

Asymptotic properties of neural network sieve estimators

Xiaoxi Shen, Chang Jiang, Lyudmila Sakhanenko & Qing Lu

To cite this article: Xiaoxi Shen, Chang Jiang, Lyudmila Sakhanenko & Qing Lu (2023): Asymptotic properties of neural network sieve estimators, Journal of Nonparametric Statistics, DOI: [10.1080/10485252.2023.2209218](https://doi.org/10.1080/10485252.2023.2209218)

To link to this article: <https://doi.org/10.1080/10485252.2023.2209218>



View supplementary material 



Published online: 13 May 2023.



Submit your article to this journal 



Article views: 48



View related articles 



View Crossmark data 



Asymptotic properties of neural network sieve estimators

Xiaoxi Shen^a, Chang Jiang^b, Lyudmila Sakhanenko^c and Qing Lu ^b

^aTexas State University, San Marcos, TX, USA; ^bUniversity of Florida, Gainesville, FL, USA; ^cMichigan State University, East Lansing, MI, USA

ABSTRACT

Neural networks have become one of the most popularly used methods in machine learning and artificial intelligence. Due to the universal approximation theorem, 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.

ARTICLE HISTORY

Received 6 June 2020

Accepted 20 April 2023

KEYWORDS


Empirical processes; entropy integral

1. Introduction

With the success of machine learning and artificial intelligence in research 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 minimisation 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, e.g., Vapnik 1998; Anthony and Bartlett 2009; Devroye et al. 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 inference based on neural networks. Throughout this paper, we will focus on the asymptotic properties of neural networks with one hidden layer.

CONTACT Qing Lu  lucienq@ufl.edu

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/10485252.2023.2209218>

One of the most important properties of neural networks is that they are universal approximators (Hornik et al. 1989), which means any continuous function on a 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. Many other series-based estimators such as splines and wavelets also possess similar approximation properties and have been extensively studied in the literature. Chen (2007) provides a comprehensive review for those methods. For nonparametric regression problems, random design and fixed design are the two main frameworks. Many existing literature on neural networks focus on random design (e.g. Chen and White 1999; Györfi et al. 2002). On the other hand, general theories on nonparametric regression under fixed design have been well studied in van de Geer (2000). Therefore, it is still worthwhile to study neural networks under fixed design.

Specifically, consider the following nonparametric regression model:

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

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. random variables defined on a complete probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{E}[\epsilon] = 0$, $\text{Var}[\epsilon] = \sigma^2$ and $\mathbb{E}[|\epsilon|^{2+\lambda}] < \infty$ for some $\lambda > 0$; $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X} \subset \mathbb{R}^d$ 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 minimises the population criterion function

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

A least squares estimator of the regression function can be obtained by minimising the empirical squared error loss $Q_n(f)$:

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

However, if the class of functions \mathcal{F} is too rich, the resulting least squares estimator may have undesired properties, such as inconsistency (Shen and Wong 1994; Shen 1997; van de Geer 2000). Instead, we can optimise 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,

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

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 $d(f, \pi_n f) \rightarrow 0$ as $n \rightarrow \infty$, 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

$$\mathbb{Q}_n(\hat{f}_n) \leq \inf_{f \in \mathcal{F}_n} \mathbb{Q}_n(f) + \mathcal{O}_p(\eta_n), \quad (1)$$

where $\eta_n \rightarrow 0$ as $n \rightarrow \infty$. We refer interested reader to Chen (2007) for a thorough discussion on sieve extremum estimators.

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

$$\mathcal{F}_{r_n} = \left\{ \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma(\gamma_j^T \mathbf{x} + \gamma_{0,j}) : \gamma_j \in \mathbb{R}^d, \alpha_j, \gamma_{0,j} \in \mathbb{R}, \right. \\ \left. \sum_{j=0}^{r_n} |\alpha_j| \leq V_n \text{ for some } V_n > 4 \text{ and } \max_{1 \leq j \leq r_n} \sum_{i=0}^d |\gamma_{i,j}| \leq M_n \text{ for some } M_n > 0 \right\}, \quad (2)$$

where $r_n, V_n, M_n \uparrow \infty$ as $n \rightarrow \infty$. For theoretical simplification, we impose boundedness assumption on the weights of the neural networks in \mathcal{F}_{r_n} , which is related to the ℓ_1 -regularisation when fitting a neural network. Such a method has been discussed in previous literatures (e.g. White 1989, 1990). In those papers, the 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. In terms of the rate of convergence, Chen and Shen (1998) obtained rate of convergence $\mathcal{O}_p((\frac{n}{\log n})^{-\frac{1+1/d}{4(1+1/(2d))}})$ for neural network sieve estimators.

Later on, Chen and White (1999) improved the convergence rate to $\mathcal{O}_p((\frac{n}{\log n})^{-\frac{1+2\alpha/(d+1)}{4(1+\alpha/(d+1))}})$, where α relates to the smoothness of the true function f_0 and in their paper, a central limit theorem for smooth functional of the estimated function is also provided. In Chen et al. (2001), rate of convergence was also obtained for stationary β -mixing data. One important characteristic of the aforementioned rate of convergence is that as $d \rightarrow \infty$, both rates become $\mathcal{O}_p((\frac{n}{\log n})^{-1/4})$. Similar results can also be found in Barron (1994). It is well known from Stone (1982) that local smoothing methods suffer from the curse of dimensionality. But such phenomenon seems to vanish for approximate sieve extremum estimators based on neural networks. Bauer and Kohler (2019) also developed general theory and conditions to justify that neural networks can be used to circumvent the issue of curse of dimensionality. Hornik et al. (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 $\|f\|_n^2 = n^{-1} \sum_{i=1}^n f^2(\mathbf{x}_i)$ (see Proposition 2.1 in the supplementary material) defined on \mathcal{F} and \mathcal{F}_{r_n} .

With the increasing popularity of deep learning, recent research also starts to focus on the statistical properties of deep neural networks. For example, Schmidt-Hieber (2020) provided the rate of convergence for sparse deep neural network with Rectified Linear Unit (ReLU) activation function under the assumption that the underlying function is a composition of functions in some Hölder space. Farrell et al. (2020) and Farrell et al. (2021) established the rate of convergence for deep neural network estimators when the underlying function belongs to a unit ball in a Sobolev space. It is worth pointing out that in these

two papers, no restrictions on the boundedness of weights in neural networks are imposed. The rate of convergence for deep ReLU neural networks has also been developed in Fabozzi et al. (2019) and Kohler and Langer (2021). It is worth mentioning that the results in Fabozzi et al. (2019) were developed based on the original version of this manuscript on ArXiv. We believe that the ultimate goal of developing these theories is to perform statistical inference based on neural networks for real-world problems and the results discussed in this paper may provide a starting point for further developments. For instance, Chen and White (1999) developed asymptotic normality for neural network sieve extremum estimators. However, to apply their result, a calculation of a series of covariance is essential, which may be hard to accomplish in practice. Recently, Horel and Giesecke (2020) developed a significance test based on neural networks. However, the theories in that paper are difficult to apply and verify in practice. Using similar techniques to be discussed in this paper, Shen et al. (2021) developed a goodness-of-fit test based on neural networks.

The remaining paper is organised as follows. 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 Sections 3 and 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 this 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 $\frac{a_n}{b_n} \rightarrow 1$ as $n \rightarrow \infty$. 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 $H(\epsilon, T, d)$.

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 $1/4$.*

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 ,*

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

Now we quote a general result from White and Wooldridge (1991) for readers who are not familiar with the theories of sieve extremum estimators. The theorem tells us that under some mild conditions, there exists a sieve approximate estimator and such an estimator is also measurable.

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 $\mathbb{Q}_n : \Omega \times \Theta_n \rightarrow \bar{\mathbb{R}}$ 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 , $n = 1, 2, \dots$. Then for each $n = 1, 2, \dots$, there exists $\hat{\theta}_n : \Omega \rightarrow \Theta_n$, $\mathcal{A}/\mathcal{B}(\Theta_n)$ -measurable such that for each $\omega \in \Omega$, $\mathbb{Q}_n(\omega, \hat{\theta}_n(\omega)) = \inf_{\theta \in \Theta_n} \mathbb{Q}_n(\omega, \theta)$.

Note that

$$\begin{aligned} \mathbb{Q}_n(f) &= \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 \\ &= \frac{1}{n} \sum_{i=1}^n (f_0(\mathbf{x}_i) + \epsilon_i - f(\mathbf{x}_i))^2 \\ &= \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2 - 2 \frac{1}{n} \sum_{i=1}^n \epsilon_i (f(\mathbf{x}_i) - f_0(\mathbf{x}_i)) + \frac{1}{n} \sum_{i=1}^n \epsilon_i^2. \end{aligned}$$

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.

Proof: For each fixed n , let $\boldsymbol{\theta}_n = [\alpha_0, \dots, \alpha_{r_n}, \gamma_{0,1}, \dots, \gamma_{0,r_n}, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_{r_n}^T]^T$ 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

$$\begin{aligned} H : (\Theta_n, \|\cdot\|_2) &\rightarrow (\mathcal{F}_{r_n}, \|\cdot\|_n) \\ \boldsymbol{\theta}_n &\mapsto H(\boldsymbol{\theta}_n) = \alpha_0 + \sum_{j=1}^{r_n} \alpha_j \sigma \left(\boldsymbol{\gamma}_j^T \mathbf{x} + \gamma_{0,j} \right) \end{aligned}$$

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

$$\begin{aligned} &\|H(\boldsymbol{\theta}_{1,n}) - H(\boldsymbol{\theta}_{2,n})\|_n^2 \\ &= \frac{1}{n} \sum_{i=1}^n \left[\alpha_0^{(1)} + \sum_{j=1}^{r_n} \alpha_j^{(1)} \sigma \left(\boldsymbol{\gamma}_j^{(1)T} \mathbf{x}_i + \gamma_{0,j}^{(1)} \right) - \alpha_0^{(2)} - \sum_{j=1}^{r_n} \alpha_j^{(2)} \sigma \left(\boldsymbol{\gamma}_j^{(2)T} \mathbf{x}_i + \gamma_{0,j}^{(2)} \right) \right]^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \left[\left| \alpha_0^{(1)} - \alpha_0^{(2)} \right| + \sum_{j=1}^{r_n} \left| \alpha_j^{(1)} \sigma \left(\boldsymbol{\gamma}_j^{(1)T} \mathbf{x}_i + \gamma_{0,j}^{(1)} \right) - \alpha_j^{(2)} \sigma \left(\boldsymbol{\gamma}_j^{(2)T} \mathbf{x}_i + \gamma_{0,j}^{(2)} \right) \right| \right]^2 \end{aligned}$$

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

Hence, for any $\epsilon > 0$, we choose $\delta = \epsilon / (\frac{V_n}{4} (1 \vee \|\mathbf{x}\|_\infty) \sqrt{r_n(d+1)})$. 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 $n = 1, 2, \dots$, there exists $\hat{f}_n : \Omega \rightarrow \mathcal{F}_{r_n}$, $\mathcal{A}/\mathcal{B}(\mathcal{F}_{r_n})$ -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.

Lemma 3.1: *Under the assumption of*

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

we have

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

Proof: For any $\delta > 0$, we have

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

It follows from the Weak Law of Large Numbers that $(I) \rightarrow 0$. Now, we are going to evaluate (II) . By using the Markov's inequality, $(II) \rightarrow 0$ holds if

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

Note that $\mathbb{E}[\epsilon] = 0$ and each $f \in \mathcal{F}_{r_n}$ has its corresponding parameterisation $\boldsymbol{\theta}_n$. Since $\boldsymbol{\theta}_n$ is in a compact set, there exists a sequence $\boldsymbol{\theta}_{n,k} \rightarrow \boldsymbol{\theta}_n$ as $k \rightarrow \infty$ 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 $f_k(\mathbf{x}) \rightarrow f(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$. From Example 2.3.4 in van der Vaart and Wellner (1996), we know that \mathcal{F}_{r_n} is P -measurable. Based on symmetrisation inequality, we have

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

where ξ_1, \dots, ξ_n are i.i.d. Rademacher random variables independent of $\epsilon_1, \dots, \epsilon_n$. Based on the Strong Law of Large Numbers, there exists $N_1 > 0$, such that for all $n \geq N_1$,

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

For fixed $\epsilon_1, \dots, \epsilon_n$, $\sum_{i=1}^n \xi_i \epsilon_i (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))$ is a sub-Gaussian process indexed by $f \in \mathcal{F}_{r_n}$. Suppose that (Ξ, \mathcal{C}, μ) is the probability space on which ξ_1, \dots, ξ_n are defined and let $Y(f, \omega) = \sum_{i=1}^n \xi_i(\omega) \epsilon_i (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))$ with $f \in \mathcal{F}_{r_n}$ and $\omega \in \Xi$. As we have shown above, we have $f_k \rightarrow f$ and by continuity, $Y(f_k, \omega) \rightarrow Y(f, \omega)$ 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 $n \geq N_1$,

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

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

where for $f, g \in \mathcal{F}_{r_n}$,

$$d(f, g) = \left(\frac{1}{n} \sum_{i=1}^n \epsilon_i^2 (f(\mathbf{x}_i) - g(\mathbf{x}_i))^2 \right)^{1/2}$$

so that the last inequality follows by noting that $\sup_{f \in \mathcal{F}_{r_n}} \|f\|_\infty \leq V_n$ and

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

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

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

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

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

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

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

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 2:

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

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

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

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 $\log \frac{[r_n(d+2)+1]V_n^2}{V_n-4} \geq \log \frac{2\sqrt{\sigma^2+1}(V_n-4)}{V_n-4} = \log(2\sqrt{\sigma^2 + 1})$. We also have

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

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

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

which implies that

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

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

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

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

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

Since

$$\begin{aligned} &\mathbb{E}_\epsilon [\sqrt{\sigma^2 + 1} \|\pi_{r_n} f_0 - f_0\|_\infty + 4\sqrt{2}KB_{r_n, d, V_n}^{1/2} n^{-1/2} V_n] \\ &= \sqrt{\sigma^2 + 1} \|\pi_{r_n} f_0 - f_0\|_\infty + 4\sqrt{2}KB_{r_n, d, V_n}^{1/2} n^{-1/2} V_n \rightarrow 0 < \infty, \end{aligned}$$

by using the Generalised Dominated Convergence Theorem, we know that

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

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*

$$[r_n(d+2) + 1]V_n^2 \log(V_n[r_n(d+2) + 1]) = o(n), \quad \text{as } n \rightarrow \infty, \quad (3)$$

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

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

Hence, based on Lemma 1, Lemma 3 and Corollary 2.6 in White and Wooldridge (1991), we have

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

Remark 3.1: We discuss the condition (3) 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 $r_n = o((n/\log n)^{1/3})$. 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 $V_n = o(n^{1/2}/\log n)$.

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).

Theorem 4.1: *Based on the above notations, if*

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

then

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

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

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

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 $\delta < 1$,

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

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

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

we have

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

Therefore, for

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

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

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

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 $r_n(d+2)V_n^2 \log[r_n V_n(d+2)] = o(n)$. Under such a condition, $r_n(d+2) \log[r_n V_n(d+2)] \leq n$, 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.

$$\begin{aligned} \mathcal{F} &= \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} : f(\mathbf{x}) = \int \exp \{ i \mathbf{a}^T \mathbf{x} \} d\mu_f(\mathbf{a}), \|\mu_f\|_1 \right. \\ &\quad \left. := \int \max(\|\mathbf{a}\|_1, 1) d|\mu_f|(\mathbf{a}) \leq C \right\}, \end{aligned} \quad (4)$$

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 . $\|\mathbf{a}\|_1 = \sum_{i=1}^d |a_i|$ for $\mathbf{a} = [a_1, \dots, a_d]^T \in \mathbb{R}^d$. 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:

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

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 $\|\hat{f}_n - f_0\|_n = \mathcal{O}_p((n/\log n)^{-1/3})$. 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 3 in the supplementary material, \mathcal{F}_{r_n} is a class of functions with bounded variation in \mathbb{R} . Therefore, the convergence rate we obtained makes sense.

5. Asymptotic normality

To establish the asymptotic normality of sieve estimator for neural networks, we follow the idea in Shen (1997) and start by calculating the Gâteaux derivative of the empirical

criterion function $\mathbb{Q}_n(f) = n^{-1} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$,

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

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

$$R_n[f - f_0] = \mathbb{Q}_n(f) - \mathbb{Q}_n(f_0) - \mathbb{Q}'_{n,f_0}[f - f_0] = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2 = \|f - f_0\|_n^2.$$

As will be seen in the proof of asymptotic normality, the rate of convergence for the empirical process $\{n^{-1/2} \sum_{i=1}^n \epsilon_i (f(\mathbf{x}_i) - f_0(\mathbf{x}_i)) : f \in \mathcal{F}_{r_n}\}$ 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, \dots, X_n be independent random variables with $X_i \sim P_i$. Define the empirical process $\{v_n(f)\}$ as*

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

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

$$H(u, \mathcal{F}_n, \|\cdot\|_\infty) \leq A_n u^{-r}, \quad (5)$$

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

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

we have

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

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

some $0 < r < 2$, we have

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

Based on the assumption of

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

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

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

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 $\tilde{K}_1 = 2^9 \epsilon^{-3/2} (2-r)^{-1} (\frac{4}{\epsilon})^{\frac{1}{r}-\frac{1}{2}} (\frac{3}{4})^{\frac{1}{2}-\frac{1}{r}}$. 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 $\tilde{K}_2 = 2^9 \epsilon^{-3/2} (2-r)^{-1} (\frac{4}{\epsilon})^{\frac{1}{r}-\frac{1}{2}} (\frac{1}{4})^{\frac{1}{2}-\frac{1}{r}}$. 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 $2^8 \epsilon^{-3/2} I(\frac{\epsilon M}{64\sqrt{n}}, t_0) < M$. By Theorem 2.1 in Alexander (1984), we have the desired result. \blacksquare

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

$$(C1) \quad r_n(d+2)V_n \log[r_n V_n(d+2)] = o(n^{1/4});$$

$$(C2) \quad n\rho_n^{-2}/V_n^\lambda = o(1),$$

we have

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

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

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

For (I), we can apply Lemma 6 directly. Note that $|\epsilon_i \mathbb{I}_{\{|\epsilon_i| \leq V_n\}} (f - f_0)(\mathbf{x})| \leq V_n(V_n + \|f_0\|_\infty) \lesssim V_n^2$ since $\|f_0\|_\infty < \infty$ and for $0 < \eta < 1$,

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

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

$$C_{r_n, d, V_n} = [r_n(d+2) + 1] \log \tilde{C}_{r_n, d, V_n} - [r_n(d+2) + 1] \sim r_n(d+2) \log[r_n V_n(d+2)].$$

Therefore, Equation (5) in the main text is satisfied with $r = 1$ and $A_n = 2C_{r_n, d, V_n}$. Following from Lemma 6, 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

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

From (C1),

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

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

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

Then it follows from the Markov inequality that

$$\begin{aligned} (II) &\leq \mathbb{P} \left(\left(\frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \mathbb{I}_{\{|\epsilon_i| > V_n\}} \right)^{1/2} \rho_n^{-1} \gtrsim M n^{-1/2} \right) \lesssim M^{-2} n \rho_n^{-2} \mathbb{E}[\epsilon^2 \mathbb{I}_{|\epsilon| > V_n}] \\ &\lesssim M^{-2} n \rho_n^{-2} \frac{\mathbb{E}[|\epsilon|^{2+\lambda}]}{V_n^\lambda}. \end{aligned}$$

Based on condition (C2), we have $(II) \rightarrow 0$, and

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

Combining the results we obtained above, we get

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

■

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

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

then

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

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 $r_n(d+2) \log n/V_n^\lambda \rightarrow 0$.

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

$$\tilde{f}_n(f) = (1 - \delta_n)f + \delta_n(f_0 + \iota), \quad (6)$$

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

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:

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

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

then

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

Proof: 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 $f \in \{f \in \mathcal{F}_{r_n} : \|f - f_0\|_n \leq$

$\rho_n^{-1}\}$,

$$\begin{aligned}\mathbb{Q}_n(f) &= \mathbb{Q}_n(f_0) + \mathbb{Q}'_{n,f_0}[f - f_0] + R_n[f - f_0] \\ &= \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 - \frac{2}{n} \sum_{i=1}^n \epsilon_i (f(\mathbf{x}_i) - f_0(\mathbf{x}_i)) + \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2.\end{aligned}\quad (7)$$

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 (7) by \hat{f}_n and $\pi_{r_n}\tilde{f}_n(f)$, we get

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

Subtracting these two equations yields

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

Now note that

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

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

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

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

$$\begin{aligned}
-\mathcal{O}_p(\delta_n^2) &\leq \inf_{f \in \mathcal{F}_{r_n}} \mathbb{Q}_n(f) - \mathbb{Q}_n(\hat{f}_n) \leq \mathbb{Q}_n(\pi_{r_n} \tilde{f}_n(f)) - \mathbb{Q}_n(\hat{f}_n) \\
&\leq (1 - \delta_n)^2 \|\hat{f}_n - f_0\|_n^2 - \|\hat{f}_n - f_0\|_n^2 + 2(1 - \delta_n) \delta_n \langle \hat{f}_n - f_0, \iota \rangle \\
&\quad + 2(1 - \delta_n) \|\hat{f}_n - f_0\|_n \|\pi_{r_n} \tilde{f}_n(f) - \tilde{f}_n(f)\|_n \\
&\quad + 2\delta_n \|\pi_{r_n} \tilde{f}_n(f) - \tilde{f}_n(f)\|_n + \|\pi_{r_n} \tilde{f}_n(f) - \tilde{f}_n(f)\|_n^2 \\
&\quad - \frac{2}{n} \sum_{i=1}^n \epsilon_i \left(\pi_{r_n} \tilde{f}_n(f)(\mathbf{x}_i) - \tilde{f}_n(f)(\mathbf{x}_i) \right) + \frac{2}{n} \delta_n \sum_{i=1}^n \epsilon_i \left(\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) \right) \\
&\quad + \frac{2}{n} \delta_n \sum_{i=1}^n \epsilon_i + \mathcal{O}_p(\delta_n^2) \\
&\leq \delta_n^2 \|\hat{f}_n - f_0\|_n^2 + 2(1 - \delta_n) \delta_n \langle \hat{f}_n - f_0, \iota \rangle \\
&\quad + 2(1 - \delta_n) \|\hat{f}_n - f_0\|_n \|\pi_{r_n} \tilde{f}_n(f) - \tilde{f}_n(f)\|_n \\
&\quad + 2\delta_n \|\pi_{r_n} \tilde{f}_n(f) - \tilde{f}_n(f)\|_n + \|\pi_{r_n} \tilde{f}_n(f) - \tilde{f}_n(f)\|_n^2 \\
&\quad - \frac{2}{n} \sum_{i=1}^n \epsilon_i \left(\pi_{r_n} \tilde{f}_n(f)(\mathbf{x}_i) - \tilde{f}_n(f)(\mathbf{x}_i) \right) + \frac{2}{n} \delta_n \sum_{i=1}^n \epsilon_i \left(\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) \right) \\
&\quad + \frac{2}{n} \delta_n \sum_{i=1}^n \epsilon_i + \mathcal{O}_p(\delta_n^2), \tag{8}
\end{aligned}$$

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

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

Combining with Theorem 2, 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

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

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

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

It follows from these observations that

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

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. ■

Let us focus on the conditions given in the theorem. Note that if (C1) holds, we have

$$r_n(d+2)V_n^2 \log[r_n V_n(d+2)] \leq [r_n(d+2)]^4 V_n^4 (\log[r_n V_n(d+2)])^4 = o(n),$$

so it is a sufficient condition to ensure the consistency of the neural network sieve estimator. As in Remark 3.1, we consider some simple scenarios here. If $V_n = \mathcal{O}(r_n)$, then $r_n(d+2)V_n \log[r_n V_n(d+2)] = \mathcal{O}(r_n^2 \log r_n)$ so that a possible growth rate for r_n is $r_n = o(n^{1/8}/(\log n)^2)$. On the other hand, if $r_n = \log V_n$, then $r_n(d+2)V_n \log[r_n V_n(d+2)] = \mathcal{O}(V_n(\log V_n)^2)$ 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 3.1. 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 $\sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$ 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 $\sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$, 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.

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

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

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

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 $0 \leq \eta_n = o(n^{-1})$. Then under the conditions as stated in Theorem 5.1, we have

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

Proof: Note that

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

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

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

Under (C3), $\|\pi_{r_n} f_0 - f_0\|_n^2 = o(\rho_n^2 \delta_n^4) = o(n^{-1/2})$ 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 $\|\hat{f}_n - f_0\|_n^2 = o_p(n^{-1/2})$. Moreover, by the same arguments as in the proof of Theorem 5.1, we can show that

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

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 in the main text, we obtain

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

6. Simulation studies

In this section, simulations were conducted to check the validity of the theoretical results obtained in the previous sections. 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. For illustration purpose, we only include the simulations where the dimension of the covariates is 1. More simulations for the multivariate cases are given in the supplementary materials.

6.1. Consistency for neural network sieve estimators

In this simulation, we are going to check the consistency result from Section 3 and the 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 α : $\sum_{i=0}^{r_n} |\alpha_i| \leq V_n$. So finding the nearly optimal function in \mathcal{F}_{r_n} for $\mathbb{Q}_n(f)$ is in fact a constrained optimisation problem. A classical way to conduct this optimisation 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 Mutapicic (2008). The basic idea is to update the parameter $\alpha_0, \dots, \alpha_{r_n}$ through

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

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 $\alpha^{(k)}$. More specifically, we take

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

Table 1. 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

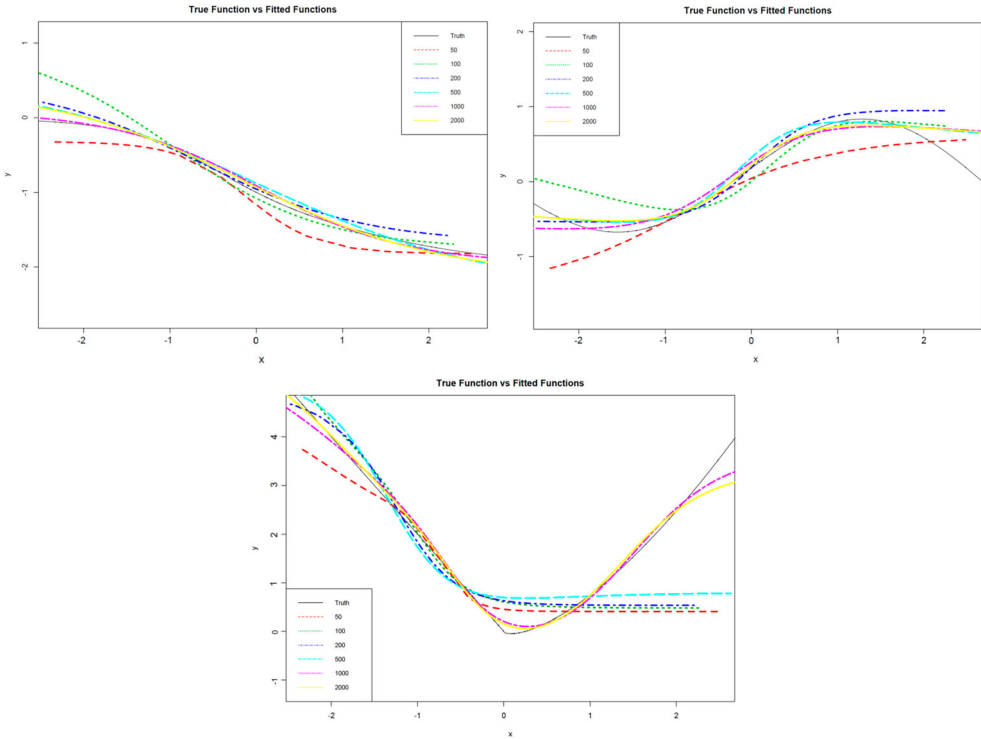


Figure 1. Comparison of the true function and the fitted function for three different types of non-linear functions. The top panel shows the scenario when the true function is a single layer neural network; the middle panel shows the scenario when the true function is a sine function, and the bottom 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.

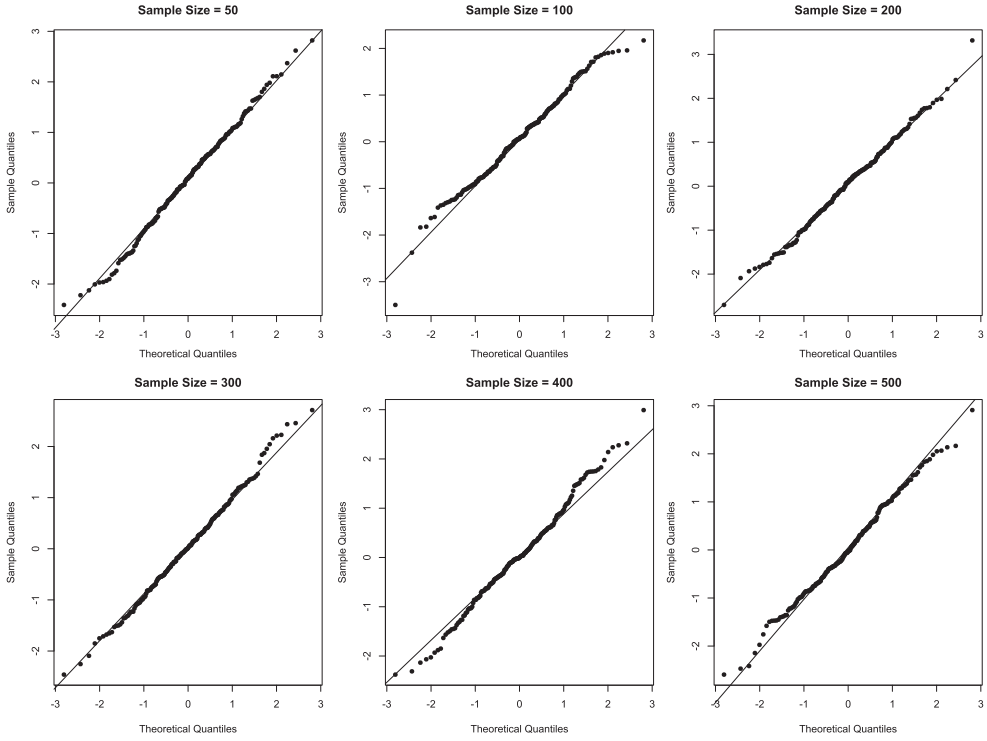


Figure 2. Normal Q-Q plot for $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(x_i) - f_0(x_i)]$ various sample sizes. The true function f_0 is a single-layer neural network with two hidden units as defined in (10).

The updating equations of $\gamma_1, \dots, \gamma_{r_n}, \gamma_{0,1}, \dots, \gamma_{0,r_n}$ remain the same as those in the classical gradient descent algorithm.

We simulated the response through the following model:

$$y_i = f_0(x_i) + \epsilon_i, \quad i = 1, \dots, n, \quad (9)$$

where the total sample size n varies from 50 to 2000, $x_1, \dots, x_n \sim \text{i.i.d. } \mathcal{N}(0, 1)$, $\epsilon_1, \dots, \epsilon_n \sim \text{i.i.d. } \mathcal{N}(0, 0.7^2)$. For the true function $f_0(x)$, we considered the following three functions:

- (1) A neural network with a single hidden layer and two hidden units:

$$f_0(x_i) = -1 - \sigma(2x_i + 1) + \sigma(-x_i + 1). \quad (10)$$

- (2) A trigonometric function:

$$f_0(x) = \sin\left(\frac{\pi}{3}x\right) + \frac{1}{3} \cos\left(\frac{\pi}{4}x + 1\right) \quad (11)$$

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

$$f_0(x) = \begin{cases} -2x & \text{if } x \leq 0 \\ \sqrt{x} \left(x - \frac{1}{4}\right) & \text{if } x > 0. \end{cases} \quad (12)$$

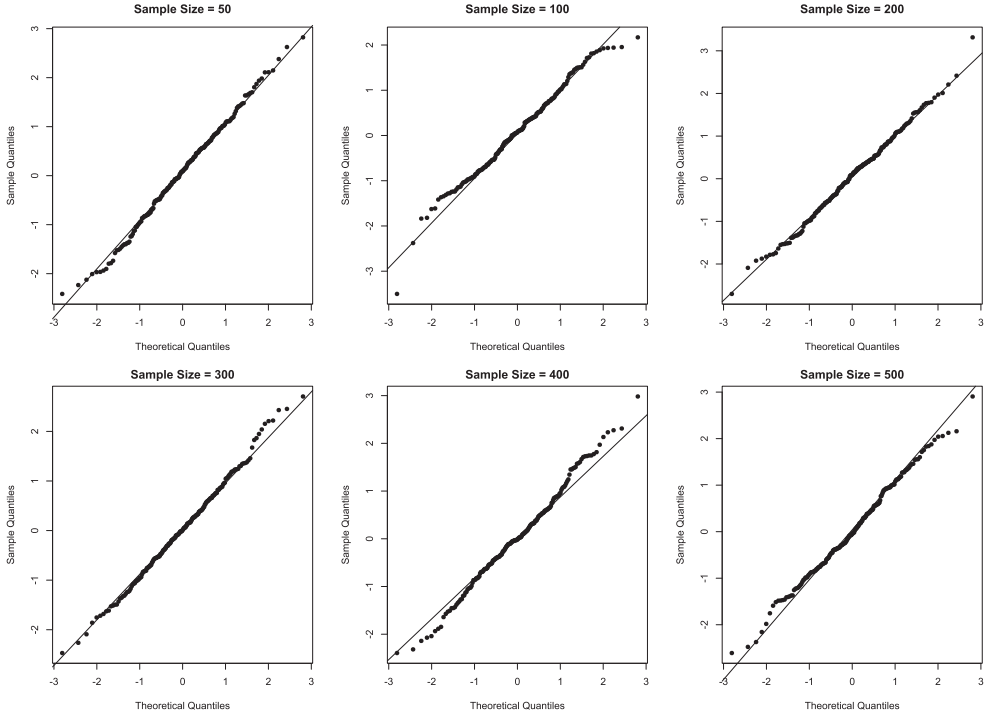


Figure 3. Normal Q-Q plot for $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(x_i) - f_0(x_i)]$ various sample sizes. The true function f_0 is a trigonometric function as defined in (11).

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 3.1 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 $Q_n(\hat{f}_n)$ under different sample sizes. The results are summarised in Table 1.

As we can see from Table 1, 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 increases 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 1 plots the fitted functions and the true function, from which we can straightforwardly visualise the result more and draw the conclusions.

6.2. 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 6.1 but sampled

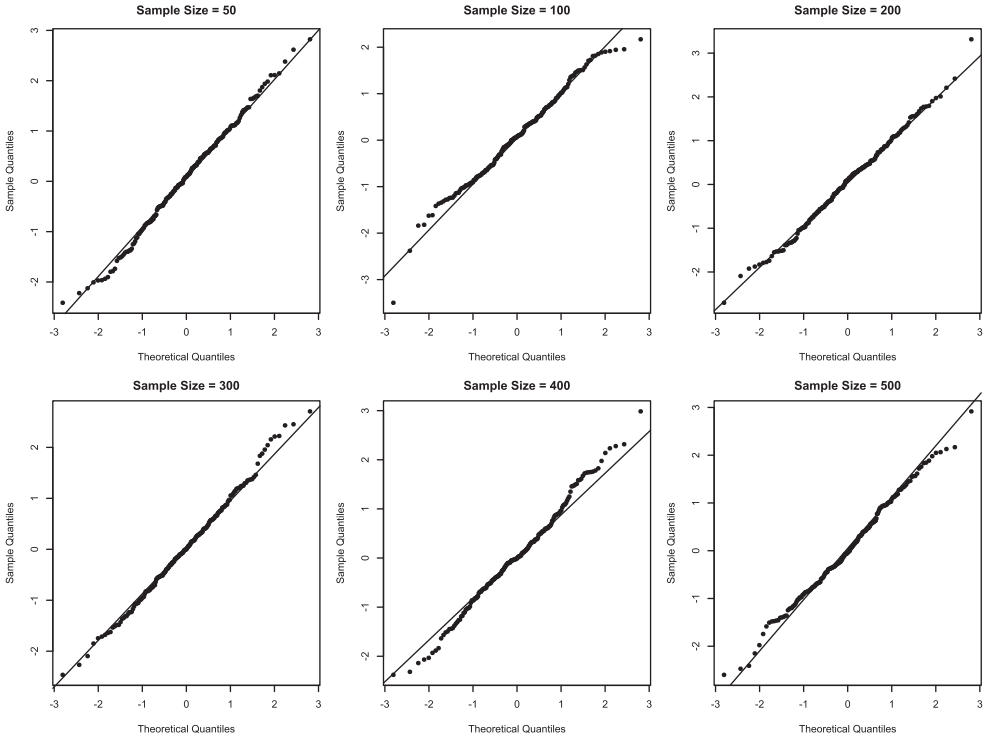


Figure 4. Normal Q–Q plot for $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$ various sample sizes. The true function f_0 is a continuous function having a non-differential point as defined in (12).

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 6.1 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. To get the normal Q–Q plot for $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$, we repeated the simulation 200 times.

Figures 2 to 4 are the normal Q–Q plots under different nonlinear functions and various sample sizes. From the figures, we found that the statistic $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$ 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 $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$ is $\mathcal{N}(0, 1)$ under all scenarios. Another implication we can obtain from the Q–Q plots is that the statistic $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$ 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 $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$, which facilitates hypothesis testing.

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

Table 2. 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 (10), a trigonometric function described in (11) and a continuous function having a non-differential point described in (12), 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

all cases, we failed to reject that $n^{-1/2} \sum_{i=1}^n [\hat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i)]$ follows the standard normal distribution.

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 units 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. Theorems 5.1 and 5.2 depend on the knowledge of the underlying function f_0 , which is typically unknown in practice. Therefore, if we assume f_0 has some certain form, the results can be applied and served as preliminary work for conducting hypothesis testing on $H_0 : f_0 = h_0$ for a fixed function h_0 . On the other hand, since multiple functions can lead to the same value of $n^{-1} \sum_{i=1}^n f_0(\mathbf{x}_i)$, the test may not be power. The asymptotic normality results are crucial in developing more sophisticated significance test methods for neural networks (Shen et al. 2022).

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 Lipschitz continuous activation functions. Therefore, it is possible to extend our results discussed in this paper to a deep neural network with Lipschitz 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).

On the other hand, although homoscedasticity is assumed in the previous discussions, it can be relaxed to take heteroscedasticity into consideration. To see this, if $\mathbb{E}[\epsilon_i^2] = \phi^2 \zeta^2(\mathbf{x}_i)$, then under the assumptions that $|\zeta(\cdot)| \leq \varsigma$ for some $\varsigma > 0$ and $\sum_i \text{Var}[\epsilon_i^2]/i^2 < \infty$, the proof of Lemma 3.1 can go through. The only modifications to be made are to use Kolmogorov Strong Law of Large Numbers to show $(I) \rightarrow 0$ and to change σ^2 to $\varsigma^2 \phi^2$.

later on. Therefore, the consistency result still holds under heteroskedasticity with the aforementioned two assumptions satisfied. Moreover, after a clear examination on the proof of Theorems 4.1, 5.1 and 5.2, it is easy to see that only the consistency part is involved with heteroskedasticity. Therefore, these results still hold under the aforementioned two assumptions.

When we train a deep neural network, we usually need to face an overfitting issue. In practice, regularisation 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 regularisation terms. By taking regularisation into account, we believe the theories could have a much broader application in real-world scenarios.

Acknowledgments

The authors are grateful to the insights and suggestions from two referees to make this manuscript better.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

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-1208238.

ORCID

Qing Lu  <http://orcid.org/0000-0002-7943-966X>

References

- Alexander, K.S. (1984), 'Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm', *The Annals of Probability*, 12(4), 1041–1067.
- Anthony, M., and Bartlett, P.L. (2009), *Neural Network Learning: Theoretical Foundations*, Cambridge University Press. <https://www.stat.berkeley.edu/~bartlett/nnl/index.html>
- Barron, A.R. (1994), 'Approximation and Estimation Bounds for Artificial Neural Networks', *Machine Learning*, 14(1), 115–133.
- Bauer, B., and Kohler, M. (2019), 'On Deep Learning As a Remedy for the Curse of Dimensionality in Nonparametric Regression', *The Annals of Statistics*, 47(4), 2261–2285.
- Boyd, S., and Mutapcic, A. (2008), 'Subgradient Methods (Notes for EE364B Winter 2006-07, Stanford University)'.
- Chen, X. (2007), 'Large Sample Sieve Estimation of Semi-nonparametric Models', *Handbook of Econometrics*, 6, 5549–5632.
- Chen, X., Racine, J., and Swanson, N.R. (2001), 'Semiparametric Arx Neural-network Models with an Application to Forecasting Inflation', *IEEE Transactions on Neural Networks*, 12(4), 674–683.
- Chen, X., and Shen, X. (1998), 'Sieve Extremum Estimates for Weakly Dependent Data', *Econometrica*, 66(2), 289–314.
- Chen, X., and White, H. (1999), 'Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators', *IEEE Transactions on Information Theory*, 45(2), 682–691.
- Devroye, L., Györfi, L., and Lugosi, G. (2013), *A Probabilistic Theory of Pattern Recognition*, Vol. 31. New York, NY: Springer Science & Business Media.

- Fabozzi, F.J., Fallahgoul, H., Franstianto, V., and Loeper, G. (2019), 'Towards Explaining Deep Learning: Asymptotic Properties of relu ffn Sieve Estimators', Available at SSRN 3499324.
- Farrell, M.H., Liang, T., and Misra, S. (2020), 'Deep Learning for Individual Heterogeneity: an Automatic Inference Framework', Preprint [arXiv:2010.14694](https://arxiv.org/abs/2010.14694).
- Farrell, M.H., Liang, T., and Misra, S. (2021), 'Deep Neural Networks for Estimation and Inference', *Econometrica*, 89(1), 181–213.
- Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016), *Deep Learning*, Vol. 1, MIT Press Cambridge. <https://www.deeplearningbook.org/>
- Grenander, U. (1981), *Abstract Inference*, New York: Wiley.
- Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002), *A Distribution-Free Theory of Nonparametric Regression*, Vol. 1. New York, NY: Springer.
- Horel, E., and Giesecke, K. (2020), 'Significance Tests for Neural Networks', *Journal of Machine Learning Research*, 21(227), 1–29.
- Hornik, K., Stinchcombe, M., and White, H. (1989), 'Multilayer Feedforward Networks Are Universal Approximators', *Neural Networks*, 2(5), 359–366.
- Kohler, M., and Langer, S. (2021), 'On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates', *The Annals of Statistics*, 49(4), 2231–2249.
- Makovoz, Y. (1996), 'Random Approximants and Neural Networks', *Journal of Approximation Theory*, 85(1), 98–109.
- Mendelson, S. (2003), 'A few notes on statistical learning theory', in *Advanced Lectures on Machine Learning*, Springer, pp. 1–40. https://link.springer.com/chapter/10.1007/3-540-36434-X_1
- Schmidt-Hieber, J. (2020), 'Nonparametric Regression Using Deep Neural Networks with Relu Activation Function', *The Annals of Statistics*, 48(4), 1875–1897.
- Shen, X. (1997), 'On Methods of Sieves and Penalization', *The Annals of Statistics*, 25(6), 2555–2591.
- Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2021), 'A Goodness-of-fit Test Based on Neural Network Sieve Estimators', *Statistics & Probability Letters*, 174, 109100.
- Shen, X., Jiang, C., Sakhanenko, L., and Lu, Q. (2022), 'A Sieve Quasi-likelihood Ratio Test for Neural Networks with Applications to Genetic Association Studies', Preprint [arXiv:2212.08255](https://arxiv.org/abs/2212.08255).
- Shen, X., and Wong, W.H. (1994), 'Convergence Rate of Sieve Estimates', *The Annals of Statistics*, 580–615. <https://www.jstor.org/stable/2242281>
- Stone, C.J. (1982), 'Optimal Global Rates of Convergence for Nonparametric Regression', *The Annals of Statistics*, 1040–1053. <https://www.jstor.org/stable/2240707>
- van de Geer, S. (2000), *Empirical Processes in M-estimation*, Vol. 6, Cambridge University Press. <https://www.cambridge.org/us/academic/subjects/statistics-probability/statistical-theory-and-methods/empirical-processes-mestimation?format=PB&isbn=9780521123259>
- van der Vaart, A.W., and Wellner, J.A. (1996), *Weak Convergence and Empirical Processes*. New York: Springer.
- Vapnik, V. (1998), *Statistical Learning Theory*, Vol. 3, New York: Wiley.
- White, H. (1989), 'Learning in Artificial Neural Networks: a Statistical Perspective', *Neural Computation*, 1(4), 425–464.
- White, H. (1990), 'Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings', *Neural Networks*, 3(5), 535–549.
- White, H., and Wooldridge, J. (1991), 'Some results on sieve estimation with dependent observations', in *Nonparametric and Semiparametric Methods in Economics*, ed. W. Barnett, J. Powell, and G. Tauchen, New York: Cambridge University Press, pp. 459–493.