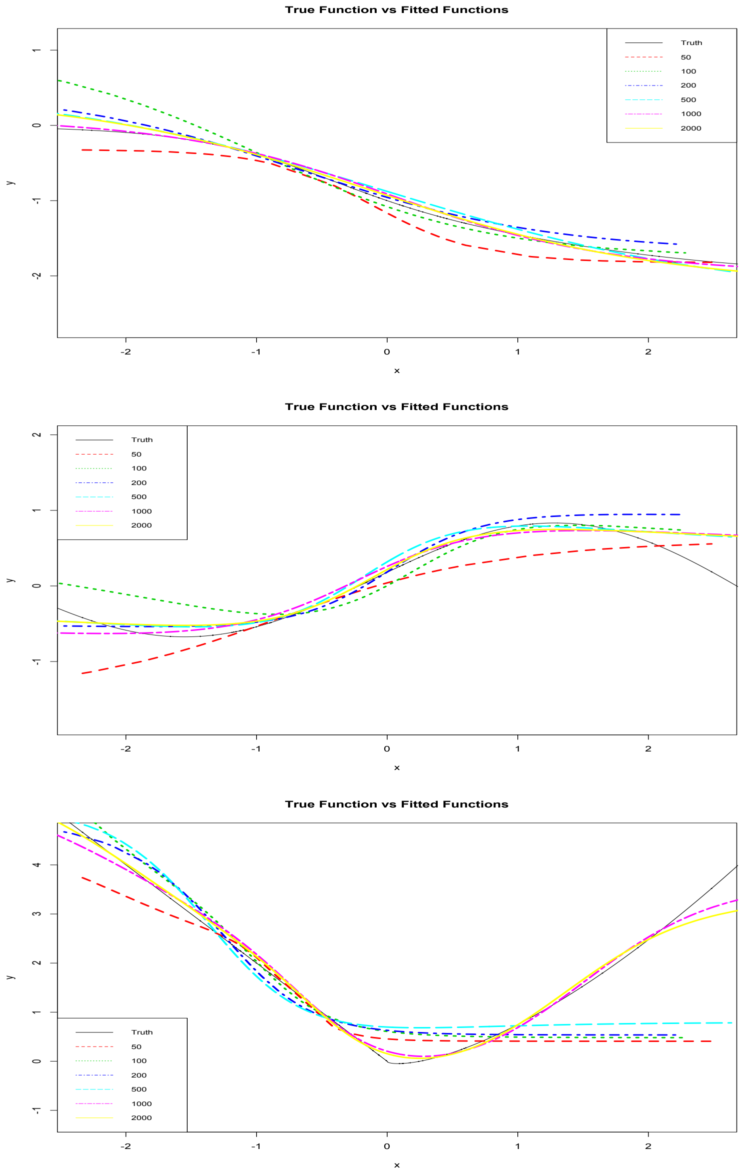
We then trained a neural network using the subgradient method mentioned above and set the number of iterations used for fitting as 20,000. We chose the growth rate on the number of hidden units $r_n = n^{1/4}$ and the upper bound for ℓ_1 norm of the weights and bias from the hidden layer to the output layer $V_n = 10n^{1/4}$. Such choice satisfies the condition mentioned in Remark 3.2 and hence satisfies the condition in Theorem 3.1. We compared the errors $\|\hat{f}_n - f_0\|_n^2$ and the least square errors $\mathbb{Q}_n(\hat{f}_n)$ under different sample sizes. The results are summarized in Table 2.

TABLE 2
Comparison of errors $\|\hat{f}_n - f_0\|_n^2$ and the least square errors $\mathbb{Q}_n(\hat{f}_n)$ after 20,000 iterations under different sample sizes.

Sample Sizes	Neural Network		Sine		Piecewise Continuous	
	$\|\hat{f}_n - f_0\|_n^2$	$\mathbb{Q}_n(\hat{f}_n)$	$\|\hat{f}_n - f_0\|_n^2$	$\mathbb{Q}_n(\hat{f}_n)$	$\|\hat{f}_n - f_0\|_n^2$	$\mathbb{Q}_n(\hat{f}_n)$
50	3.33E-2	0.519	6.04E-2	0.513	6.20E-1	1.124
100	2.79E-2	0.552	3.04E-2	0.587	3.20E-1	0.920
200	6.05E-3	0.500	1.05E-2	0.501	2.51E-1	0.786
500	8.15E-3	0.484	1.19E-2	0.499	3.26E-1	0.769
1000	3.02E-3	0.475	1.54E-2	0.480	2.98E-2	0.489
2000	2.88E-3	0.500	9.72E-3	0.506	1.69E-2	0.515

As we can see from Table 2, the errors $\|\hat{f}_n - f_0\|_n^2$ overall has a decreasing pattern as the sample size increases. There are some cases where the error becomes a little bit larger when the sample sizes increases (e.g. the errors using 500 samples in all scenarios is larger than those errors using 200 sample). One explanation is that the number of hidden units increase from 3 (for 200 samples) to 4 (for 500 samples) under our simulation setup, which adds variation to the estimation performance. Overall, the table shows that the estimated function $\hat{f}_n$ is indeed consistent in the sense that $\|\hat{f}_n - f_0\|_n = o_p^*(1)$. Figure 2 plots the fitted functions and the true function, from which we can straightforwardly visualize the result more and draw the conclusions.

6.3. *Asymptotic Normality for Neural Network Sieve Estimators.* The last part of the simulation focuses on the asymptotic normality derived in Theorem 5.1. We still considered the same types of true functions as described in section 6.2 but sampled the random errors from the standard normal distribution. In this simulation, we still used the subgradient method to obtain the fitted model. The number of iterations used for fitting was set at 20,000. What is different from section 6.2 is the growth rates for r_n and V_n set in this simulation. As mentioned in section 5, the growth rates required for asymptotic normality are slower than those required for consistency. Therefore, we chose $r_n = n^{1/8}$ and $V_n = 10n^{1/10}$. Such choice satisfies the condition (C1) in Theorem 5.1. In order to get the normal Q-Q plot for



True Function vs Fitted Functions

True Function vs Fitted Functions

True Function vs Fitted Functions

Fig 2: Comparison of the true function and the fitted function for three different types of non-linear functions. The top panel shows the scenario when the true function is a single layer neural network; the middle panel shows the scenario when the true function is a sine function, and the bottom panel shows the scenario when the true function is a continuous function having a non-differentiable point. As we can see from all the cases, the fitted curve becomes closer to the truth as the sample size increases.

$n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$, we repeated the simulation 200 times.

Figure 3 to Figure 5 are the normal Q-Q plots under different nonlinear functions and various sample sizes. From the figures, we found that the statistic $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$ fit the normal distribution pretty well under all simulation scenarios. It is also worth to note that the Q-Q plots looks similar under all simulation scenarios. This is what we would expect since the limiting distribution for the statistic $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$ is $\mathcal{N}(0,1)$ under all scenarios. Another implication we can obtain from the Q-Q plots is that the statistic $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$ is robust to the choice of f_0. Therefore, as long as the true function f_0 is continuous, $\mathcal{N}(0,1)$ is a good asymptotic distribution for $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$, which facilitates hypothesis testing.

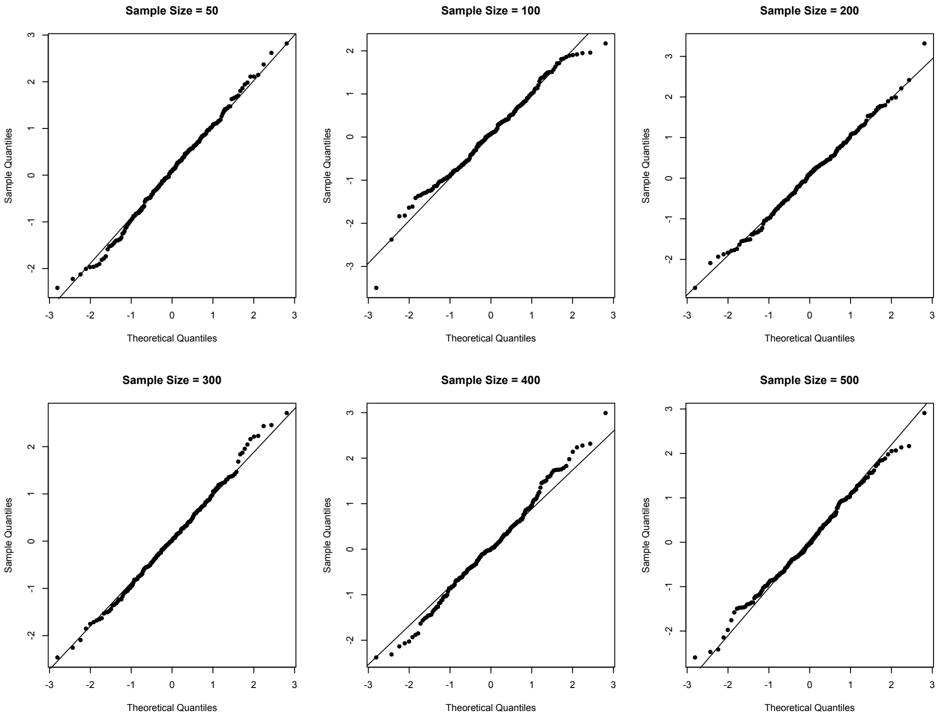


Fig 3: Normal Q-Q plot for $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$ various sample sizes. The true function f_0 is a single-layer neural network with 2 hidden units as defined in (6.2).

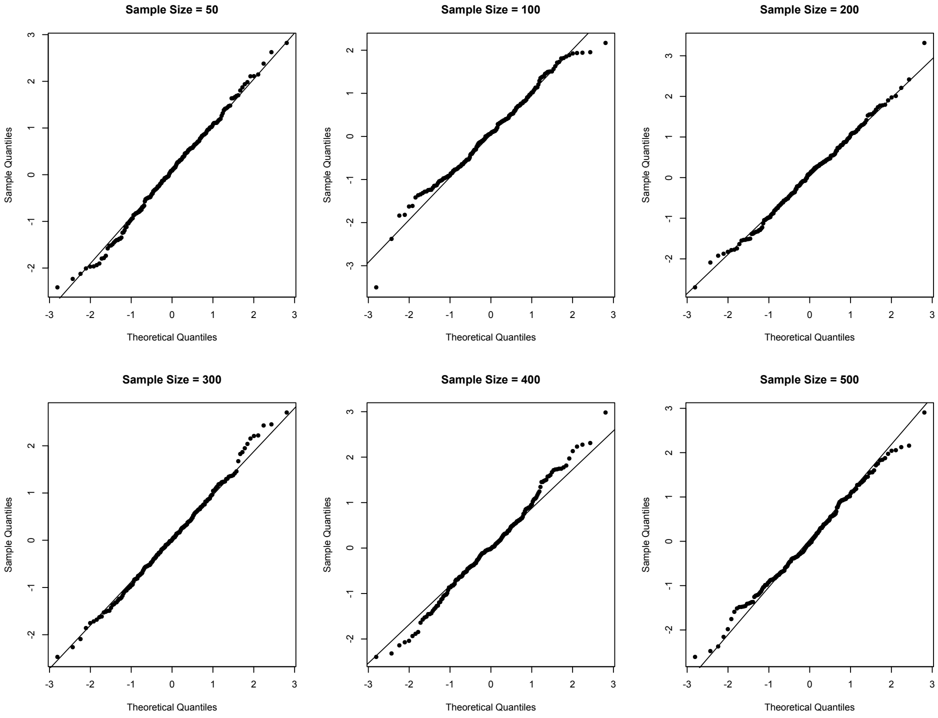


Fig 4: Normal Q-Q plot for $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$ various sample sizes. The true function f_0 is a trigonometric function as defined in (6.3).

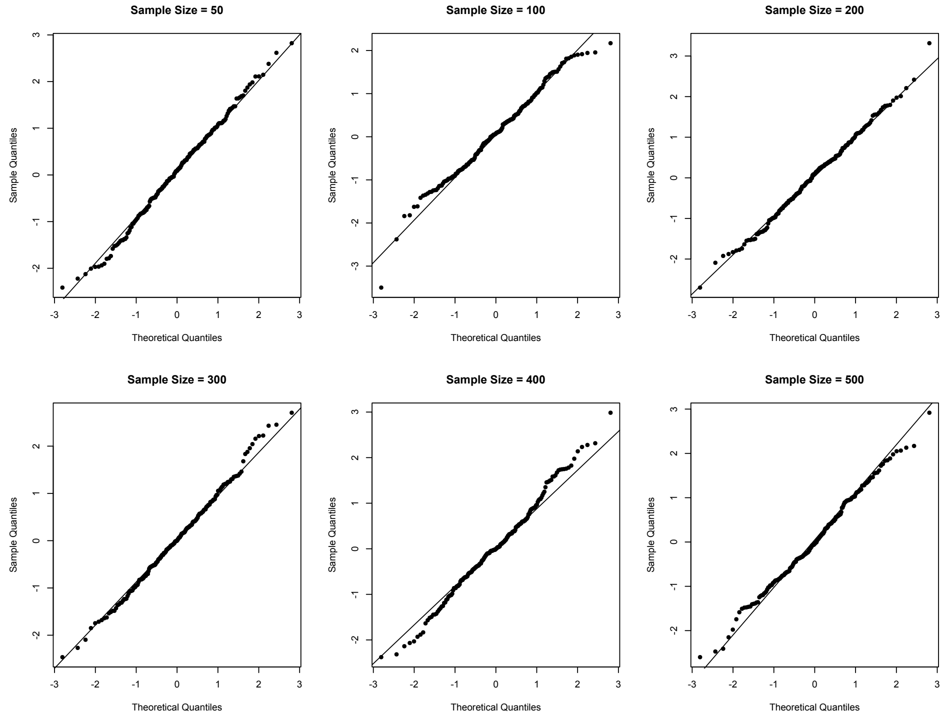


Fig 5: Normal Q-Q plot for $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i) - f_0(\boldsymbol{x}_i)\right]$ various sample sizes. The true function f_0 is a continuous function having a non-differential point as defined in (6.4).

Besides the Q-Q plots, we also conducted the normality tests to check whether $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right]$ follows the standard normal distribution. Specifically, we used the Shapiro-Wilks test and the Kolmogorov-Smirnov test to perform the normality test. Table 3 summarizes the p-values for both normality tests. As we observed from Table 3, in all cases, we failed to reject that $n^{-1/2}\sum_{i=1}^{n}\left[\hat{f}_n(\boldsymbol{x}_i)-f_0(\boldsymbol{x}_i)\right]$ follows the standard normal distribution.

TABLE 3

Summary of results from the Shapiro-Wilks test and the Kolmogorov-Smirnov test. We use "NN", "TRI" and "ND" to denote a neural network described in (6.2), a trigonometric function described in (6.3) and a continuous function having a non-differential point described in (6.4), respectively.

Sample Sizes	Shapiro-Wilks Test			Kolmogorov-Smirnov Test		
	NN	TRI	ND	NN	TRI	ND
50	0.878	0.884	0.881	0.584	0.597	0.595
100	0.098	0.095	0.095	0.472	0.508	0.484
200	0.940	0.944	0.944	0.731	0.719	0.708
300	0.884	0.888	0.872	0.976	0.986	0.973
400	0.514	0.525	0.513	0.670	0.754	0.708
500	0.768	0.778	0.768	0.733	0.769	0.733

7. Discussion.

We have investigated the asymptotic properties, including the consistency, rate of convergence, and asymptotic normality for neural network sieve estimators with one hidden layer. While in practice, the number of hidden unites is often chosen ad hoc, it is important to note that the conditions in the theorems provide theoretical guidelines on choosing the number of hidden units for a neural network with one hidden layer to achieve the desired statistical properties. The validity of the conditions made in the theorems has also been checked through simulation results. Theorem 5.1 and Theorem 5.2 can be served as preliminary work for conducting hypothesis testing on $H_0 : f_0 = h_0$ for a fixed function h_0. However, there is currently no simple way to check conditions (C3) and (C4) in the theorem, which requires further researches on local entropy numbers for classes of neural networks.

The work conducted in this paper mainly focuses on sieve estimators based on neural networks with one hidden layer and standard sigmoid activation function. The work presented in this paper can be extended in several ways. The main theorems in this paper depend heavily on the covering number or the entropy number of the function class consisting of neural network with one hidden layer. Theorem 14.5 in Anthony and Bartlett (2009) provides a general upper bound for the covering number of a function class consisting of deep neural networks with Lipshitz continuous activation functions. Therefore, it is possible to extend our results discussed in this paper to a deep neural network with Lipshitz continuous activation functions. It is also worthwhile to investigate asymptotic properties of other commonly used deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

When we train a deep neural network, we usually need to face an overfitting issue. In practice, regularization is frequently used to reduce overfitting. Another natural extension of the work discussed in this paper is to modify the loss function by involving some regularization terms. By taking regularization into account, we believe the theories could have a much broader application in real world scenarios.

Acknowledgment. X. Shen, C. Jiang, and Q. Lu are supported by the National Institute on Drug Abuse (Award No. K01DA033346 and No. R01DA043501). L. Sakhanenko is supported by NSF grant DMS- 1742881.

Appendix. In this appendix, we are going to explore some basic properties of the parameter space $(\mathcal{F}, \|\cdot\|_n)$ discussed in the main text.

PROPOSITION 7.1. *The space $(\mathcal{F}, \|\cdot\|_n)$ is a pseudo-normed space.*

PROOF. Note that $\|f\|_n = \left(\frac{1}{n}\sum_{i=1}^n f^2(\boldsymbol{x}_i)\right)^{1/2}$.

(i) Based on the definition of $\|\cdot\|_n$, it is clear that $\|f\|_n \geq 0$, for any $f \in \mathcal{F}$.

(ii) For any $\lambda \in \mathbb{R}$ and $f \in \mathcal{F}$,

$$\|\lambda f\|_n = \left(\frac{1}{n}\sum_{i=1}^n \lambda^2 f^2(\boldsymbol{x}_i)\right)^{1/2} = |\lambda|\|f\|_n.$$

(iii) For any $f, g \in \mathcal{F}$,

$$\begin{aligned}\|f+g\|_n &= \left(\frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{x}_i) + g(\boldsymbol{x}_i))^2\right)^{1/2} = \left(\sum_{i=1}^n \left(\frac{1}{\sqrt{n}}f(\boldsymbol{x}_i) + \frac{1}{\sqrt{n}}g(\boldsymbol{x}_i)\right)^2\right)^{1/2} \\ &\leq \left(\sum_{i=1}^n \left(\frac{1}{\sqrt{n}}f(\boldsymbol{x}_i)\right)^2\right)^{1/2} + \left(\sum_{i=1}^n \left(\frac{1}{\sqrt{n}}g(\boldsymbol{x}_i)\right)^2\right)^{1/2} \\ &= \|f\|_n + \|g\|_n,\end{aligned}$$

where we applied the triangle inequality to the classical Euclidean norm.

Therefore, we showed that $(\mathcal{F}, \|\cdot\|_n)$ is a pseudo-normed space. □

Proposition 7.2. *There is an pseudo-inner product on $\mathcal{F}$ such that $\|f\|^2 = \langle f, f\rangle$ for any $f \in \mathcal{F}$. Moreover, the pseudo-inner product is given by*

$$\langle f, g\rangle = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)g(\boldsymbol{x}_i), \quad \forall f, g \in \mathcal{F}.$$

Proof. Based on the theorem attributed to Fréchet, von Neumann and Jordan (see for example, Proposition 14.1.2 in Blanchard and Brüning (2015)), to show the existence of the inner product, it suffices to check the parallelogram law of the pseudo-norm and the corresponding pseudo-inner product can be obtained via the polarization identity. To check to validity of the parallelogram law, we note that for any $f, g \in \mathcal{F}$,

$$\begin{aligned}
\|f+g\|_n^2 + \|f-g\|_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i)+g(\boldsymbol{x}_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i)-g(\boldsymbol{x}_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} f^2(\boldsymbol{x}_i) + \frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)g(\boldsymbol{x}_i) + \frac{1}{n}\sum_{i=1}^{n} g^2(\boldsymbol{x}_i) \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n} f^2(\boldsymbol{x}_i) - \frac{2}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)g(\boldsymbol{x}_i) + \frac{1}{n}\sum_{i=1}^{n} g^2(\boldsymbol{x}_i) \\
&= \frac{2}{n}\sum_{i=1}^{n} f^2(\boldsymbol{x}_i) + \frac{2}{n}\sum_{i=1}^{n} g^2(\boldsymbol{x}_i) \\
&= 2\|f\|_n^2 + 2\|g\|_n^2.
\end{aligned}$$

Hence, the parallelogram law is satisfied based on the pseudo-norm, and the pseudo-inner product does exist. By using the polarization identity, for any

$f, g \in \mathcal{F}$, we have

$$\begin{aligned}\langle f, g\rangle &= \frac{1}{4}\left(\|f+g\|_n^2 - \|f-g\|_n^2\right)\\ &= \frac{1}{4}\Bigg(\frac{1}{n}\sum_{i=1}^n f^2(\boldsymbol{x}_i) + \frac{2}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)g(\boldsymbol{x}_i) + \frac{1}{n}\sum_{i=1}^n g^2(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^n f^2(\boldsymbol{x}_i)\\ &\qquad + \frac{2}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)g(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^n g^2(\boldsymbol{x}_i)\Bigg)\\ &= \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i)g(\boldsymbol{x}_i).\end{aligned}$$

□

Let

$$\mathcal{G} = \left\{g : \mathbb{R} \to \mathbb{R}, \int \left|g'(z)\right| \mathrm{d}z \le M\right\}$$

be the class of functions of bounded variation in $\mathbb{R}$ (see Example 9.3.3 in van de Geer (2000)). The following proposition shows that $\mathcal{F}_{r_n} \subset \mathcal{G}$ for a fixed n.

PROPOSITION 7.3. *For a fixed n, $\mathcal{F}_{r_n} \subset \mathcal{G}$.*

PROOF. For any $f \in \mathcal{F}_{r_n}$, we have

$$f(x) = \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma\left(\gamma_j x + \gamma_{0,j}\right),$$

so that

$$f'(x) = \sum_{j=1}^{r_n} \alpha_j \gamma_j \sigma\left(\gamma_j x + \gamma_{0,j}\right)\left[1 - \sigma\left(\gamma_j x + \gamma_{0,j}\right)\right].$$

Without loss of generality, we assume that $\gamma_j \neq 0$ for $j = 1, \ldots, r_n$. Note that

$$\begin{aligned}\int |f'(x)|\mathrm{d}x &= \int \left|\sum_{j=1}^{r_n} \alpha_j \gamma_j \sigma\left(\gamma_j x + \gamma_{0,j}\right)\left[1 - \sigma\left(\gamma_j x + \gamma_{0,j}\right)\right]\right| \mathrm{d}x\\ &\le \sum_{j=1}^{r_n} |\alpha_j||\gamma_j| \int \sigma(\gamma_j x + \gamma_{0,j})[1 - \sigma(\gamma_j x + \gamma_{0,j})]\mathrm{d}x\\ &\le \sum_{j=1}^{r_n} |\alpha_j| \frac{|\gamma_j|}{\gamma_j} \int \sigma(u_j)(1 - \sigma(u_j))\mathrm{d}u_j,\end{aligned}$$

where in the last inequality, we let $u_j = \gamma_j x + \gamma_{0,j}$. Clearly, $|\gamma_j|/\gamma_j = \operatorname{sign}(\gamma_j)$. Moreover, since

$$\begin{aligned}\int \sigma(x)(1-\sigma(x))\mathrm{d}x &= \int \frac{e^x}{(1+e^x)^2}\mathrm{d}x \\ &= \int_0^\infty \frac{1}{(1+u)^2}\mathrm{d}u \quad \text{(by letting } u = e^x\text{)} \\ &= -(1+u)^{-1}\Big|_0^\infty \\ &= 1,\end{aligned}$$

for a fixed n, we have

$$\int |f'(x)|\mathrm{d}x \leq \sum_{j=1}^{r_n} |\alpha_j| \operatorname{sign}(\gamma_j) \leq \sum_{j=1}^{r_n} |\alpha_j| \leq V_n.$$

Therefore, $f \in \mathcal{G}$ and the desired result follows. □

References.

Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *The Annals of Probability* 1041–1067.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations.* cambridge university press.

Blanchard, P. and Brüning, E. (2015). *Mathematical methods in Physics: Distributions, Hilbert space operators, variational methods, and applications in quantum physics* **69**. Birkhäuser.

Boyd, S. and Mutapcic, A. (2008). Subgradient Methods (notes for EE364B Winter 2006-07, Stanford University).

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica* 289–314.

Devroye, L., Györfi, L. and Lugosi, G. (2013). *A probabilistic theory of pattern recognition* **31**. Springer Science & Business Media.

Fukumizu, K. (1996). A regularity condition of the information matrix of a multilayer perceptron network. *Neural networks* **9** 871–879.

Fukumizu, K. et al. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *The Annals of Statistics* **31** 833–851.

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

Grenander, U. (1981). *Abstract Inference.* Wily, New York.

Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks* **2** 359–366.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics* **31** 807–832.

Makovoz, Y. (1996). Random approximants and neural networks. *Journal of Approximation Theory* **85** 98–109.

Mendelson, S. (2003). A few notes on statistical learning theory. In *Advanced lectures on machine learning* 1–40. Springer.

Shen, X. (1997). On methods of sieves and penalization. *The Annals of Statistics* 2555–2591.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics* 580–615.

van de Geer, S. (2000). *Empirical Processes in M-estimation* **6**. Cambridge university press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Vapnik, V. (1998). *Statistical learning theory. 1998* **3**. Wiley, New York.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural computation* **1** 425–464.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural networks* **3** 535–549.

White, H. and Racine, J. (2001). Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates. *IEEE Transactions on Neural Networks* **12** 657–673.

White, H. and Wooldridge, J. (1991). Some results on sieve estimation with dependent observations. In *Nonparametric and Semiparametric Methods in Economics* (W. A. Barnett, J. Powell and G. Tauchen, eds.) 459–493. Cambridge University Press New York.

Wu, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *The Annals of Statistics* 501–513.

Zhu, H. and Zhang, H. (2006). Asymptotics for estimation and testing procedures under loss of identifiability. *Journal of Multivariate Analysis* **97** 19–45.

Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA
E-mail: shenxia4@stt.msu.edu

Department Biostatistics
University of Florida
Gainesville, FL 32611, USA
E-mail: jiangc@ufl.edu

Department of Statistics and Probability
Michigan State University
East Lansing, MI 48824, USA
E-mail: luda@stt.msu.edu

Department Biostatistics
University of Florida
Gainesville, FL 32611, USA
E-mail: lucienq@ufl.edu