Towards explaining the ReLU Feed-Forward Network*

Hasan Fallahgoul† Vincentius Franstianto‡ Grégoire Loeper§

Monash University Monash University Monash University

December 29, 2019
Preliminary and incomplete

Abstract

A multi-layer, multi-node ReLU network is a powerful, efficient, and popular tool in statistical prediction tasks. However, in contrast to the great emphasis on its empirical applications, its statistical properties are rarely investigated. To help closing this gap, we establish three asymptotic properties of the ReLU network: consistency, sieve-based convergence rate, and asymptotic normality. To validate the theoretical results, a Monte Carlo analysis is provided.

Keywords: Consistency, Rate of Convergence, Sieve Estimators, Rectified Linear Unit
JEL classification: G12; C21.

*Monash Center for Quantitative Finance and Investment Strategies has been supported by BNP Paribas.

†Hasan Fallahgoul, Monash University, School of Mathematics and Center for Quantitative Finance and Investment Strategies, 9 Rainforest Walk, 3800 Victoria, Australia. E-mail: hasan.fallahgoul@monash.edu.

‡Vincentius Franstianto, Monash University, School of Mathematics and Center for Quantitative Finance and Investment Strategies, 9 Rainforest Walk, 3800 Victoria, Australia. E-mail: vincentius.franstianto@monash.edu.

§Grégoire Loeper, Monash University, School of Mathematics and Center for Quantitative Finance and Investment Strategies, 9 Rainforest Walk, 3800 Victoria, Australia. E-mail: gregoire.loeper@monash.edu.

1 Introduction

The asymptotic properties of feed-forward networks (FFN) with rectified linear unit (ReLU) activation functions are rarely explored.[1] Although it has a great accuracy in statistical regression and classification tasks, the lack of these properties have left ReLU networks, as many other neural networks, as "black boxes". This lack of knowledge also barricades the development of statistical inferences of the ReLU network regressions.

The objective of a neural network is to transform an input data to some output through approximating an unknown function, *target function*, possibly non-linear and time-varying. Each layer of a neural network can be seen as an approximation for the unknown target function. For a complicated target function, it is likely the output of the first-layer neural network does not match the target function. Generally, there are two ways to overcome this problem: increasing the number of nodes (neurons) or layers.[2] Adding another layer to the neural network is equivalent to endow the model with another chance for a better approximation of the objective function. Theoretically, increasing the layers leads to better approximation. By adding more layers to the network, the model gets more powerful. Practically, however, it leads to several problems: (i) training the model is difficult and takes longer time; (ii) it is likely that the trained model becomes "overfitted"; (iii) to use the trained model for transforming unseen data in future, there will be a greater level of regularization to obtain a reasonable

[1] With the notable exception of Farrell, Liang, and Misra (2019).
[2] Adding more nodes and layers mean increasing network width and depth, respectively. There are other approaches such as using different activation function, optimization techniques, etc. Detailed information about a feed-forward network can be found in Anthony and Bartlett (2009).

validation metric.[3]

Farrell, Liang, and Misra (2019) have obtained a groundbreaking result on the asymptotic properties of deep neural networks. They prove probabilistic convergence rate for multi-layer ReLU networks, under the assumption that the number of hidden layers, i.e., depth of the network, grows with the sample size. However, two criticisms remain on the ground. Firstly, perhaps more importantly, practitioners fix the depth of the network, as growing the depth makes the training harder and larger tendency to eventual overfit.[4] Secondly, the paper does not prove the existence of the exact solution of least squares or logistic losses. It assumes the solution does exist. Since the neural network function is a form of sieve estimator,[5] it would be natural to explore the solution in the strong formulation. This paper tries to tackle these criticisms.

Many of the existing works such as Anthony and Bartlett (2009), Akpinar, Kratzwald, and Feuerriegel (2019), and Bartlett, Harvey, Liaw, and Mehrabian (2019) focus more on sample complexities, Vapnik-Chervonenkis and Pollard dimension upper bounds, which do not give much insight about statistical explainability in the way that asymptotic properties do. The other theoretical-oriented works such as Shen, Jiang, Sakhanenko, and Lu (2019) and Horel and Giesecke (2019) focus on asymptotic properties and statistical tests, respectively, of one-layer sigmoid networks. However, sigmoid networks are rarely used in practice compared to ReLU networks, as the latter have sparse representations and non-vanishing gradients that help speeding up training computations.

It is thus natural to explore the ReLU asymptotic properties using conventional ap-

[3]See Liu, Shi, Li, Li, Zhu, and Liu (2016), Sun, Chen, Wang, Liu, and Liu (2016), and references therein.
[4]See Sun, Chen, Wang, Liu, and Liu (2016), among others.
[5]See Chen and Shen (1998).

proaches such as consistency, probabilistic rate of convergence, and asymptotic normality, as neural network regression can be seen as a specific class of sieve extremum estimation, see Grenander (1981). The contribution of this paper is threefold. Firstly, motivated by Shen, Jiang, Sakhanenko, and Lu (2019), we prove that the ReLU network with fixed depth is consistent. This result is in the line of Farrell, Liang, and Misra (2019) while in our setting depth of network is fixed and does not diverge with the sample size.

Secondly, we establish nonasymptotic bounds for nonparametric estimation which we refer to as the rate of convergence of the ReLU neural network estimator. Finally, the asymptotic normality of a sieve estimator for the ReLU feed-forward network is provided. The asymptotic normality appears to be new to the literature and is one of the main theoretical contributions of this paper.

Our results are among the first inference of multi-layer ReLU networks built on the sieve space of continuous functions. The sieve estimation underlies many parametric and non-parametric estimating methods, such as time-series and quantile regressions. As these classical methods have been known before machine learning methods, the sieve-based inferences are more commonly known than the new machine-learning-specific inferences. The asymptotic properties of ReLU networks from the sieve framework will open a new possibilities of the adaptation of existing sieve inferences to ReLU neural networks.

Furthermore, it has been known that fixed-depth ReLU networks are easier to train than their growing-depth counterpart.[6] With the same number of sample data and itera-

[6]Detailed information can be found in Sun, Chen, Wang, Liu, and Liu (2016) and Liu, Shi, Li, Li, Zhu, and Liu (2016).

tions in the stochastic gradient descent, the fixed-depth networks give better convergence results. Considering the availability of easy-to-use packages for training neural networks such as Keras in Python, having a sieve-based convergence result will give an intuitive understanding of the accuracy of ReLU networks among communities that are unfamiliar with machine learning.

The paper is organized as follows. Section 2 provides an overview of the ReLU feed-forward network. Section 3 states the theoretical setting needed to prove asymptotic properties of ReLU neural networks. Section 4 presents the main theoretical results of the paper: the consistency, sieve-based convergence rate, and asymptotic normality. Section 5 explores the validity of theoretical results in simulations. Section 6 concludes.

2 ReLU feed-forward network

In this section, we discuss the architecture of a ReLU feed-forward network. We refer interested readers to Anthony and Bartlett (2009) and Goodfellow, Bengio, and Courville (2016) for detailed exposition.

In regression, we observe the observations y_i whose values are driven by the underlying target function f_0, which is a function of a d-dimensional vector $\boldsymbol{x_i}$. Each element of $\boldsymbol{x_i}$ is an observed predictor/independent variable. As the exact functional form of f_0 is rarely known, the target function f_0 is estimated using a specific function of $\boldsymbol{x_i}$. In our context, this estimating function is a multi-layer neural network with rectified linear unit (ReLU) activation function.

Figure 1 is an example of a ReLU network. This is an example of a feed-forward

network where information propagates only forward.[7] This network begins by taking input from d initial nodes. The number of initial nodes in Figure 1 is two, i.e., $d = 2$. Each initial node corresponds to each element of the $d-$dimensional predictor vector x_i, and the layer consisting of these nodes is called the input layer. These initial nodes can be seen as impulse receptors in biological neural networks.

The inputs are then transformed into signals that are propagated forward to the next layer called the first hidden layer, equivalent to neurons in the biological counterpart. This layer contains H_n hidden nodes, and each of them are connected to all nodes in the input layer. The value of each node, Y, is specified in the following way. First, one needs to calculate the sum of weighted of previous nodes plus a bias. More specifically,

$$Y = \sum_i w_i x_i + b$$

where w_i, x_i, and b are the weight, value of each node, and bias, respectively. Since the value of Y can belong to $(-\infty, \infty)$, one needs to use a function to decide whether the neuron associated with Y to be "fired" or not. An activation function is used for this purpose. There are various types of activation function. In this paper, we use the most popular one, rectified linear unit, which is given by $ReLU(x) = \max(x, 0)$.

In plain words, the input signals are taken in the form of linear combinations of all input nodes, whose weights depend on each hidden nodes. Then, these hidden nodes may be activated based on the used activation function. The activation is the same for all nodes in the hidden layer. In our case, the activation function is $ReLU(x) = \max(x, 0)$.

[7] There are other classes of deep neural networks which the information does not only propagate forward, see Anthony and Bartlett (2009) and Goodfellow, Bengio, and Courville (2016) for detailed exposition. The results of this paper is for feed-forward network, however, the validity of the results of this paper for other neural networks is an open question.

We refer to this network, ReLU network. If the input is positive, then the hidden node is activated and produces a positive output. If it is not, then the node is not activated and produces a zero output. In Figure 1, the first and second hidden layer have three nodes.

The outputs of this hidden layer will be then used as new inputs to the next layer with nodes, and again each of them are connected to the inputing nodes from the first hidden layer. If the next layer is also a hidden layer, then the input signals from the previous hidden layer will be processed by each node of this next hidden layer and turned into output in the same manner as the previous hidden layer nodes process input.[8] If the next layer is the output layer, equivalent to response to the impulse in its biological counterpart, the input signals are still linear aggregations, but no activation function applied on them. They are taken plainly as linear combinations. We denote the number of hidden layer in ReLU network by L_d which in Figure 1, $L_d = 2$.

Notation: For the rest of the paper, vectors are denoted by either bold or capital fonts. $N(\epsilon, \mathcal{D}, \rho)$ is a covering number for a pseudo-metric space $(\mathcal{D}, \rho)$. Asymptotic inequality $a_n \lesssim b_n$ means $\exists R > 0, N_1 \in \mathbb{N}$ such that $\forall n \geq N_1$, $a_n \leq R b_n$, $a_n \sim b_n$ implies $\lim_{n\to\infty} \frac{a_n}{b_n}$ = 1. The covering number $N(\epsilon, \mathcal{D}, \rho)$ for a pseudo-metric space $(\mathcal{D}, \rho)$ is the minimum number of ϵ-balls of the pseudo-metric ρ needed to cover $\mathcal{D}$, with possible overlapping. $\mathcal{C}^0([0,1]^d)$ is the set of all continuous function being defined on $[0,1]^d$, and $\mathcal{W}^{k,\infty}([0,1]^d)$ denotes the Sobolev space of order k on $[0,1]^d$. We denote big-O and small-O as $\mathcal{O}$ and o, respectively. Also, both $\mathcal{O}_{\mathbb{P}}$ and $o_{\mathbb{P}}$ denote big-O and small-O in probability.

[8]The input signals have linear form and are subject to the ReLU activation function.

3 The setting

In this section, we discuss the setting that is need to establish the main results in Section 4. Suppose that the true non-parametric regression model is

$$y_i = f_0(\boldsymbol{x_i}) + \epsilon_i$$

where ϵ_1, ϵ_2,...., ϵ_n are independent identically distributed defined on a complete probability space, $(\Omega, \mathcal{A}, \mathbb{P})$, $\mathbb{E}[\epsilon_i] = 0$, $\text{Var}(\epsilon_i) = \sigma^2 < \infty$, $\boldsymbol{x_1}$, $\boldsymbol{x_2}$,...., $\boldsymbol{x_n} \in \mathcal{X} = [0,1]^d$ are vectors of predictors, and $f_0 \in \mathcal{F} := \{f \in \mathcal{C}^0 \mid f : [0,1]^d \to \mathbb{R}\}$. Define the sample squared loss on $f \in \mathcal{F}$ and the population criterion function, respectively, as

$$\mathbb{Q}_n(f) := \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x_i}))^2 = \frac{1}{n}\sum_{i=1}^{n}(f_0(\boldsymbol{x_i}) - f(\boldsymbol{x_i}))^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i(f_0(\boldsymbol{x_i}) - f(\boldsymbol{x_i})) + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2$$

$$Q_n(f) := \mathbb{E}[\mathbb{Q}_n(f)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x_i}))^2\right] = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}))^2 + \sigma^2.$$

In regression, we are interested to find $\hat{f}$ such that

$$\hat{f} := \arg\min_{f \in \mathcal{F}} \mathbb{Q}_n(f).$$

However, if $\mathcal{F}$ is too rich, $\hat{f}$ may be inconsistent.[9] Hence, we are interested in finding a sequence of nested function spaces $\mathcal{F}_n$, which satisfies

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

where $\forall f \in \mathcal{F}$, $\exists f_n \in \mathcal{F}_n$ *s.t.* $\lim_{n\to\infty} \rho(f, f_n) = 0$. More precisely, $\mathcal{F}_n$ is dense in $\mathcal{F}$. $\mathcal{F}_n$ itself is called a *sieve space of* f *with respect to the pseudo-metric* ρ, and the sequence $\{f_n\}$ is called

[9] Inconsistency here means ($\hat{f} \xrightarrow{\mathbb{P}} f_0$)

a *sieve*. Take $\rho \equiv \rho_n$, where $\rho_n : \mathcal{F} \times \mathcal{F} \to [0, \infty)$

$$\rho_n(f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x_i})^2}.$$

where ρ_n is a pseudo-norm.[10] Hence, instead of $\hat{f}$, we are looking for an *approximate sieve estimator* $\hat{f}_n$ that satisfies

$$\mathbb{Q}_n(\hat{f}_n) \leq \inf_{f \in \mathcal{F}_n} \mathbb{Q}_n(f) + \mathcal{O}_{\mathbb{P}}(\eta_n)$$

where $\lim_{n\to\infty} \eta_n = 0$.

The next question is how to construct $\mathcal{F}_n$ that makes it dense in $\mathcal{F}$. First of all, consider the following fixed-width ReLU feed-forward network function space indexed by W_n

$$\mathcal{F}_{W_n} \coloneqq \left\{ h_{L_d+1,1}(\boldsymbol{x}) : \boldsymbol{x} \in [0,1]^d \right\}$$

where $h_{u,j}(\boldsymbol{x})$ is the output of the j^{th} node of the layer u in the ReLU network with input $\boldsymbol{x}$, $u = 0$ or $u = L_d + 1$ correspond to the input and output layers, respectively, and $1 \leq u \leq L_d$ correspond to the u^{th} hidden layer. We also have $j \in \{1, 2, ..., H_{n,u}\}$, where $H_{n,u}$ is the number of nodes in the u^{th} layer, $H_{n,0} = d$, and $H_{n,L_d+1} = 1$. For $1 \leq u \leq L_d$, the formula for $h_{u,j}(\boldsymbol{x})$ is

$$h_{u,j}(\boldsymbol{x}) = ReLU\left(\sum_{k=1}^{H_{n,u-1}} \gamma_{u,j,k} \cdot h_{u-1,k}(\boldsymbol{x}) + \gamma_{u,j,0} \right)$$

where $h_{0,k}(\boldsymbol{x}) = x_k$, the k^{th} element of $\boldsymbol{x}$.

We use the upper bound $\max_{1\leq j\leq H_{n,u}} \sum_{k=0}^{H_{n,u-1}} |\gamma_{u,j,k}| \leq M_{n,u}$, $\forall\ 1 \leq u \leq L_d + 1$, where $M_{n,u} > 1$, $M_{n,u}$ can depend on n, and $M_{n,0} = 1$ as $\mathcal{X} = [0,1]^d$. This upper bound is used in the entropy number upper bound. W_n itself is the number of parameters $\gamma_{u,j,k}$ in a single

[10] Detailed information about why ρ_n is a pseudo-norm can be found in Shen, Jiang, Sakhanenko, and Lu (2019).

ReLU network, with $W_n = \sum_{u=0}^{L_d} (H_{n,u} + 1) H_{n,u+1}$.

According to Proposition 1 in Yarotsky (2018), $\forall f_0 \in \mathcal{F}$, $\exists \pi_{W_n} f_0 \in \mathcal{F}_n$ s.t.

$$\|\pi_{W_n} f_0 - f_0\|_\infty := \sup_{\boldsymbol{x} \in [0,1]^d} |\pi_{W_n} f_0(\boldsymbol{x}) - f_0(\boldsymbol{x})| \leq \mathcal{O}\left(\omega_{f_0}\left(\mathcal{O}\left(W_n^{-1/d}\right)\right)\right)$$

with $\omega_{f_0} : [0, \infty] \to [0, \infty)$, $\omega_{f_0}(r) = max\left\{|f_0(\boldsymbol{x}) - f_0(\boldsymbol{y})| : \boldsymbol{x}, \boldsymbol{y} \in [0,1]^d, |\boldsymbol{x} - \boldsymbol{y}| < r\right\}$. It is clear that $\|\cdot\|_\infty$ is a pseudo-metric, $(\mathcal{F}_{W_n}, \|\cdot\|_\infty)$ is a sieve space of f_0, and $\{\pi_{W_n} f_0\}$ is the related sieve. It is imperative that $W_n^{-1/d} \to 0$ as $n \uparrow \infty$ to ensure dense $\mathcal{F}_{W_n}$. Also, for $\Gamma_n := \left\{\gamma_{u,j,k} : \forall u, j, k\right\}$,[11] we need to ensure that the range of the parameters inside $\cup_n \Gamma_n$ can span through $\mathbb{R}^{W_n}$. These requirements for having a dense $\mathcal{F}_{W_n}$ can be summarized as

$$W_n \uparrow \infty \text{ and } \Gamma_n \uparrow \mathbb{R}^{W_n}, \text{ as } n \uparrow \infty$$

and the chosen set Γ_n is a compact set

$$\Gamma_n = \left[-M_{n,u,j,k}^{(\gamma)}, M_{n,u,j,k}^{(\gamma)}\right]^{W_n}, \forall 1 \leq u \leq L_d + 1, \forall 1 \leq j \leq H_{n,u}, \forall 0 \leq k \leq H_{n,u-1}$$

such that $\forall u, j, k$, $\left|\gamma_{u,j,k}\right| \leq M_{n,u,j,k}^{(\gamma)}$, and also $\sum_{j=1}^{H_{n,u}} \sum_{k=0}^{H_{n,u-1}} M_{n,u,j,k}^{(\gamma)} = M_{n,u}$. Hence, the two requirements of the denseness of $\mathcal{F}_{W_n}$ will be gotten under the following assumption.

Assumption 3.1 (Assumption for Dense $\mathcal{F}_{W_n}$). $H_n, M_{n,u,j,k}^{(\gamma)} \uparrow \infty$ *as* $n \uparrow \infty$, $\forall 1 \leq u \leq L_d + 1$, $\forall 1 \leq j \leq H_{n,u}$, $\forall 0 \leq k \leq H_{n,u-1}$.

For the following next sections, the results and proofs are obtained by mimicking the proofs in Shen, Jiang, Sakhanenko, and Lu (2019). More specifically, the main things that are changed are the upper bound of the supremum of $\|\cdot\|_\infty$ among $\mathcal{F}_{W_n}$ elements and the upper bound of the entropy number, and some other things that are related to them.

[11] This is the set of all parameters in the ReLU network

4 Main results

In this section we discuss our three main theoretical results in this section; sieve-estimator consistency, convergence rate, and asymptotic normality. All proofs are provided in Appendix A.

4.1 Existence

Theorem 4.1.1 (Existence). *There exists an approximate sieve estimator $\hat{f}_n$ in $\mathcal{F}_{W_n}$.*

The following remark is the tool for proving the Existence Theorem.

Remark 4.1.1 (Existence Conditions). (Remark 2.1. in Chen (2007)). *There exists an approximate sieve estimator $\hat{f}_n$ inside $\mathcal{F}_{W_n}$ if the following statements hold*

EC1. $\mathbb{Q}_n(f)$ *is measurable function of the data* $(\boldsymbol{x_i}, y_i), i \in \{1, 2, ..., n\}$

EC2. $\mathbb{Q}_n(f)$ *is lower semicontinuous on* $\mathcal{F}_{W_n}$ *under the pseudo-metric* ρ_n*, for each* $\omega \in \Omega$ *fixing the sequence* $\{\boldsymbol{x_i}, y_i(\omega))\}_{i=1}^n$

EC3. $\mathcal{F}_{W_n}$ *is a sieve of* $\mathcal{F}$ *and compact under* ρ_n*.*

Note that in Chen (2007), the notation $\widehat{Q}_n(f) = -\mathbb{Q}_n(f)$ is used. This explains why EC2 of the Remark 4.1.1 is lower semicontinuous instead of upper semicontinous. EC1 is obviously satisfied by $\mathbb{Q}_n(f)$, as $y_i = f_0(\boldsymbol{x_i}) + \epsilon_i$. Note also that fixing $\{\boldsymbol{x_i}, \boldsymbol{y_i}(\omega)\}$ is equivalent to fixing $\{\epsilon_i(\omega)\}, \forall\, \omega \in \Omega$. To prove EC2 and EC3, we make use of the following lemma.

Lemma 4.1.1. *For each* n, $1 \leq u \leq L_d + 1$, *and* $1 \leq j \leq H_{n,u}$,

$$\sup_{1 \leq j \leq H_{n,u}} \|h_{n,j}\|_\infty \leq M_{n,i}^* := \prod_{i=0}^{u} M_{n,i} \geq 1$$

and this implies

$$\sup_{f \in \mathcal{F}_{W_n}} \|f\|_\infty \leq M^*_{n,L_d+1} = \prod_{u=0}^{L_d+1} M_{n,u} \geq 1.$$

Please refer to the Appendix A for the proof of this lemma, EC2, and EC3. Note also that $\mathbb{Q}_n(f)$ can be proven to be continuous on $(\mathcal{F}_{W_n}, \rho_n)$, which is stronger than EC2. Thus the existence of $\hat{f}_n$ is justified.

We can consider the bounded ReLU function

$$h^*_{L_d+1,1}(\boldsymbol{x}) = \min\left(UB_{f_0}, \max\left(LB_{f_0}, h_{L_d+1,1}(\boldsymbol{x})\right)\right)$$

where LB_{f_0} and UB_{f_0} are lower and upper bounds of f_0 such that $LB_{f_0} < \min_{\boldsymbol{x} \in [0,1]^d} f_0$ and $UB_{f_0} > \max_{\boldsymbol{x} \in [0,1]^d} f_0$, respectively. The existence of both bounds are guaranteed by the Extreme Value Theorem.

If $\pi^*_{W_n} f_0$ is the sequence of functions $h^*_{L_d+1,1}(\boldsymbol{x})$ sharing the same parameters with $\pi_{W_n} f_0$, then the fact that $\|\pi_{W_n} f_0 - f_0\|_\infty \to 0$ as $n \to \infty$ gives $\pi_{W_n} f_0 = \pi^*_{W_n} f_0$, where $n > M$, $\exists M > 0$. Thus $\pi^*_{W_n} f_0$ is also a sieve sequence, and the pseudo-metric space $\left(\mathcal{F}^*_{W_n}, \rho_n\right)$ being composed of functions $h^*_{L_d+1,1}(\boldsymbol{x})$ is also a sieve space of $\mathcal{F}$.

As $\min\left(UB_{f_0}, x\right)$ and $\max\left(LB_{f_0}, x\right)$ are Lipschitz continuous functions with Lipschitz constant 1, all Existence Conditions are satisfied and $\mathbb{Q}_n(f)$ are still continuous on $\left(\mathcal{F}^*_{W_n}, \rho_n\right)$. The key difference is now $\sup_{f \in \mathcal{F}^*_{W_n}} \|f\|_\infty = M^{**}_{n,L^*_d+1} := \min\left(M^*_{n,L_d+1}, \max\left(\left|LB_{f_0}\right|, \left|UB_{f_0}\right|\right)\right)$, with $L^*_d = L_d + 2$ is the depth of the new bounded ReLU network.

The advantage of considering such bounded ReLU network space will be clear in the next subsection, where we show that $\hat{f}_n$ is also consistent under a product probability space.

4.2 Consistency

Define the product space $(\Omega^*, \mathcal{A}^*, \mathbb{P}^*) = \prod_{i=1}^{n} (\Omega, \mathcal{A}, \mathbb{P}) \times (\mathcal{Z}, \mathcal{C}, \mathbb{P}_{\mathcal{Z}})$, with the last probability measure containing additional random variables independent of $\prod_{i=1}^{n} (\Omega, \mathcal{A}, \mathbb{P})$. The consistency of $\hat{f}_n$ is satisfied under this probability measure, with a condition on the parameter number growth.

Theorem 4.2.1 (Consistency). *Define*

$$M_{n,L_d+1}^{(all)} := \max_{1 \le i \le L_d+1} M_{n,i} > 1,$$

$$C_{n,d,L_d+1,W_n}^{*} := W_n \, ln \left(d \, M_{n,L_d+1}^{*} \, W_n \, (M_{n,L_d+1}^{(all)})^{L_d} \right),$$

If $(M_{n,L_d+1}^{*})^2 C_{n,d,L_d+1,W_n}^{*} = o(n)$,

$$\operatorname*{plim}_{n \to \infty} \rho_n(\hat{f}_n - f_0) = 0, \textit{under } (\Omega^*, \mathcal{A}^*, \mathbb{P}^*).$$

The conditions for consistency are given in the following remark.

Remark 4.2.1 (Consistency Conditions). (Remark 3.1.(3) in Chen (2007)). *The approximate sieve estimator* $\hat{f}_n$ *in the sieve space* $\mathcal{F}_{W_n}$ *of* $\mathcal{F}$ *satisfies*

$$\operatorname*{plim}_{n \to \infty} \rho_n(\hat{f}_n - f_0) = 0$$

if the following conditions are satisfied

CC1. $Q_n(f)$ *is continuous at* f_0 *in* $\mathcal{F}$, $Q_n(f_0) < \infty$

CC2. For all $\zeta > 0$, $Q_n(f_0) < \inf_{\{f \in \mathcal{F} : \rho_n(f - f_0) \ge \zeta\}} Q_n(f)$

CC3. $\mathbb{Q}_n(f)$ *is a measurable function of the data* $(\boldsymbol{x_i}, y_i)$, $i \in \{1, 2, ..., n\}$

CC4. $\mathbb{Q}_n(f)$ *is lower semicontinuous on* $\mathcal{F}_{W_n}$ *under* ρ_n*, for each* $\omega \in \Omega$ *fixing the sequence* $\{\boldsymbol{x_i}, y_i(\omega))\}_{i=1}^n$

CC5. $(\mathcal{F}_{W_n}, \rho_n)$ *is compact sieve space*

CC6. (Uniform convergence) $\operatorname*{plim}_{n\to\infty} \sup_{f\in\mathcal{F}_{W_n}} |\mathbb{Q}_n(f) - Q_n(f)| = 0$*, for each* W_n*.*

Note that $Q_n(f) = \dfrac{1}{n}\sum_{i=1}^{n} (f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}))^2 + \sigma^2$ is continuous on $\mathcal{F}$. The proof of its continuity is very similar to the proof of the lower semicontinuity of $\mathbb{Q}_n(f)$, with the related constant is now $\widehat{D}_{n,L_d+1} = n^{-1}\left(2M^*_{n,L_d+1} + 2\max_{\boldsymbol{x}\in[0,1]^d} |f_0(\boldsymbol{x})|\right)$, and hence CC1 is satisfied. It is clear that CC2 is satisfied as f_0 minimizes Q_n in $\mathcal{F}$. As CC3, CC4, and CC5 are the Existence Conditions, we already have them. The last thing that needs to be dealt with is CC6.

Lemma 4.2.1 (CC6 Satisfaction). *If* $(M^*_{n,L_d+1})^2 C^*_{n,d,L_d+1,W_n} = o(n)$*, then CC6 is satisfied under* $(\Omega^*, \mathcal{A}^*, \mathbb{P}^*)$.

We can also consider $\mathcal{F}^*_{W_n}$ for consistency. CC1-CC5 are clearly satisfied. For CC6, our new conditions are

$$(M^{**}_{n,L^*_d+1})^2 C^{**}_{n,d,L^*_d+1,W^*_n} = o(n)$$

where

$$C^{**}_{n,d,L^*_d+1,W^*_n} \coloneqq W^*_n \, ln\left(d\, M^{**}_{n,L^*_d+1}\, W^*_n\, (M^{(all)*}_{n,L^*_d+1})^{L^*_d}\right)$$

where $M^{(all)*}_{n,L^*_d+1} \coloneqq \max\left(M^{(all)}_{n,L_d+1}, \left|LB_{f_0}\right|, \left|UB_{f_0}\right|\right)$ and $W^*_n = W_n + 4$. The reason W^*_n is taken such this is both functions $\min\left(UB_{f_0}, v\right)$ and $\max\left(LB_{f_0}, v\right)$ are activation functions that can be seen to take linear aggregation of v ($1\cdot$v + 0) as an input, and thus requires two

additional weights.

Remark 4.2.2. *As M^*_{n,L_d+1} is bounded above by* $\max\left(\left|LB_{f_0}\right|, \left|UB_{f_0}\right|\right)$*, the consistency condition satisfaction is only dependent on the growth rate of $W^*_n = \mathcal{O}(W_n)$. This implies we have more flexibility in adjusting the growth rate of W_n. Note that UB_{f_0} and LB_{f_0} can be taken to be very large positive or negative real numbers, respectively (such as $\pm 10^{100,000}$). As most f_0 encountered in practice are rarely that large, we can treat bounded ReLU networks as if they are unbounded ReLU networks in most applications.*

In the next subsection, we show that the convergence rate of $\hat{f}_n$ can be bounded by η_n convergence rate.

4.3 Rate of convergence

Theorem 4.3.1. (**Rate of Convergence**) *Suppose that*

$$\eta_n = \mathcal{O}\left(\max\left\{\rho_n\left(\pi_{W_n} f_0 - f_0\right)^2, \left(\frac{C^*_{n,d,L_d+1,W_n}}{n}\right)^{2/3}\right\}\right)$$

where C^*_{n,d,L_d+1,W_n} defined in the Consistency Theorem and $\left(M^*_{n,L_d+1}\right)^2 C^*_{n,d,L_d+1,W_n} = o(n)$. Then

$$\rho_n\left(\hat{f}_n - f_0\right) = \mathcal{O}_{\mathbb{P}^*}\left(\max\left\{\rho_n\left(\pi_{W_n} f_0 - f_0\right), \left(\frac{C^*_{n,d,L_d+1,W_n}}{n}\right)^{1/3}\right\}\right).$$

The following remark underlies the Rate of Convergence Theorem proof.

Remark 4.3.1 (Convergence Rate of $\rho_n(\hat{f}_n - \pi_{W_n} f_0)$). (Theorem 3.4.1 in Vaart and Wellner (1996)) *For each n, let δ_n satisfying $0 \leq \delta_n \leq \alpha$ be arbitary (δ_n is typically a multiple of $\rho_n(\pi_{W_n} f_0 - f_0)$)). Suppose that, for every n and $\delta_n < \delta \leq \alpha$,*

$$\sup_{\substack{\delta/2 \leq \rho_n(f - \pi_{W_n} f_0) \leq \delta \\ f \in \mathcal{F}_{W_n}}} Q_n(\pi_{W_n} f_0) - Q_n(f) \leq -\delta^2$$

$$\mathbb{E}_{\mathbb{P}^*}\left[\sup_{\substack{\delta/2\leq\rho_n(f-\pi_{W_n}f_n)\leq\delta\\ f\in\mathcal{F}_{W_n}}}\sqrt{n}\left[\left(\mathbb{Q}_n-Q_n\right)\left(\pi_{W_n}f_0\right)-\left(\mathbb{Q}_n-Q_n\right)\left(f\right)\right]\right]\lesssim\phi_n(\delta)$$

for functions ϕ_n such that $\delta\mapsto\phi_n(\delta)/\delta^\beta$ is decreasing on (δ_n,α) for some $\beta<2$. Let $r_n\lesssim\delta_n^{-1}$ satisfy

$$r_n^2\phi_n\left(\frac{1}{r_n}\right)\leq\sqrt{n}, \text{ for every } n.$$

If the approximate sieve estimator $\hat{f}_n$ satisfies $\mathbb{Q}_n(\hat{f}_n)\leq\mathbb{Q}_n(\pi_{W_n}f_0)+\mathcal{O}_{\mathbb{P}}\left(r_n^{-2}\right)$ and $\rho_n(\hat{f}_n-\pi_{W_n}f_0)$ converges to zero in outer probability defined in $(\Omega^*,\mathcal{A}^*,\mathbb{P}^*)$, then

$$\rho_n(\hat{f}_n-\pi_{W_n}f_0)=\mathcal{O}_{\mathbb{P}^*}\left(r_n^{-1}\right).$$

If the displayed conditions are valid for $\alpha=\infty$, then the condition that $\hat{f}_n$ is consistent is unnecessary.

The two supremum-upper-bound conditions in Remark 4.3.1 have been proven in Shen, Jiang, Sakhanenko, and Lu (2019). We will state them in the remark below

Remark 4.3.2. (Lemma 4.1 and Lemma 4.2 in Shen, Jiang, Sakhanenko, and Lu (2019))

- *For every n and $\delta>8\rho_n\left(\pi_{W_n}f_0-f_0\right)$, we have*

$$\sup_{\substack{\delta/2\leq\rho_n(f-\pi_{W_n}f_n)\leq\delta\\ f\in\mathcal{F}_{W_n}}}Q_n(\pi_{W_n}f_0)-Q_n(f)\lesssim-\delta^2$$

- *For every sufficiently large n and $\delta>8\rho_n\left(\pi_{W_n}f_0-f_0\right)$, we have*

$$\mathbb{E}_{\mathbb{P}^*}\left[\sup_{\substack{\delta/2\leq\rho_n(f-\pi_{W_n}f_0)\leq\delta\\ f\in\mathcal{F}_{W_n}}}\sqrt{n}\left[\left(\mathbb{Q}_n-Q_n\right)\left(\pi_{W_n}f_0\right)-\left(\mathbb{Q}_n-Q_n\right)\left(f\right)\right]\right]$$

$$\lesssim \int_0^{\delta} \sqrt{ln\left(N\left(\eta, \mathcal{F}_{W_n}, \rho_n\right)\right)} d\eta.$$

We can then finish the proof of the major theorem regarding the rate of convergence of $\hat{f}_n$.

4.4 Asymptotic Normality of The Approximate Sieve Estimator

We show that $\hat{f}_n - f_0$ is indeed asymptotically Gaussian under certain assumptions. We follow the same procedure as the proof of one-layer sigmoid network normality in Shen, Jiang, Sakhanenko, and Lu (2019), which is inspired by the General Theory on Asymptotic Normality in Shen et al. (1997).

This general asymptotic normality requires stronger growth regulations than consistency. To achieve this, we depart from our usual assumption that each ReLU network requires fixed depth. The networks now are allowed to have growing depths. We note that this do *not* change any proofs or circumstances above to achieve consistency and sieve-based rate of convergence. This is done to ensure that we can use Theorem 1 from Yarotsky (2017), which requires depth growing ReLU network classes. Then, we have a flexible sieve sequence rate $\rho_n\left(\pi_{W_n} f_0 - f_0\right)$ that can be adjusted for asymptotic Gaussianity requirement. We replace L_d with L_n for clarity.

We also require that $f_0 \in \left\{f \in \mathcal{C}^0([0,1]^d) \cap \mathcal{W}^{k,\infty}([0,1]^d) \mid \|f\|_{\mathcal{W}^\infty} \leq M_{\mathcal{W}}\right\}$, $\exists M_{\mathcal{W}}>0$, and $k \in \mathbb{N}$. $\mathcal{W}^{k,\infty}([0,1]^d)$ is the Sobolev space defined in $[0,1]^d$, composed of functions whose derivatives up to order k are defined in *weak* sense, in terms of; *partial integration* for d=1, or, *distribution* for d>1. This space is a Banach space with respect to the norm

$\|f\|_{\mathcal{W}^\infty} := \max_{\boldsymbol{k}:0\leq|\boldsymbol{k}|\leq k} \|D^{\boldsymbol{k}} f(x)\|_{L^\infty([0,1]^d)}$, where $\boldsymbol{k}\in(\mathbb{N}\cup\{0\})^d$, $D^{\boldsymbol{k}} f(\boldsymbol{x}) := \dfrac{\partial^{|\boldsymbol{k}|} f}{\partial x_1^{k_1}\partial x_2^{k_2}\ldots\partial x_d^{k_d}}$ is the related weak derivative, $x_1,\ldots,x_d$ and $k_1,\ldots,k_d$ are the elements of vectors $\boldsymbol{x}$ and $\boldsymbol{k}$, respectively.

We use the Gâteaux derivative of $\mathbb{Q}_n(f)$ at f_0 in the direction of $f - f_0$. For further algebraic details, please refer to Section 5 of Shen, Jiang, Sakhanenko, and Lu (2019).

$$\begin{aligned} d\mathbb{Q}_n(f_0; f - f_0) &= \lim_{\tau\to 0} \frac{\mathbb{Q}_n\left(f_0 + \tau(f - f_0)\right) - \mathbb{Q}_n\left(f_0\right)}{\tau} \\ &= \lim_{\tau\to 0} \frac{\sum_{i=1}^n \left[y_i - f_0(\boldsymbol{x_i}) - \tau\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right]^2 - \sum_{i=1}^n \left[y_i - f_0(\boldsymbol{x_i})\right]^2}{n\tau} \\ &= -\frac{2}{n}\sum_{i=1}^n \left[\epsilon_i\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right] \end{aligned}$$

with the related first-order Taylor remainder term

$$\begin{aligned} \mathcal{R}_1\left(f_0; f - f_0\right) &= \mathbb{Q}_n(f) - \mathbb{Q}_n(f_0) - d\mathbb{Q}_n(f_0; f - f_0) \\ &= \frac{1}{n}\sum_{i=1}^n \left[y_i - f(\boldsymbol{x_i})\right]^2 - \sum_{i=1}^n \left[y_i - f_0(\boldsymbol{x_i})\right]^2 + \frac{2}{n}\left[\epsilon_i\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^n \left[\epsilon_i + f_0(\boldsymbol{x_i}) - f(\boldsymbol{x_i})\right]^2 - \sum_{i=1}^n \epsilon_i^2 + \frac{2}{n}\left[\epsilon_i\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^n \left[f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right]^2 = \rho_n\left(f - f_0\right)^2. \end{aligned}$$

Note that Gâteaux derivatives of $\mathbb{Q}_n$ can be defined as $\mathcal{F}$ is a convex vector space. We define a pseudo-scalar product $\langle\cdot,\cdot\rangle_{\rho_n} : \mathcal{F}\times\mathcal{F}\to\mathbb{R}$, with the mapping rule

$$\langle f, g\rangle_{\rho_n} = \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x_i})g(\boldsymbol{x_i})$$

where the subscript ρ_n indicates $\langle f - g, f - g\rangle_{\rho_n} = \rho_n(f - g)^2$. The proof that $\langle\cdot,\cdot\rangle_{\rho_n}$ is indeed a pseudo-scalar product is the proof of Proposition 6.2 in Shen, Jiang, Sakhanenko, and Lu (2019).

We also use of the following remark, which is useful to bound the empirical process $\sqrt{n}\, d\mathbb{Q}_n(f_0; f - f_0)$ in the proof.

Remark 4.4.1. (Lemma 5.1. in Shen, Jiang, Sakhanenko, and Lu (2019)). *Let $X_1, ..., X_n$ be independent random variables, X_i is under probability measure $\mathbb{P}_i$. Define the empirical process $\{\nu_n(g)\}$ as*

$$\nu_n(g) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[g\left(X_i\right) - \mathbb{E}_{\mathbb{P}_i}\left[g\left(X_i\right)\right]\right].$$

Let $\mathcal{G}_n = \{g : \|g\|_\infty \leq M_n\}$, $\epsilon > 0$ and $V \geq \sup_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\left[g\left(X_i\right)\right]$ be arbitrary. Define $\psi\left(B, n, V\right) := B^2 / \left[2V\left(1 + \frac{BM_n}{2\sqrt{n}V}\right)\right]$. If $\ln\left(N\left(u, \mathcal{G}_n, \|\cdot\|_\infty\right)\right) \leq A_n u^{-r}$ for some $0 < r < 2$ and $u \in (0, a]$, where a is a small positive number, and, there exists a positive constant $K_i = K_i(r, \epsilon)$ 1 = 1,2 such that

$$B \geq K_1 A_n^{\frac{2}{r+2}} M_n^{\frac{2-r}{r+2}} n^{\frac{r-2}{2(r+2)}} \vee K_2 A_2^{1/2} V^{\frac{2-r}{4}}.$$

Then

$$\mathbb{P}^*\left(\sup_{g\in\mathcal{G}_n} |\nu_n(g)| > B\right) \leq 5\exp\left(-(1-\epsilon)\psi(B, n, V)\right).$$

Now, we are ready to state the asymptotic Gaussianity exhibited by $\hat{f}_n$.

Theorem 4.4.1. **(Asymptotic Normality)** *Suppose that $\eta_n = o\left(r_n^{-2}\right)$, and also*

$$\begin{aligned} r_n^{-1} &= o(n^{-1/2}) \\ \sqrt{M^*_{n,L_n+1}}\, C^*_{n,d,L_n+1,W_n} &= o\left(n^{1/4}\right) \\ \rho_n\left(\pi_{W_n} f_0 - f_0\right) &= o\left(\min\left\{n^{-1/4}, n^{-1/6}\left(C^*_{n,d,L_n+1,W_n}\right)^{-1/3}\right\}\right) \end{aligned}$$

then the distribution of the statistics

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\hat{f}_n(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)$$

approaches normal distribution when $n \to \infty$.

4.4.1 Asymptotic Normality Conditions Satisfaction for Sufficiently Smooth f_0

In this section, we discuss the growth rate requirements that satisfy the Asymptotic Normality Theorem conditions, which are very challenging. As stated before, we use the following remark, which is just a reciting of Theorem 1 from Yarotsky (2017).

Remark 4.4.2. (Theorem 1 from Yarotsky (2017)). *For any function*

$$f \in \mathcal{G}^* := \left\{ f \in \mathcal{W}^{k,\infty}([0,1]^d) \mid \|f\|_{\mathcal{W}^\infty} \leq 1 \right\}$$

and any k,d$\in\mathbb{N}$, $\varepsilon\in(0,1)$, *there is a feed-forward ReLU network, whose layers may be connected with layers after their adjacent layers, with a weight assignment that*

- *is capable of expressing* f *with error* ε
- *has the depth at most* $c(\ln(1/\varepsilon)+1)$ *and at most* $c\varepsilon^{d/k}(\ln(1/\varepsilon)+1)$ *weights and hidden layer nodes, with some constant c = c(d,k).*

This remark makes it possible to construct a sieve sequence $\{\pi_{W_n} f_0\}$ that satisfies $\|\pi_{W_n} f_0 - f_0\|_\infty = \varepsilon_n$. Note that we can have any function f satisfying $\|f\|_{\mathcal{W}^\infty} \leq K, \forall K > 0$ to be related to $f^* \in \mathcal{G}^*$ by $f^* = \frac{1}{K} f$. We can then take ε_n to be a sequence in $(0,1)$ such that $\varepsilon_n \downarrow 0$. One might question the possibilities of getting such weight assignments from a compact Γ_n that can adjust ε_n. We emphasize that Γ_n can be made as large as possible and be replaced by other compact sets. For example, one can just take $M^\gamma_{n,u,j,k} = M'$, where M' can be taken *arbitarily* large, and replace Γ_n constructed from element-wise bounds on $\gamma_{u,j,k}$ by the set made from ℓ^1-norm bounds for $\sum_{k=0}^{H_{n,u-1}} |\gamma_{u,j,k}|$, where the bounds

are $M_{n,u} = \sum_{j=1}^{H_{n,u}} \sum_{k=0}^{H_{n,u-1}} M^{'}$. The resulting Γ_n is still compact. However, as M' can be made as large as possible, the sieve sequence $\{\pi_{W_n} f_0\}$ has its tail in Γ_n for sufficiently large n, as it converges to f_0.

We show that the ReLU network required by the remark above can be contained in a ReLU network with layers connected only to their adjacent layers, which is called a multi-layer perceptron. Our multi-layer perceptron also assume a hidden layer node is connected to all previous nodes. Our idea is similar to Lemma 1 in Farrell, Liang, and Misra (2019), although we use hidden-layer-nodes upper bound instead of weight bound. Note that all ReLU networks described in previous sections are multi-layer perceptrons. We give them explicit name in this section for the sake of clarity. In other cases, what we mean by ReLU networks are multi-layer perceptrons, as they are the most commonly used ReLU networks in practice.

Lemma 4.4.1. *If θ is a ReLU feed-forward network with non adjacent layer connections with N_n hidden layer nodes and L_n hidden layers, then there is a ReLU multi-layer perceptron θ' with full previous-layer connections, H_n nodes per hidden layer, and L_n hidden layers such that $\theta(x) = \theta'(x)$, where x is the input vector, and $H_n \leq N_n L_n + d$.*

From this point, we refer multi-layer perceptrons as ReLU networks again. First, for bounded ReLU networks, $M^{**}_{n,L_d+1} \leq \max(|LB_{f_0}|, |UB_{f_0}|)$, and both $|LB_{f_0}|$ and $|UB_{f_0}|$ can be taken arbitarily large. We emphasize again that arbitary, large magnitude bounds for the bounded ReLU networks enables them to be regarded as standard ReLU networks in practice both during training and predicting. However, as the bounds no longer dependent on n, it enables the flexibility of convergence rate settings for satisfying Asymptotic

Normality conditions. Now, we instead work on the bounded ReLU networks under the assumption of very large $|LB_{f_0}|$ and $|UB_{f_0}|$.

Now, we are going to derive the conditions for satisfaction of all conditions of the Asymptotic Normality Theorem. Suppose f_0 satisfies $k = ud$, $u \in \mathbb{N}$, and a sieve-sequence error that $\varepsilon_n = n^{-a}$, $\exists a > 0$. If we assume n-polynomial growth condition on ε_n, then both H_n and L_n have n-polynomial growth rate. Therefore, $W_n = \mathcal{O}(H_n^2 L_n)$. As $\rho_n(\pi_{W_n} f_0 - f_0) \leq \varepsilon_n$ and $C^*_{n,d,L_n+1,W_n} = \mathcal{O}(W_n L_n)$, the two Asymptotic Gaussianity rate conditions can thus be written as

$$H_n^2 L_n^2 = o(n^{1/4})$$

$$\varepsilon_n = o(\min\{n^{-1/4}, n^{-1/6}(C^*_{n,d,L_n+1,W_n})^{-1/3}\})$$

where $H_n = \mathcal{O}(N_n L_n)$ by Lemma 4.4.1, where N_n is the number of hidden unit nodes in the original, possibly non-multi-layer-perceptron ReLU networks from which the ReLU sieve sequence $\{\pi_{W_n} f_0\}$ is constructed. Remark 4.4.2 tells us that

$$N_n = c n^{a/u} \left(\frac{a}{u} \ln(n) + 1\right)$$

$$L_n = c \left(\frac{a}{u} \ln(n) + 1\right)$$

and these together with the rewritten rate conditions yield

$$N_n^2 L_n^3 = \mathcal{O}\left(c^5 n^{2a/u} \left(\frac{a}{u} \ln(n) + 1\right)^5\right)$$

$$n^{-a} = o(n^{-1/4})$$

$$n^{-a} = o\left(n^{-1/6} c^{-5/3} n^{-2a/3u} \left(\frac{a}{u} \ln(n) + 1\right)^{-5/3}\right)$$

and these conditions lead to

$$n^{2a/u} < n^{1/4}, n^a > n^{1/4} \text{ and } n^{-a} < n^{-1/6} n^{-2a/3u}$$

which then simplify to

$$\frac{1}{4} < a < \frac{u}{8} \text{ and } a > \frac{u}{6u-4}. \tag{4.1}$$

The last two conditions are satisfied for every $u \geq 3$, as the function $b : (0, \infty) \to \mathbb{R}$, $b(x) = \dfrac{x}{6x-4}$ is decreasing on $[1, \infty)$.

5 Monte Carlo analysis

This simulations are meant to confirm that multi-layer ReLU network sieve estimator $\hat{f}_n$ does indeed converge to the true regression function f_0, and also their difference is asymptotically normal. As f_0 rarely has the same form as the estimating neural network $\hat{f}_n$, parameter comparisons such as those in Section 4.1 of Shen, Jiang, Sakhanenko, and Lu (2019) cannot be done for $\hat{f}_n$ and f_0 with different functional forms. Considering this impracticality of studying parameter consistency, one can study the asymptotic properties of the estimating function without bothering the parameter consistency.

5.1 Consistency of ReLU feed-forward network

We conduct a simulation of $y_i = f_0(x_i) + \epsilon_i$ for showing the probabilistic convergence of $\hat{f}_n$. We simulate x_i from the uniform distribution in $[0, 1]$, i.e., $x_i \sim \mathcal{U}[0, 1]$, and residuals are independent and identically distributed as normal distribution with mean zero and standard deviation 0.7, i.e., $\epsilon_i \sim$ i.i.d $\mathcal{N}(0, 0.7^2)$. The functions that serve as the true mean function $f_0(x)$ are

- A sigmoid function

$$f_0(x) = 5 + 18\sigma(9x - 2) - 12\sigma(2x - 9)$$

- A periodic function

$$f_0(x) = \sin(2\pi x) + \frac{1}{3}\cos(3\pi x + 3)$$

- A non-differentiable function

$$f_0(x) = \begin{cases} 8\left(\frac{1}{2} - x\right), & \text{if } x \in \left[0, \frac{1}{2}\right] \\ 10\sqrt{x - \frac{1}{2}}(2 - x), & \text{if } x \in \left(\frac{1}{2}, 1\right] \end{cases}$$

- A superposition between sigmoid and periodic function

$$f_0(x) = 5\sin(8\pi x) + 18\sigma(9x - 2) - 12\sigma(2x - 9).$$

Note that we have chosen functions that have similar functional form with simulation functions in Shen, Jiang, Sakhanenko, and Lu (2019), but with larger parameter values. Although being defined in a very short range $[0, 1]$, they have significant value variations. They are harder to fit than those functions used in Shen, Jiang, Sakhanenko, and Lu (2019), which are much gentler. This difficulty in fitting makes the comparison between the f_0 and $\hat{f}_n$ more interesting, as these two functions are much more likely to have different plots for smaller values of n. Also, as we compare the performance between multi-layer ReLU and one-layer sigmoid networks, the neural network with better performance is also more likely to show significantly better numerical and visual convergence if we use f_0 that are challenging to fit.

To conduct the simulation, we take $M^{(\gamma)}_{n,u,j,k}$ for ReLU networks and V_n for sigmoid networks (see Shen, Jiang, Sakhanenko, and Lu (2019)) to be M' and $M'r_n$, respectively,

where M' is a very large number, and one possible example of its values is $M' = 10^{100,000}$. We can replace the original Γ_n with the new compact set made by using ℓ^1-norm bounds for $\sum_{k=0}^{H_{n,u-1}} |\gamma_{u,j,k}|$, and the bounds are $M_{n,u} = \sum_{j=1}^{H_{n,u}} \sum_{k=0}^{H_{n,u-1}} M'$. Remark 4.2.2 guarantees that our output-unbounded ReLU networks can be seen as output-bounded with the upper and lower bounds that are very large and small, respectively, and these bounds are independent of sample size n.

By bounding the parameter sets and the output with a large real number, we can conduct the training minimization as an unbounded optimization, as the values of parameters are always be contained in the set in the coding implementations. This is meant to simplify the implementations, as one can use common gradient descent algorithms instead of the subgradient projection algorithm when doing unbounded minimization. The subgradient projection algorithm projects each point in each gradient descent iteration to the convex set, e.g., such as Γ_n, where the parameters are assumed to belong, and thus it simplifies to the standard gradient/subgradient descent if in each iteration the parameter stays inside the convex set.[12]

The training is done by using Keras 2.2.4 for Python 3.7 in Spyder 3.3.4. The gradient algorithm being used is Nadam with learning rate 0.001. The simulation is conducted by setting the growth rate of H_n and r_n (for one-layer sigmoid networks) to be $n^{0.4}$. For multi-layer ReLU networks, $L_d = 2$. The number of samples are $n \in \{2 \times 10^3, 5 \times 10^3, 2 \times 10^4, 5 \times 10^4\}$. For the superposition f_0, as the convergence is slower, we also consider $n \in \{2 \times 10^5, 5 \times 10^5\}$. The results can be seen in Tables 1. For $\mathbb{Q}_n(\hat{f}_n)$, the values are considered good if it is close to $\mathbb{Q}_n(f_0) = 0.49$.

[12]See Zhou and Feng (2018) and references therein.

An inspection on the errors, $\rho_n(\hat{f}_n - f_0)^2$, and least square errors, $\mathbb{Q}_n(\hat{f}_n)$, reveals two major points. Firstly, by increasing the sample size, both $\rho_n(\hat{f}_n - f_0)^2$ and $\mathbb{Q}_n(\hat{f}_n)$ converge to zero where the activation function is either ReLU or sigmoid across all simulated functions, i.e., f_0. In fact, the errors $\rho_n(\hat{f}_n - f_0)^2$ has a decreasing pattern as sample size increases for both ReLU and sigmoid. Secondly, when the simulated function f_0 has more complicated structure, the two-layer ReLU neural network outperform the one-layer sigmoid in terms of convergent rate. Overall, the consistency of estimated function $\hat{f}_n$ is confirmed by the results that are provided in Table 1.

We close this section with a detailed comparison of two-layer ReLU networks to one-layer sigmoid networks as they were used in Shen, Jiang, Sakhanenko, and Lu (2019). An inspection on Table 1 reveals that the one-layer sigmoid networks convergence speed matches the multi-layer ReLU network when f_0 is the sigmoid. This result does hold both numerically, the left half of Table 1, and visually, Figure 2. This is not surprising as the sigmoid neural networks themselves are linear combinations of sigmoid functions.

As evidenced by Figures 3-5, the two-layer ReLU can detect the fluctuating patterns and the non-differentiable point better and quicker than the one-layer sigmoid. The sigmoid networks somehow become wavy and less accurate when approaching the point of non-differentiability at larger n, Figure 4. As expected, the ReLU networks also have faster numerical convergence speed for fluctuating and non-differentiable f_0, indicated by Tables 1.

5.2 Asymptotic Normality of Sieve Estimator for ReLU Neural Network

This part focuses on the simulation of the Asymptotic Normality Theorem. For conducting the simulation, the number of nodes per hidden layer is chosen to be $H_n = 9n^{0.1}(0.1\ln(n)+1)^2$, and the hidden layer depth is $L_n = 3(0.1\ln(n)+1)$. This growth rate follows the bounding argument from Remark 4.4.2. As deep neural networks are notorious for their training difficulties, we conduct the training with batch size = 4 and epoch size = 40. Thus the training iterations are more than 8 times of those of the fixed-depth ReLU networks. The training is still conducted with the same device, operating system, Python library, method and learning rate as the consistency simulations. The training is done with the data sample size $n \in \{2\times10^3, 5\times10^3, 2\times10^4, 5\times10^4, 2\times10^5\}$.

For this simulation, the true regression functions f_0 are the first two functions in consistency simulation, and the sigmoid and periodic superposition

$$f_0(x) = 10\sin(16\pi x) + 12\sigma(2x-9) - 18\sigma(9x-2),\ x\in[0,1].$$

We choose them as all of these functions are they are infinitely differentiable, to satisfy the smoothness requirement of asymptotic normality in (4.1). As before, the true target functions f_0 used in this asymptotic normality are significantly steeper than those used in the normality simulation of Shen, Jiang, Sakhanenko, and Lu (2019). This makes getting the right estimation accuracy and stability more challenging. After the training is done, we repeat the data simulation 200 times (similar to Shen, Jiang, Sakhanenko, and Lu (2019)) to get samples of the statistics $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{f}_n(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)$. Note that the $\rho_n(\hat{f}_n - f_0)$ and

$\mathbb{Q}_n(\hat{f}_n)$ values from Tables 2 verify the consistency of the increasing-depth ReLU networks.

Then, after standardizing the samples, we construct the Q-Q plots by comparing them against $\mathcal{N}(0,1)$ and conduct normality tests on them. The statistical tests used are Kolmogorov-Smirnov, Shapiro-Wilk and d'Agostino-Pearson. We do normalize the data even for Kolmogorov-Smirnov. Our interest here is to see the form of asymptotic distribution themselves, not its parametric mean and variance, which explains why the standardization is done. Here, Kolmogorov-Smirnov is used to only to check the normality of the distribution.

The Q-Q plots from Figure 6 definitely indicate the normality of the 200 statistics' samples.[13] Almost all statistical tests results in Tables 2 do not reject the normality of these samples at 5% significance level. The notable exception is for the sigmoid f_0 when $n = 5 \times 10^4$ (Table 2), where Shapiro-Wilk and d'Agostino-Pearson reject the normality of the statistics. Our explanation is the existence of two extreme outliers that are separated from other samples in the case of sigmoid f_0 when $n = 5 \times 10^4$ (Figure 6). This makes the tail a little bit heavier. This little tail heaviness creates a problem for Shapiro-Willk test that considers the variance of the samples in testing, and also d'Agostino-Pearson, which considers samples' skewness and kurtosis. However, the general population remains on the line and thus exhibits normality. Note also that for other values of n, both of these test do not reject the normality.

[13] We got the same results for other test functions. Results are available upon request.

6 Summary and future research

We have derived three unexplored asymptotic properties of a parallel ReLU network sieve space in $\mathcal{C}^1([0,1]^d)$; consistency, sieve-based convergence rate, and asymptotic normality in the product probability space $(\Omega^*, \mathcal{A}^*, \mathbb{P}^*)$ of i.i.d errors.

We also conduct the simulations to compare the convergence of the multi-layer ReLU and the one-layer sigmoid neural networks. Overall, although both of them converge, the multi-layer ReLU networks have better convergence and pattern recognition in all cases except the sigmoid f_0, in which both of them are equally good. It is thus worthy to have the exploration of the statistical properties of ReLU neural networks, which is still uncommon.

There are several directions for future investigations. Although consistency can be derived, it requires the expansion of the errors' probability space. One can explore the consistency under the single probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Furthermore, exploring the asymptotic distribution and statistical test of multi-layer ReLU networks is also an interesting direction for future research.

Table 1: Error and least square errors for different functions of f_0.

$\rho_n(\hat{f}_n - f_0)^2$ and $\mathbb{Q}_n(\hat{f}_n)$ are the error and least square errors, respectively. n is the sample size of training data. $\hat{f}_n$ is the approximated sieve estimator. $\sigma(.)$ is a Sigmoid function. ReLU: rectified linear unit. The visualizations of the convergence for the *sigmoid*, *periodic*, *non-differentiable*, and *sigmoid and periodic* f_0 are in Figures 2, 3, 4, and 5, respectively.

	$f_0(x) = 18\sigma(9x-2) - 12\sigma(2-9x) + 5$				$f_0(x) = \sin(2\pi x) + \frac{1}{3}\cos(3\pi x + 3)$			
	$\rho_n(\hat{f}_n - f_0)^2$		$\mathbb{Q}_n(\hat{f}_n)$		$\rho_n(\hat{f}_n - f_0)^2$		$\mathbb{Q}_n(\hat{f}_n)$	
n	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid
2×10^3	13.9306	51.6624	14.6018	52.8889	0.1469	0.4475	0.6433	0.9428
5×10^3	13.6968	13.4850	14.1075	13.9171	0.0378	0.4472	0.5305	0.9281
2×10^4	0.0340	0.0439	0.5223	0.5330	0.0018	0.4413	0.4907	0.9299
5×10^4	0.0070	0.0140	0.4950	0.5025	0.0079	0.4134	0.4958	0.9077
	$f_0(x) = -8\left(x - \frac{1}{2}\right)\mathbb{1}_{\{0 \le x \le 0.5\}} + 10\sqrt{x - \frac{1}{2}}\,(2-x)\,\mathbb{1}_{\{0.5 < x \le 1\}}$				$f_0(x) = 18\sigma(9x-2) - 12\sigma(2-9x) + 5\sin(8\pi x)$			
	$\rho_n(\hat{f}_n - f_0)^2$		$\mathbb{Q}_n(\hat{f}_n)$		$\rho_n(\hat{f}_n - f_0)^2$		$\mathbb{Q}_n(\hat{f}_n)$	
n	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid	ReLU	Sigmoid
2×10^3	0.8408	3.6109	1.3705	4.1554	26.7791	56.6896	27.5223	57.9058
5×10^3	0.5048	2.3753	1.0156	2.9082	14.7668	25.8588	15.3108	26.3610
2×10^4	0.0187	0.8677	0.5058	1.3548	8.4030	12.3960	8.8938	12.8759
5×10^4	0.0194	0.1739	0.5076	0.6694	8.2855	11.4466	8.7745	11.9541
2×10^5					0.9574	7.8232	1.4476	8.3020
5×10^5					0.1372	6.4662	0.6274	6.9519

Table 2: Goodness-of-fit tests results for different functions of f_0.

$\rho_n(\hat{f}_n - f_0)^2$ and $\mathbb{Q}_n(\hat{f}_n)$ are the error and least square errors, respectively. n is the sample size of training data. $\hat{f}_n$ is the approximated sieve estimator. $\sigma(.)$ is a sigmoid function. KS: Kolmogorov-Smirnov test. SW: Shapiro-Wilk test. AP: d'Agostino-Pearson test. The Q-Q plots of the standardized data is provided in Figure 6

	$f_0(x) = 18\sigma(9x-2) - 12\sigma(2-9x) + 5$				
n	$\rho_n(\hat{f}_n - f_0)^2$	$\mathbb{Q}_n(\hat{f}_n)$	*KS(p-value)*	*SW(p-value)*	*AP(p-value)*
2×10^3	0.0722	0.5801	0.0505 (0.7300)	0.9967 (0.9485)	0.0103 (0.9948)
5×10^3	0.1054	0.5870	0.0427 (0.8668)	0.9955 (0.8232)	1.3818 (0.5011)
2×10^4	0.0897	0.5775	0.0495 (0.7050)	0.9954 (0.8200)	0.2040 (0.9029)
5×10^4	0.0722	0.5549	0.0332 (0.9783)	0.9849 (0.0314)	7.8790 (0.0194)
2×10^5	0.0490	0.5416	0.0491 (0.7105)	0.9936 (0.5432)	0.4241 (0.8089)
	$f_0(x) = \sin(2\pi x) + \frac{1}{3}\cos(3\pi x + 3)$				
n	$\rho_n(\hat{f}_n - f_0)^2$	$\mathbb{Q}_n(\hat{f}_n)$	*KS(p-value)*	*SW(p-value)*	*AP(p-value)*
2×10^3	0.0103	0.5032	0.0403 (0.9230)	0.9939 (0.5901)	1.5961 (0.4501)
5×10^3	0.0470	0.5323	0.0435 (0.8507)	0.9948 (0.7213)	2.2750 (0.3206)
2×10^4	0.0075	0.4953	0.0376 (0.9376)	0.9928 (0.4426)	1.0288 (0.5978)
5×10^4	0.0148	0.4992	0.0449 (0.8060)	0.9912 (0.2695)	5.0601 (0.0796)
2×10^5	0.0048	0.4967	0.0442 (0.8209)	0.9957 (0.8561)	0.5117 (0.7742)
	$f_0(x) = 12\sigma(2-9x) - 18\sigma(9x-2) + 10\sin(16\pi x)$				
n	$\rho_n(\hat{f}_n - f_0)^2$	$\mathbb{Q}_n(\hat{f}_n)$	*KS(p-value)*	*SW(p-value)*	*AP(p-value)*
2×10^3	41.3207	41.6786	0.0485 (0.7752)	0.9949 (0.7458)	0.0539 (0.9733)
5×10^3	21.0600	21.6952	0.0582 (0.5192)	0.9929 (0.4553)	2.3510 (0.3086)
2×10^4	0.5636	1.0433	0.0465 (0.7750)	0.9939 (0.5947)	1.2024 (0.5481)
5×10^4	0.9429	1.4291	0.0348 (0.9662)	0.9942 (0.6418)	1.3650 (0.5053)
2×10^5	0.2481	0.7400	0.0317 (0.9866)	0.9959 (0.8807)	1.5539 (0.4597)

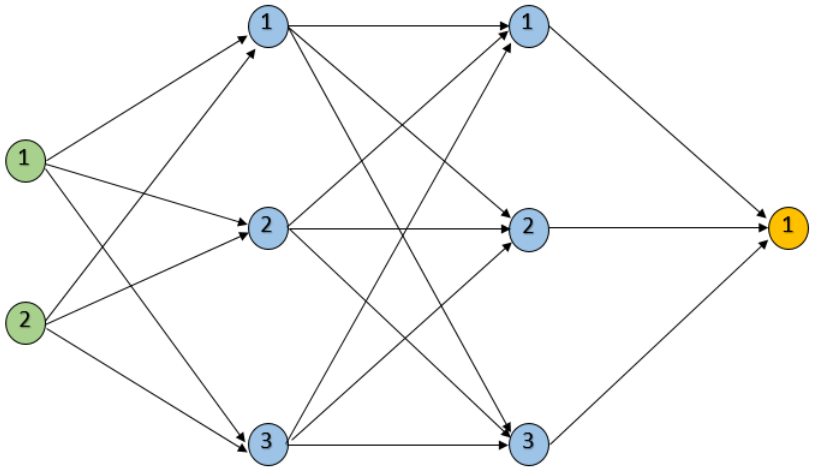

Figure 1: ReLU feed forward network.

Example of the multi-layer ReLU FFN being described. The green, blue, and yellow layers indicate input, hidden, and output layers, respectively. The number of hidden layers is L_d = 2, and it has H_n = 3 nodes per hidden layer. For the input layer, the node indices indicate the predictors (node 1 means the first predictor, 2 means the second, and so on). For the hidden and output layers, the indices indicate the ReLU function $h_{u,j}$ associated with the related nodes, with u and j is the layer and node indices, respectively, where u = 0 or u = 3 implies the input and output layers, respectively. For example, node 3 in the second hidden layer is the function $h_{2,3}(x) = ReLU\left(\sum_{k=1}^{H_n=3} \gamma_{2,3,k} \cdot h_{1,k}(x) + \gamma_{2,3,0}\right)$. A directed arrow going from node k in layer $u-1$ to node j in layer u is the parameter $\gamma_{u,j,k}$. As an example, the arrow from node 2 in the first hidden layer to node 3 in the second means parameter $\gamma_{2,3,2}$.

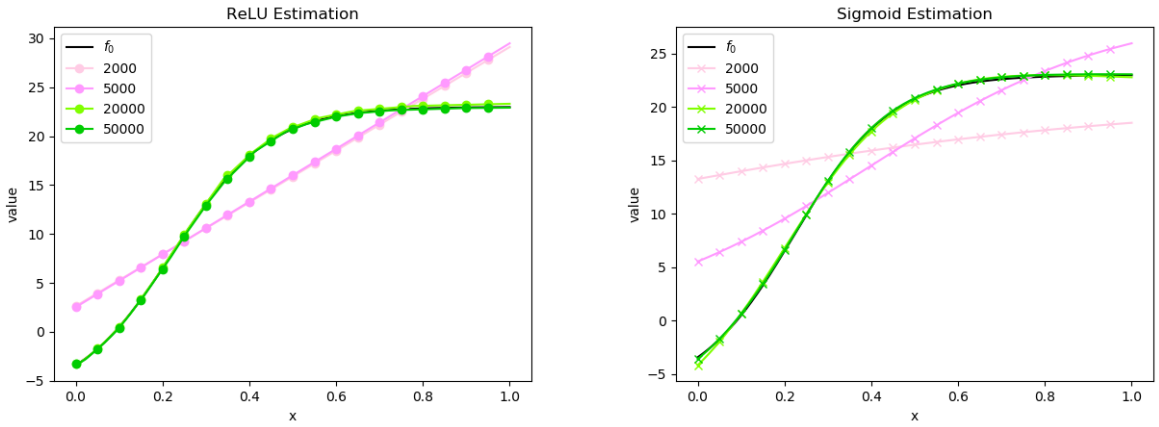

Figure 2: The pictures of multi-layer ReLU and one-layer sigmoid neural networks approximating $f_0(x) = 18\sigma(9x-2) - 12\sigma(2-9x) + 5$ for different sample size. Both batch and epoch numbers being used during the training are 32. The numbers of nodes per layer after rounding the nearest integer of $n^{0.4}$ for each choice of n are 21, 30, 53, and 76, respectively, where n is the sample size. For ReLU, the number of hidden layer L_d is 2.

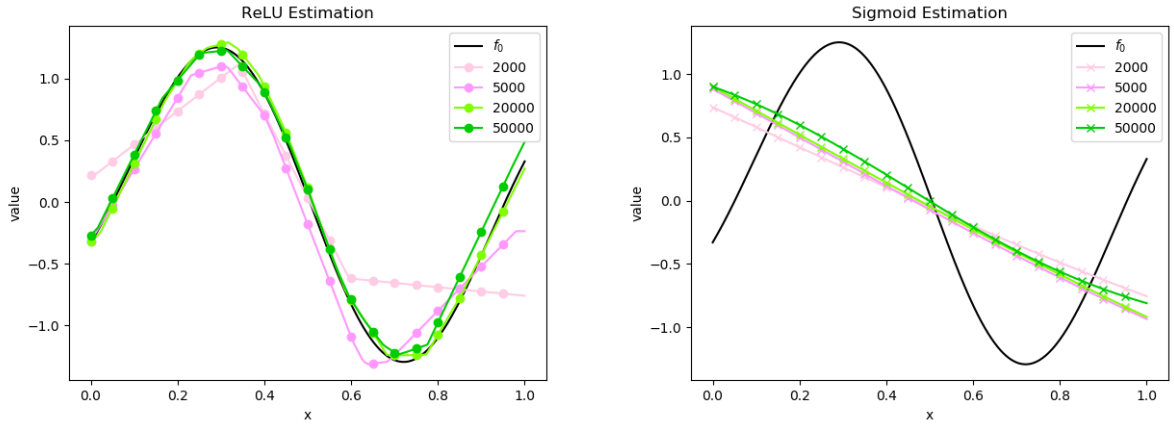

Figure 3: The pictures of multi-layer ReLU and one-layer sigmoid neural networks approximating $f_0(x) = \sin(2\pi x) + \frac{1}{3}\cos(3\pi x + 3)$ for different sample size. Both batch and epoch numbers being used during the training are 32. The numbers of nodes per layer after rounding the nearest integer of $n^{0.4}$ for each choice of n are 21,30,53, and 76, respectively, where n is the sample size. For ReLU, the number of hidden layer L_d is 2.

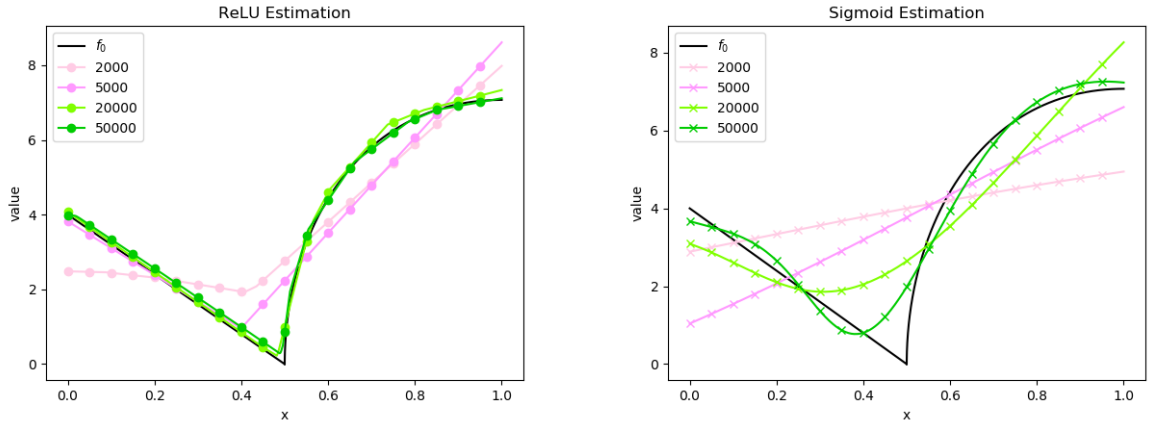

Figure 4: The pictures of multi-layer ReLU and one-layer sigmoid neural networks approximating $f_0(x) = -8\left(x - \frac{1}{2}\right)\mathbb{1}_{\{0 \le x \le 0.5\}} + 10\sqrt{x - \frac{1}{2}}\,(2 - x)\,\mathbb{1}_{\{0.5 < x \le 1\}}$ for different sample size. Both batch and epoch numbers being used during the training are 32. The numbers of nodes per layer after rounding the nearest integer of $n^{0.4}$ for each choice of n are 21, 30, 53, and 76, respectively, where n is the sample size. For ReLU, the number of hidden layer L_d is 2.

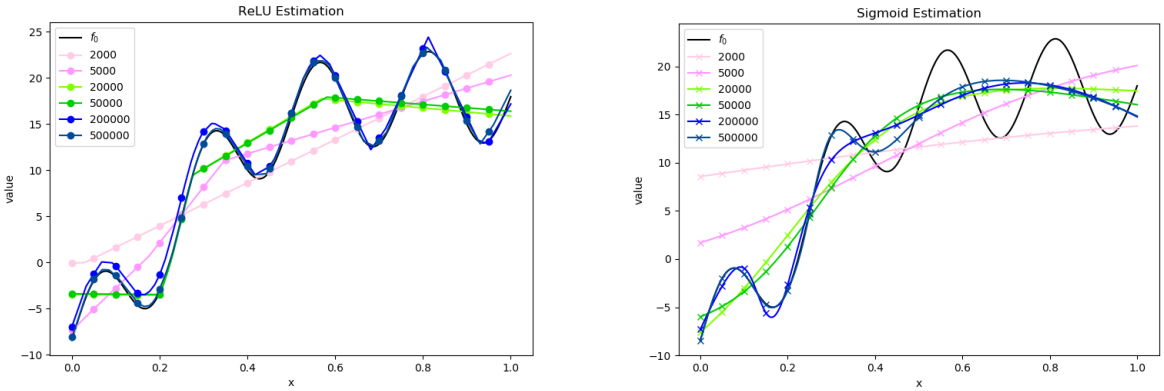

Figure 5: The pictures of multi-layer ReLU and one-layer sigmoid neural networks approximating $f_0(x) = 18\sigma(9x - 2) - 12\sigma(2 - 9x) + 5\sin(8\pi x)$ for different sample size. Both batch and epoch numbers being used during the training are 32. The numbers of nodes per layer after rounding the nearest integer of $n^{0.4}$ for each choice of n are 21, 30, 53, 76, 132, and 190, respectively, where n is the sample size. For ReLU, the number of hidden layer L_d is 2.

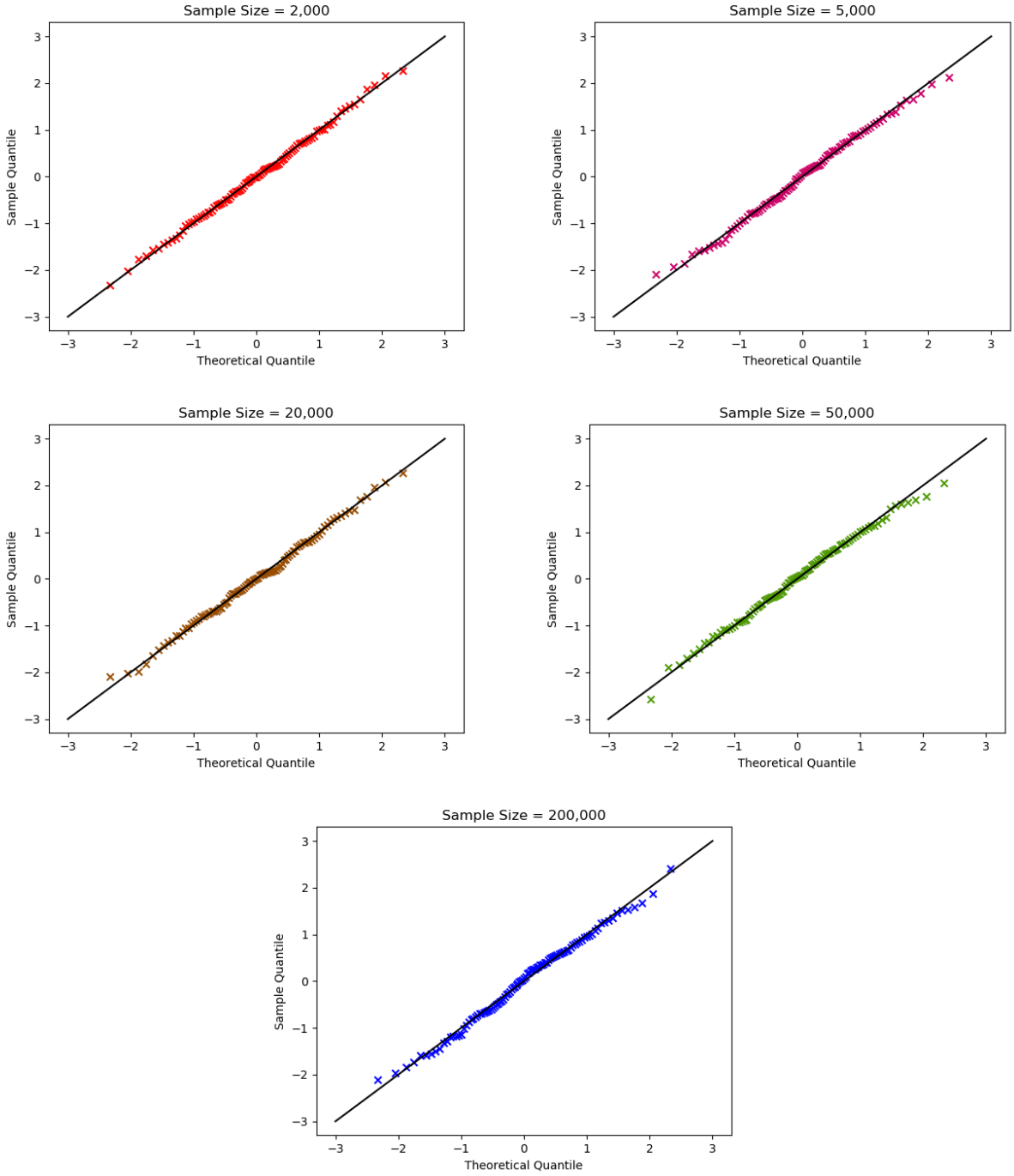

Figure 6: The Q-Q plots for multi ReLU network estimation of $f_0(x) = 5 + 18\sigma(9x - 2) - 12\sigma(2x - 9)$, $x \in [0, 1]$, for different sample sizes. The theoretical quantile is $\mathcal{N}(0, 1)$. The batch and epoch numbers used in the training are 4 and 40, respectively. The number of nodes per hidden layer and the depth of the ReLU networks are $H_n = 9n^{0.1}(0.1\ln(n) + 1)^2$ and $L_n = 3(0.1\ln(n) + 1)$, respectively, where n is the sample size.

A Proofs

Proof of Lemma 4.1.1. Suppose now $u = 1$. As a ReLU function is Lipschitz with constant 1, we have

$$\begin{aligned}\forall 1 \leq j \leq H_{n,1}, \left\|h_{1,j}\right\|_\infty &= ReLU\left(\sum_{k=1}^{d} \gamma_{1,j,k} x_k + \gamma_{1,j,0}\right) \\ &\leq \left|\sum_{k=1}^{d} \gamma_{1,j,k} x_k + \gamma_{1,j,0}\right| \leq M_{n,1}\end{aligned}$$

as we have $\mathcal{X} = [0,1]^d$. Thus we have $\underset{1\leq j\leq H_{n,1}}{sup} \left\|h_{1,j}\right\|_\infty \leq M_{n,1} = \prod_{i=0}^{2} M_{n,i}$ as $M_{n,0} = 1$

Suppose now $2 \leq u \leq L_d + 1$. Then $\forall 1 \leq j \leq H_{n,u}$,

$$\begin{aligned}\left\|h_{u,j}\right\|_\infty &= ReLU\left(\sum_{k=1}^{H_{n,u-1}} \gamma_{u,j,k} h_{u-1,k}(\boldsymbol{x}) + \gamma_{u,j,0}\right) \\ &\leq \left|\sum_{k=1}^{H_{n,u-1}} \gamma_{u,j,k} h_{u-1,k}(\boldsymbol{x}) + \gamma_{u,j,0}\right| \\ &\leq \left(\underset{1\leq k\leq H_{n,u-1}}{sup} \left\|h_{u-1,k}\right\|_\infty \vee 1\right) \sum_{k=0}^{H_n} \gamma_{u,j,k} \\ &\leq \left(\underset{1\leq k\leq H_{n,u-1}}{sup} \left\|h_{u-1,k}\right\|_\infty \vee 1\right) M_{n,u}\end{aligned}$$

We will use induction. If $u = 2$, then we have $\left\|h_{2,j}\right\|_\infty \leq \prod_{i=1}^{2} M_{n,i}$ as $M_{n,2} > 1$. Next, if we have $2 \leq u \leq L_d + 1$, then we have $\forall j$, $\left\|h_{u+1,j}\right\|_\infty \leq \left(\underset{1\leq k\leq H_{n,u}}{sup} \left\|h_{u,k}\right\|_\infty \vee 1\right) M_{n,u+1}$, and thus $\left\|h_{u+1,j}\right\|_\infty \leq \prod_{i=0}^{u+1} M_{n,i}$ as $\forall u$, $M_{n,u} \geq 1$.

Hence, $\left\|h_{L_d+1,1}\right\|_\infty \leq \prod_{i=1}^{L_d+1} M_{n,i}$, and the conclusion follows. □

Proof of EC2 and EC3 Satisfaction. For each fixed $\omega \in \Omega$, we have $\mathbb{Q}_n^{(\omega)}$:$(\mathcal{F}_{W_n}, \rho_n) \to ([0,\infty), |\cdot|)$ a mapping from pseudo-metric space to metric space. By Triangle Inequality,

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

$$\begin{aligned}
\left|\mathbb{Q}_n^{(\omega)}(f) - \mathbb{Q}_n^{(\omega)}(g)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left[(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}))^2 - (g(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}))^2\right]\right. \\
&\qquad \left. -\frac{2}{n}\sum_{i=1}^{n}[\epsilon_i(\omega)(f(\boldsymbol{x_i}) - g(\boldsymbol{x_i}))]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n}[(f(\boldsymbol{x_i}) - g(\boldsymbol{x_i}))(f(\boldsymbol{x_i}) + g(\boldsymbol{x_i}) - 2f_0(\boldsymbol{x_i}))]\right| \\
&\quad + \left|\frac{2}{n}\sum_{i=1}^{n}[\epsilon_i(\omega)(f(\boldsymbol{x_i}) - g(\boldsymbol{x_i}))]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}[|f(\boldsymbol{x_i}) - g(\boldsymbol{x_i})|\,(|f(\boldsymbol{x_i}) + g(\boldsymbol{x_i}) + 2f_0(\boldsymbol{x_i})|)] \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}[|\epsilon_i(\omega)||f(\boldsymbol{x_i}) - g(\boldsymbol{x_i})|]
\end{aligned}$$

As f_0 is continuous function on a compact domain $[0,1]^d$, we can use Extreme Value Theorem to get

$$\begin{aligned}
\left|\mathbb{Q}_n^{(\omega)}(f) - \mathbb{Q}_n^{(\omega)}(g)\right| &\leq \frac{2M^*_{n,L_d+1} + 2\underset{\boldsymbol{x}\in[0,1]^d}{max}|f_0(\boldsymbol{x})| + 2\underset{1\leq i\leq n}{sup}\,\epsilon_i(\omega)}{n}\sum_{i=1}^{n}[|f(\boldsymbol{x_i}) - g(\boldsymbol{x_i})|] \\
&\leq D_{n,L_d+1,\omega}\|f - g\|_1
\end{aligned}$$

with $D_{n,L_d+1,\omega} = n^{-1}\left(2M^*_{n,L_d+1} + 2\underset{\boldsymbol{x}\in[0,1]^d}{max}|f_0(\boldsymbol{x})| + 2\underset{1\leq i\leq n}{sup}\,\epsilon_i(\omega)\right)$, and $\|\cdot\|_k$ indicates the k^{th} Euclidean norm.

From the equivalence of Euclidean norms, $\exists V^* > 0$ such that $\|f - g\|_1 \leq V^*\|f - g\|_2$, which leads to

$$\left|\mathbb{Q}_n^{(\omega)}(f) - \mathbb{Q}_n^{(\omega)}(g)\right| \leq \sqrt{n}D_{n,L_d+1,\omega}V^*\rho_n(f - g)^2$$

Hence, $\forall\,\zeta > 0$, we can take $\delta = \zeta^{1/2}\left(\sqrt{n}D_{n,L_d+1,\omega}V^*\right)^{-1/2} > 0$ such that $\forall\, f, g \in \mathcal{F}_{W_n}$, $\left|\mathbb{Q}_n^{(\omega)}(f) - \mathbb{Q}_n^{(\omega)}(g)\right| < \zeta$ whenever $\rho_n(f - g) < \delta$, implying continuity and hence EC2.

Next, we will prove EC3. Note that

$$\rho_n(\pi_{W_n} f_0 - f_0) \leq \|\pi_{W_n} f_0 - f_0\|_\infty \leq \mathcal{O}\left(\omega_{f_0}\left(\mathcal{O}\left(W_n^{-1/d}\right)\right)\right)$$

and thus it is clear that $\mathcal{F}_{W_n}$ is still a sieve space of $\mathcal{F}$ under ρ_n, as long as we have the Denseness Assumption.

To prove the compactness of the pseudo-metric space $(\mathcal{F}_{W_n}, \rho_n)$, use the following mapping

$$F : (\Gamma_n, \|\cdot\|_2) \to (\mathcal{F}_{W_n}, \rho_n)$$

$$[\gamma_{u,j,k}] \mapsto F([\gamma_{u,j,k}]) = h_{L_d+1,1}(\boldsymbol{x} \mid [\gamma_{u,j,k}])$$

with $h_{L_d+1,1}(\boldsymbol{x} \mid [\gamma_{u,j,k}])$ uses $[\gamma_{u,j,k}]$ as its parameters in ReLU linear combinations. Definitely $\mathcal{F}_{W_n} = F(\Gamma_n)$.

Now, we prove the continuity of F. For every $[\gamma^{(1)}_{u,j,k}]$, $[\gamma^{(2)}_{u,j,k}] \in \Gamma_n$, we have

$$\begin{aligned}\rho_n\left(F([\gamma^{(1)}_{u,j,k}]) - F([\gamma^{(2)}_{u,j,k}])\right)^2 =& \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{H_{n,L_d}} \gamma^{(1)}_{L_d+1,1,k} \cdot h_{L_d,k}(\boldsymbol{x_i} \mid [\gamma^{(1)}_{u,j,k}]) + \gamma^{(1)}_{L_d+1,1,0}\right.\\ &\left. - \sum_{k=1}^{H_{n,L_d}} \gamma^{(2)}_{L_d+1,1,k} \cdot h_{L_d,k}(\boldsymbol{x_i} \mid [\gamma^{(2)}_{u,j,k}]) - \gamma^{(2)}_{L_d+1,1,0}\right)^2\end{aligned}$$

Take $\gamma_{u,j}$ and h_u as column matrices with element orders corresponding to $k = 1, 2, ..., H_{n,u}$. This notations together with the Triangle Inequality and the fact that ReLU is Lipschitz with constant 1 lead to

$$\begin{aligned}&\rho_n\left(F([\gamma^{(1)}_{u,j,k}]) - F([\gamma^{(2)}_{u,j,k}])\right)^2\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{H_{n,L_d}} \gamma^{(1)}_{L_d+1,1,k} \cdot ReLU\left(\left(\gamma^{(1)}_{L_d,k}\right)^\top \cdot h_{L_d-1}(\boldsymbol{x_i} \mid [\gamma^{(1)}_{u,k,[1:H_{n,L_d-1}]}]) + \gamma^{(1)}_{L_d,k,0}\right) + \gamma^{(1)}_{L_d+1,1,0}\right.\\ &\quad\left. - \sum_{k=1}^{H_{n,L_d}} \gamma^{(2)}_{L_d+1,1,k} \cdot ReLU\left(\left(\gamma^{(2)}_{L_d,k}\right)^\top \cdot h_{L_d-1}(\boldsymbol{x_i} \mid [\gamma^{(2)}_{u,k,[1:H_{n,L_d-1}]}]) + \gamma^{(2)}_{L_d,k,0}\right) - \gamma^{(2)}_{L_d+1,1,0}\right)^2\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\Bigg(\left|\gamma^{(1)}_{L_d+1,1,0}-\gamma^{(2)}_{L_d+1,1,0}\right|+\sum_{k=1}^{H_{n,L_d}}\Bigg|\gamma^{(1)}_{L_d+1,1,k}\cdot ReLU\left(\left(\gamma^{(1)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(1)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(1)}_{L_d,k,0}\right)\\
&\quad -\gamma^{(2)}_{L_d+1,1,k}\cdot ReLU\left(\left(\gamma^{(2)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(2)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(2)}_{L_d,k,0}\right)\Bigg|\Bigg)^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Bigg(\left|\gamma^{(1)}_{L_d+1,1,0}-\gamma^{(2)}_{L_d+1,1,0}\right|+\\
&\quad \sum_{k=1}^{H_{n,L_d}}\Bigg|\gamma^{(1)}_{L_d+1,1,k}\left(ReLU\left(\left(\gamma^{(1)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(1)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(1)}_{L_d,k,0}\right)\right.\\
&\quad \left.-ReLU\left(\left(\gamma^{(2)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(2)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(2)}_{L_d,k,0}\right)\right)\Bigg|+\\
&\quad \sum_{k=1}^{H_{n,L_d}}\Bigg[\left|\gamma^{(1)}_{L_d+1,1,k}-\gamma^{(2)}_{L_d+1,1,k}\right|\times\\
&\quad \left|ReLU\left(\left(\gamma^{(2)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(2)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(2)}_{L_d,k,0}\right)\right|\Bigg]\Bigg)^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Bigg(\left|\gamma^{(1)}_{L_d+1,1,0}-\gamma^{(2)}_{L_d+1,1,0}\right|+\\
&\quad \sum_{k=1}^{H_{n,L_d}}\Bigg|\gamma^{(1)}_{L_d+1,1,k}\left(\left(\gamma^{(1)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(1)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(1)}_{L_d,k,0}\right.\\
&\quad \left.-\left(\gamma^{(2)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(2)}_{u,k,[1:H_{n,L_d-1}]}])-\gamma^{(2)}_{L_d,k,0}\right)\Bigg|+\\
&\quad \sum_{k=1}^{H_{n,L_d}}\left[\left|\gamma^{(1)}_{L_d+1,1,k}-\gamma^{(2)}_{L_d+1,1,k}\right|\left|\left(\gamma^{(2)}_{L_d,k}\right)^{\top}\cdot h_{L_d-1}(\boldsymbol{x_i}\mid[\gamma^{(2)}_{u,k,[1:H_{n,L_d-1}]}])+\gamma^{(2)}_{L_d,k,0}\right|\right]\Bigg)^2
\end{aligned}
$$

By Lemma 4.1.1 and the Triangle Inequality, we have

$$
\begin{aligned}
&\rho_n\left(F([\gamma^{(1)}_{u,j,k}])-F([\gamma^{(2)}_{u,j,k}])\right)^2\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\Bigg(\left|\gamma^{(1)}_{L_d+1,1,0}-\gamma^{(2)}_{L_d+1,1,0}\right|+\\
&\quad M^{*}_{n,L_d-1}\sum_{k=1}^{H_{n,L_d}}\left[\left|\gamma^{(1)}_{L_d+1,1,k}\right|\left\|\gamma^{(1)}_{L_d,k}-\gamma^{(2)}_{L_d,k}\right\|_1+\left|\gamma^{(1)}_{L_d+1,1,k}\right|\left|\gamma^{(1)}_{L_d,k,0}-\gamma^{(2)}_{L_d,k,0}\right|\right]+
\end{aligned}
$$

$$
\begin{aligned}
& M^*_{n,L_d-1} \sum_{k=1}^{H_{n,L_d}} \left[\left| \gamma^{(1)}_{L_d+1,1,k} - \gamma^{(2)}_{L_d+1,1,k} \right| \left(\left\| \gamma^{(2)}_{L_d,k} \right\|_1 + \left| \gamma^{(2)}_{L_d,k,0} \right| \right) \right] \Bigg)^2 \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \Bigg(\left| \gamma^{(1)}_{L_d+1,1,0} - \gamma^{(2)}_{L_d+1,1,0} \right| + \\
& M^*_{n,L_d-1} \sum_{k=1}^{H_{n,L_d}} \left[\left| \gamma^{(1)}_{L_d+1,1,k} \right| \left\| \gamma^{(1)}_{L_d,k} - \gamma^{(2)}_{L_d,k} \right\|_1 + \left| \gamma^{(1)}_{L_d+1,1,k} \right| \left| \gamma^{(1)}_{L_d,k,0} - \gamma^{(2)}_{L_d,k,0} \right| \right] + \\
& M^*_{n,L_d-1} \sum_{k=1}^{H_{n,L_d}} \left[\left| \gamma^{(1)}_{L_d+1,1,k} - \gamma^{(2)}_{L_d+1,1,k} \right| \left(\left\| \gamma^{(2)}_{L_d,k} \right\|_1 + \left| \gamma^{(2)}_{L_d,k,0} \right| \right) \right] \Bigg)^2 \\
\leq & \frac{1}{n} \sum_{i=1}^{n} \Bigg(\left| \gamma^{(1)}_{L_d+1,1,0} - \gamma^{(2)}_{L_d+1,1,0} \right| + \\
& M^*_{n,L_d-1} M_{n,L_d+1} \sum_{k=1}^{H_{n,L_d}} \left[\left\| \gamma^{(1)}_{L_d,k} - \gamma^{(2)}_{L_d,k} \right\|_1 + \left| \gamma^{(1)}_{L_d,k,0} - \gamma^{(2)}_{L_d,k,0} \right| \right] + \\
& M^*_{n,L_d-1} M_{n,L_d} \sum_{k=1}^{H_{n,L_d}} \left[\left| \gamma^{(1)}_{L_d+1,1,k} - \gamma^{(2)}_{L_d+1,1,k} \right| \right] \Bigg)^2 \\
\leq & \left(M^*_{n,L_d+1} \right)^2 \Bigg(\left| \gamma^{(1)}_{L_d+1,1,0} - \gamma^{(2)}_{L_d+1,1,0} \right| + \sum_{k=1}^{H_{n,L_d}} \left[\left\| \gamma^{(1)}_{L_d,k} - \gamma^{(2)}_{L_d,k} \right\|_1 + \left| \gamma^{(1)}_{L_d,k,0} - \gamma^{(2)}_{L_d,k,0} \right| \right] + \\
& \sum_{k=1}^{H_{n,L_d}} \left[\left| \gamma^{(1)}_{L_d+1,1,k} - \gamma^{(2)}_{L_d+1,1,k} \right| \right] \Bigg)^2 \leq \left(M^*_{n,L_d+1} \right)^2 \left\| [\gamma^{(1)}_{u,j,k}] - [\gamma^{(2)}_{u,j,k}] \right\|_1^2
\end{aligned}
$$

Using the equivalence of Euclidean norms, $\exists V^* > 0$ such that $\left\| [\gamma^{(1)}_{u,j,k}] - [\gamma^{(2)}_{u,j,k}] \right\|_1 \leq V^* \left\| [\gamma^{(1)}_{u,j,k}] - [\gamma^{(2)}_{u,j,k}] \right\|_2$, and hence $\rho_n \left(F([\gamma^{(1)}_{u,j,k}]) - F([\gamma^{(2)}_{u,j,k}]) \right)^2 \leq \left(M^*_{n,L_d+1} V^* \left\| [\gamma^{(1)}_{u,j,k}] - [\gamma^{(2)}_{u,j,k}] \right\|_2 \right)^2$.

Hence, for every $\zeta > 0$, we can take $\delta = \left(M^*_{n,L_d+1} V^* \right)^{-1} \zeta > 0$ such that $\left\| [\gamma^{(1)}_{u,j,k}] - [\gamma^{(2)}_{u,j,k}] \right\|_2 < \delta$ leads to $\rho_n \left(F([\gamma^{(1)}_{u,j,k}]) - F([\gamma^{(2)}_{u,j,k}]) \right)^2 < \zeta$, implying the continuity of F.

It is obvious that $(\Gamma_n, \| \cdot \|_2)$ is compact. Hence, as every continuous image of a compact set is compact, we prove the third existence condition. □

Proof of the CC6 Satisfaction Lemma. The proof is almost the same with the proof of Lemma

3.2. in Shen, Jiang, Sakhanenko, and Lu (2019). For any $\delta > 0$ and W_n, we have

$$\begin{aligned}
&\mathbb{P}^* \left(\sup_{f \in \mathcal{F}_{W_n}} |\mathbb{Q}_n(f) - Q_n(f)| > \delta \right) \\
&= \mathbb{P}^* \left(\sup_{f \in \mathcal{F}_{W_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i^2 - \sigma^2 \right] - \frac{2}{n} \sum_{i=1}^{n} \left[\epsilon_i \left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}) \right) \right] \right| > \delta \right) \\
&\leq \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i^2 - \sigma^2 \right] \right| > \frac{\delta}{2} \right) + \mathbb{P}^* \left(\sup_{f \in \mathcal{F}_{W_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i \left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}) \right) \right] \right| > \frac{\delta}{4} \right)
\end{aligned}$$

Because $\mathbb{E}\left[\epsilon_i^2\right] = \sigma^2$, the Weak Law of Large Numbers gives $\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i^2 - \sigma^2 \right] \right| > \frac{\delta}{2} \right)$ $\to 0$ as $n \to \infty$.

Now we need to show that

$$\lim_{n \to \infty} \mathbb{P}^* \left(\sup_{f \in \mathcal{F}_{W_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i \left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}) \right) \right] \right| > \frac{\delta}{4} \right) = 0$$

By the Markov's Inequality, this is satisfied if

$$\lim_{n \to \infty} \mathbb{E}_{\mathbb{P}^*} \left[\sup_{f \in \mathcal{F}_{W_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \left[\epsilon_i \left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i}) \right) \right] \right| \right] = 0$$

Our intention is to use the Symmetrization Inequality (Lemma 2.3.1. in Vaart and Wellner (1996)) to get a Rademacher sequence upper bound. Define $Y_1\left(\epsilon, f(\boldsymbol{x})\right) \coloneqq \epsilon_i \left(f(\boldsymbol{x}) - f_0(\boldsymbol{x}) \right)$, for every $f \in \mathcal{F}_{W_n}$. Clearly, $\mathbb{E}\left[Y_1\left(\epsilon, f(\boldsymbol{x})\right)\right] = 0$. We need to show that $Y_1\left(\epsilon, f(\boldsymbol{x})\right)$ is measurable w.r.t. $(\Omega, \mathcal{A}, \mathbb{P})$.

Suppose $\left[\gamma_{u,j,k}\right]_n \in \Gamma_n$ from (1). As Γ_n is compact, we have a sequence $\left[\gamma_{u,j,k}\right]_{n,m} \in \mathbb{Q}^{W_n}$ $\cap\, \Gamma_n$ converging to $\left[\gamma_{u,j,k}\right]_n$ when $m \to \infty$. The function F in EC2 and EC3 Satisfaction proof maps $\left[\gamma_{u,j,k}\right]_{n,m}$ to $f_{m,n} \in \mathcal{F}_{W_n}$.

Since F is continuous, $f_{m,n} \to f$ pointwise, the Example 2.3.4 in Vaart and Wellner (1996) implies $\mathcal{F}_{W_n}$ members are measurable w.r.t. $(\Omega, \mathcal{A}, \mathbb{P})$, and thus so is $Y_1\left(\epsilon, f(\boldsymbol{x})\right)$.

Hence, by the Symmetrization Inequality

$$\mathbb{E}_{\mathbb{P}^*}\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right]=\mathbb{E}_{\mathbb{P}^*}\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[Y_1\left(\epsilon_i,f(\boldsymbol{x_i})\right)\right]\right|\right]$$
$$\leq 2\mathbb{E}_{\mathbb{P}^*,\epsilon}\left[\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[R_iY_1\left(\epsilon_i,f(\boldsymbol{x_i})\right)\right]\right|\right]\right]$$
$$\leq 2\mathbb{E}_{\mathbb{P}^*,\epsilon}\left[\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[R_i\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right]\right]$$

with R_i are i.i.d Rademacher random variables living in $(\mathcal{Z},\mathcal{C},\mathbb{P}_{\mathcal{Z}})$, independent of $\prod_{i=1}^{n}(\Omega,\mathcal{A},\mathbb{P})$, and also, $\mathbb{E}_{\mathbb{P}^*,\epsilon}$ and $\mathbb{E}_R$ mean the expectation is taken w.r.t. ϵ and the Rademacher variables, respectively.

Define $Y_2\left(\omega',f,n\right):=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(R_i(\omega')\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right)$, for each $f\in\mathcal{F}_{W_n}$, $\omega'\in\mathcal{Z}$. Fix ϵ_1, ϵ_2,...., ϵ_n. Hence, $Y_2\left(\omega',f,n\right)$ is a sub-Gaussian process indexed by f. It is easy to see that $\forall\ \omega'\in\mathcal{Z}$, $Y_2\left(\omega',f,n\right)$ is continuous w.r.t. $(\mathcal{F}_{W_n},\rho_n)$. Hence, with the sequence $f_{n,m}$ converging pointwise to f_n above, we have $Y_2\left(\omega',f_{m,n},n\right)\to Y_2\left(\omega',f_n,n\right)$, for every $\omega'\in\mathcal{Z}$.

By Section 2.3.3* and Corollary 2.2.8. in Vaart and Wellner (1996), respectively, $Y_2\left(\omega',f,n\right)$ is a separable sub-Gaussian process defined on $(\mathcal{F}_{W_n},\rho_n)$, and $\exists\ C>0$ such that $\forall\ f_n^*\in\mathcal{F}_{W_n}$, we have

$$\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\left[R_i\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right]$$

$$=\frac{1}{\sqrt{n}}\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|Y_2\left(\cdot,f,n\right)\right|\right]$$

$$\leq\mathbb{E}_R\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left[R_i\epsilon_i\left(f_n^*(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right]+C\int_0^{\infty}\sqrt{\frac{\ln\left(N\left(\frac{1}{2}\eta,\mathcal{F}_{W_n},d_n\right)\right)}{n}}d\eta$$

with $d_n(f,g) \coloneqq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\epsilon_i^2\right)}$. It is clear that d_n is also a pseudo-distance. The term $N\left(\frac{1}{2}\eta, \mathcal{F}_{W_n}, d_n\right)$ here denotes the minimum number of balls with radius $\frac{1}{2}\eta$ and distance concept d_n being required to cover $\mathcal{F}_{W_n}$. Its natural logarithm is a metric entropy of $\mathcal{F}_{W_n}$.

Obviously, $d_n(f,g) \leq \|f-g\|_\infty \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2}\right)$. The Strong Law of Large Numbers tells us that for any $n \geq N_1$, $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 < \sigma^2+1$ almost everywhere except when ϵ_i take values in a null set E of $\prod_{i=1}^{n}(\Omega, \mathcal{A}, \mathbb{P})$.

By the Cauchy-Schwartz inequality, for any $n \geq N_1$, we have

$$\begin{aligned}
\mathbb{E}_R\left[\left|\frac{1}{n}\sum_{i=1}^{n} R_i\epsilon_i\left(f_n^*(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right|\right] &\leq \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|\,|f_n^*(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})| \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_n^*(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)^2} \\
&\leq \sqrt{\sigma^2+1}\,\|f_n^* - f_0\|_\infty \text{ a.e.}
\end{aligned}$$

Take $f_n^* = \pi_{W_n} f_0$. Since $\|\pi_{W_n} f_0 - f_0\|_\infty \to 0$ as $n \to \infty$, for every $\zeta > 0$, there exists $N_2 > 0$ such that for any $n \geq N_2$

$$\|\pi_{W_n} f_0 - f_0\|_\infty < \frac{\zeta}{\sqrt{\sigma^2+1}}$$

Thus every $n > N_1 \vee N_2$ satisfies

$$\mathbb{E}_R\left[\left|\frac{1}{n}\sum_{i=1}^{n} R_i\epsilon_i\left(\pi_{W_n} f_0(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right|\right] < \zeta \text{ a.e.}$$

Next, we are going to bound the second integral term with a term that converges to 0 when $n \to \infty$ almost everywhere. We know that $d_n(f,g) \leq \|f-g\|_\infty \sqrt{\sigma^2+1}$ almost everywhere. Therefore, a ball $B_{d_n}\left(f; \frac{1}{2}\eta\right) \coloneqq \left\{g \in \mathcal{F}_{W_n} : d_n(f,g) < \frac{1}{2}\eta\right\} \supset$

$$\left\{g \in \mathcal{F}_{W_n} : \|f-g\|_\infty < \frac{\eta}{2\sqrt{\sigma^2+1}}\right\} =: B_{\|\cdot\|_\infty}\left(f; \frac{\eta}{2\sqrt{\sigma^2+1}}\right) \text{ almost everywhere, for every } f, g \in \mathcal{F}_{W_n}.$$

This implies

$$\int_0^\infty \sqrt{\frac{ln\left(N\left(\frac{1}{2}\eta, \mathcal{F}_{W_n}, d_n\right)\right)}{n}} d\eta \leq \int_0^\infty \sqrt{\frac{ln\left(N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right)}{n}} d\eta$$
$$\leq \int_0^{2M^*_{n,L_d+1}} \sqrt{\frac{\ln\left(N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right)}{n}} d\eta$$

as we have $\|f-g\|_\infty \leq 2M^*_{n,L_d+1}$, for every $f, g \in \mathcal{F}_{W_n}$.

From Theorem 14.5 in Anthony and Bartlett (2009), $\forall\ \eta \leq 2M^*_{n,L_d+1}$,

$$\ln\left(N\left(\frac{\eta}{2\sqrt{\sigma^2+1}}, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right) \leq W_n \cdot ln\left(\frac{8\sqrt{\sigma^2+1}\cdot e\cdot d\cdot M^*_{n,L_d+1}\cdot W_n\cdot\left(M^{(all)}_{n,L_d+1}\right)^{L_d+1}}{\eta\cdot\left(M^{(all)}_{n,L_d+1}-1\right)}\right)$$

Define

$$\widetilde{U}_{n,d,L_d+1,W_n} := \left(\frac{8\sqrt{\sigma^2+1}\cdot e\cdot d\cdot M^*_{n,L_d+1}\cdot W_n\cdot\left(M^{(all)}_{n,L_d+1}\right)^{L_d+1}}{M^{(all)}_{n,L_d+1}-1}\right)^{W_n}$$

$$U_{n,d,L_d+1,W_n} := \ln\left(\widetilde{U}_{n,d,L_d+1,W_n}\right) - W_n$$
$$= W_n\cdot\left(\ln\left(\frac{8\sqrt{\sigma^2+1}\cdot e\cdot d\cdot M^*_{n,L_d+1}\cdot W_n\cdot\left(M^{(all)}_{n,L_d+1}\right)^{L_d+1}}{M^{(all)}_{n,L_d+1}-1}\right)-1\right)$$
$$= W_n\cdot\left(\ln\left(\frac{d\cdot M^*_{n,L_d+1}\cdot W_n\cdot\left(M^{(all)}_{n,L_d+1}\right)^{L_d+1}}{M^{(all)}_{n,L_d+1}-1}\right)+ln\left(8\sqrt{\sigma^2+1}\right)\right)$$
$$\leq 2W_n\cdot\ln\left(\frac{d\cdot M^*_{n,L_d+1}\cdot W_n\cdot\left(M^{(all)}_{n,L_d+1}\right)^{L_d+1}}{M^{(all)}_{n,L_d+1}-1}\right), \text{ for each } n \geq N_1 \vee N_3$$

by choosing $N_3 > 0$ that allows $d\cdot M^*_{n,L_d+1}\cdot W_n \geq 8\sqrt{\sigma^2+1}$. As we have $M^{(all)}_{n,L_d+1} > 1$,

$\frac{\left(M_{n,L_d+1}^{(all)}\right)^{L_d+1}}{M_{n,L_d+1}^{(all)}-1} > 1$, and thus $ln\left(\frac{d \cdot M_{n,L_d+1}^{*} \cdot W_n \cdot \left(M_{n,L_d+1}^{(all)}\right)^{L_d+1}}{M_{n,L_d+1}^{(all)}-1}\right) > ln\left(8\sqrt{\sigma^2+1}\right)$.

Since for every $\eta \leq 2M_{n,L_d+1}^{*}$,

$$\begin{aligned}\ln\left(N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right) &\leq U_{n,d,L_d+1,W_n} + W_n \cdot \ln\left(\frac{1}{\eta}\right) \\ &\leq U_{n,d,L_d+1,W_n} + \frac{W_n}{\eta} \\ &\leq U_{n,d,L_d+1,W_n}\left(1+\frac{1}{\eta}\right)\end{aligned}$$

we have $\forall\, n \geq N_1 \vee N_3$,

$$\begin{aligned}\int_0^{2M_{n,L_d+1}^{*}} \sqrt{\frac{\ln\left(N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right)}{n}} d\eta &\leq \sqrt{\frac{U_{n,d,L_d+1,W_n}}{n}} \int_0^{2M_{n,L_d+1}^{*}} \sqrt{1+\frac{1}{\eta}} d\eta \\ &\leq 5\sqrt{2}\sqrt{\frac{U_{n,d,L_d+1,W_n}}{n}} \cdot M_{n,L_d+1}^{*}\end{aligned}$$

Now, for large $M_{n,L_d+1}^{(all)}$, we have $\frac{\left(M_{n,L_d+1}^{(all)}\right)^{L_d+1}}{M_{n,L_d+1}^{(all)}-1} \sim \left(M_{n,L_d+1}^{(all)}\right)^{L_d}$. Hence, $\exists\, N_4 > 0$ such that for any $n \geq N_4$, we have

$$\begin{aligned}&\int_0^{2M_{n,L_d+1}^{*}} \sqrt{\frac{\ln\left(N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right)}{n}} d\eta \\ \lesssim\, &10\sqrt{\frac{\left(M_{n,L_d+1}^{*}\right)^2 \cdot W_n \cdot ln\left(d \cdot M_{n,L_d+1}^{*} \cdot W_n \cdot \left(M_{n,L_d+1}^{(all)}\right)^{L_d}\right)}{n}} \\ \lesssim\, &10\sqrt{\frac{\left(M_{n,L_d+1}^{*}\right)^2 C_{n,d,L_d+1,W_n}^{*}}{n}}\end{aligned}$$

Therefore, if $\left(M^*_{n,L_d+1}\right)^2 \cdot C^*_{n,d,L_d+1,W_n} = o(n)$, every $n \geq N_1 \vee N_2 \vee N_3 \vee N_4$ satisfies

$$\int_0^{2M^*_{n,L_d+1}} \sqrt{\frac{ln\left(N\left(\frac{1}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right)}{n}} d\eta \to 0 \text{ as } n \to \infty \text{ a.e.}$$

which implies

$$\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^n\left[R_i\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right] \to 0 \text{ as } n \to \infty \text{ a.e.}$$

Thus for any $n \geq N_1 \vee N_2 \vee N_3 \vee N_4$, we have

$$\begin{aligned}
&\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^n\left[R_i\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right] \\
&\leq \sqrt{\sigma^2+1}\left\|\pi_{W_n}f_0 - f_0\right\|_\infty + 10C\sqrt{\frac{\left(M^*_{n,L_d+1}\right)^2 C^*_{n,d,L_d+1,W_n}}{n}} \to 0 \text{ a.e.}
\end{aligned}$$

It is clear that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^*,\epsilon}\left[\sqrt{\sigma^2+1}\left\|\pi_{W_n}f_0 - f_0\right\|_\infty + 10C\sqrt{\frac{\left(M^*_{n,L_d+1}\right)^2 C^*_{n,d,L_d+1,W_n}}{n}}\right] \\
&=\leq \sqrt{\sigma^2+1}\left\|\pi_{W_n}f_0 - f_0\right\|_\infty + 10C\sqrt{\frac{\left(M^*_{n,L_d+1}\right)^2 C^*_{n,d,L_d+1,W_n}}{n}} \to 0 < \infty.
\end{aligned}$$

Therefore, we can use Generalized Dominated Convergence Theorem (from the completeness of $(\Omega, \mathcal{A}, \mathbb{P})$) to conclude

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}^*}\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^n\left[\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right] \\
&\leq 2\mathbb{E}_{\mathbb{P}^*,\epsilon}\left[\mathbb{E}_R\left[\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^n\left[R_i\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right|\right]\right] \to 0 \text{ a.e.}
\end{aligned}$$

and this certainly implies that under $(\Omega^*, \mathcal{A}^*, \mathbb{P}^*)$),

$$\operatorname*{plim}_{n\to\infty}\sup_{f\in\mathcal{F}_{W_n}}\left|\frac{1}{n}\sum_{i=1}^n\left[\epsilon_i\left(f(\boldsymbol{x_i})-f_0(\boldsymbol{x_i})\right)\right]\right| = 0$$

and as we have $\ln\left(N\left(\frac{\eta}{2\sqrt{\sigma^2+1}}\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right) = o(n)$, the equality above holds for each W_n, which finishes the proof. □

Proof of the Rate of Convergence Theorem. Take $\delta_n = \rho_n(\pi_{W_n} f_0 - f_0)$, which definitely decreases to 0 as $n \to \infty$. Suppose that all hypotheses of the Rate of Convergence Theorem hold.

Now, note that for any $\eta \leq 2M^*_{n,L_d+1}$

$$\begin{aligned} ln\left(N\left(\eta, \mathcal{F}_{W_n}, \rho_n\right)\right) &\leq ln\left(N\left(\eta, \mathcal{F}_{W_n}, \|\cdot\|_\infty\right)\right) \\ &\leq W_n \cdot ln\left(\frac{4 \cdot e \cdot d \cdot M^*_{n,L_d+1} \cdot W_n \cdot \left(M^{(all)}_{n,L_d+1}\right)^{L_d+1}}{\eta \cdot \left(M^{(all)}_{n,L_d+1} - 1\right)}\right) \\ &\lesssim C^*_{n,d,L_d+1,W_n} + W_n\left(ln\left(\frac{4}{\eta}\right) + 1\right) \\ &\lesssim C^*_{n,d,L_d+1,W_n} \cdot \left(\frac{1+\eta}{16\eta}\right) \end{aligned}$$

as for n sufficiently large, $C^*_{n,d,L_d+1,W_n} > W_n$, and the difference keeps increasing as n getting larger.

Thus, for any $\delta \leq 1$, we have

$$\begin{aligned} \int_0^\delta \sqrt{ln\left(N\left(\eta, \mathcal{F}_{W_n}, \rho_n\right)\right)}d\eta &\lesssim \sqrt{C^*_{n,d,L_d+1,W_n}} \int_0^\delta \sqrt{\frac{1+\eta}{16\eta}}d\eta \\ &\lesssim \frac{\sqrt{C^*_{n,d,L_d+1,W_n}}}{4} \int_0^\delta \sqrt{1+\frac{1}{\eta}}d\eta \\ &\lesssim \frac{\sqrt{2 \cdot C^*_{n,d,L_d+1,W_n}}}{4} \int_0^\delta \sqrt{\frac{1}{\eta}}d\eta \\ &\lesssim \sqrt{C^*_{n,d,L_d+1,W_n} \cdot \delta} =: \phi_n(\delta) \end{aligned}$$

Clearly, $\phi_n(\delta)$ defined above makes $\delta \mapsto \phi_n(\delta)/\delta^\beta$ decreasing in $(0,1)$, for any $\frac{1}{2} < \beta < 2$. Let $r_n \lesssim \rho_n\left(\pi_{W_n} f_0 - f_0\right)^{-1}$. For each n under consideration, we would like to have

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) = \sqrt{C^*_{n,d,L_d+1,W_n}}\, r_n^{3/2} \leq \sqrt{n}$$

which will be satisfied if we have $r_n \leq \sqrt[3]{n/C^*_{n,d,L_d+1,W_n}}$.

From

$$\begin{aligned}
\mathbb{Q}_n(\hat{f}_n) &\leq \inf_{f \in \mathcal{F}_{W_n}} \mathbb{Q}_n(f) + \mathcal{O}_{\mathbb{P}}\left(\eta_n\right) \\
&\leq \mathbb{Q}_n\left(\pi_{W_n} f_0\right) + \mathcal{O}_{\mathbb{P}}\left(\eta_n\right) \\
&\leq \mathbb{Q}_n\left(\pi_{W_n} f_0\right) + \mathcal{O}_{\mathbb{P}}\left(r_n^{-2}\right)
\end{aligned}$$

we require $\eta_n \leq r_n^{-2}$.

Because no other requirements of the lower bounds of r_n^{-1} beside the three inequalities above, we can take

$$r_n^{-2} = \max\left\{\rho_n\left(\pi_{W_n} f_0 - f_0\right)^2, \left(\frac{C^*_{n,d,L_d+1,W_n}}{n}\right)^{2/3}\right\}$$

which justifies the upper bounding of η_n in the Rate of Convergence Theorem.

Hence, by the Convergence Rate of $\rho_n(\hat{f}_n - \pi_{W_n} f_0)$ Remark, we have $\rho_n\left(\hat{f}_n - \pi_{W_n} f_0\right)$ $= \mathcal{O}_{\mathbb{P}^*}\left(r_n^{-1}\right)$ the Triangle inequality, we have

$$\begin{aligned}
\rho_n\left(\hat{f}_n - f_0\right) &\leq \rho_n\left(\hat{f}_n - \pi_{W_n} f_0\right) + \rho_n\left(\pi_{W_n} f_0 - f_0\right) \\
&\leq \mathcal{O}_{\mathbb{P}^*}\left(r_n^{-1}\right) + \rho_n\left(\pi_{W_n} f_0 - f_0\right) \\
&\leq \mathcal{O}_{\mathbb{P}^*}\left(\max\left\{\rho_n\left(\pi_{W_n} f_0 - f_0\right), \left(\frac{C^*_{n,d,L_d+1,W_n}}{n}\right)^{1/3}\right\}\right).
\end{aligned}$$

□

Proof of the Theorem **??**. For any $f \in \mathcal{F}$, the first-order functional Taylor expansion of $\mathbb{Q}_n$ in

f is

$$\begin{aligned}\mathbb{Q}_n(f) &= \mathbb{Q}_n(f_0) + d\mathbb{Q}_n(f_0; f - f_0) + \mathcal{R}_1(f_0; f - f_0) \\ &= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right) + \frac{1}{n}\sum_{i=1}^{n}\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(f(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right) + \rho_n(f - f_0)^2 \end{aligned} \tag{A.1}$$

For $\delta_n = \sqrt{\eta_n} = o\left(r_n^{-1}\right)$ and $\iota(\boldsymbol{x}) \equiv 1$, define the local alternatives

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

$$\pi_{W_n}\tilde{f}_n := (1 - \delta_n)\,\hat{f}_n + \delta_n\,(\pi_{W_n}f_0 + \iota)$$

Then, by using (A.1), we have

$$\mathbb{Q}_n(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\left(\hat{f}_n(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right) + \rho_n(\hat{f}_n - f_0)^2$$

$$\mathbb{Q}_n(\pi_{W_n}\tilde{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\left(\pi_{W_n}\tilde{f}_n(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right)\right) + \rho_n(\pi_{W_n}\tilde{f}_n - f_0)^2$$

and subtracting these two equations gives

$$\mathbb{Q}_n(\hat{f}_n) = \mathbb{Q}_n(\pi_{W_n}\tilde{f}_n) + \frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\left(\pi_{W_n}\tilde{f}_n(\boldsymbol{x_i}) - \hat{f}_n(\boldsymbol{x_i})\right)\right) + \rho_n(\hat{f}_n - f_0)^2 - \rho_n(\pi_{W_n}\tilde{f}_n - f_0)^2$$

Next, by using the local alternative definition of $\pi_{W_n}\tilde{f}_n$, the pseudo-scalar product distributive-w.r.t-addition property, and the Cauchy-Schwartz inequality we have

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

Using this expression and the fact

$$\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{W_n}\tilde{f}_n(\boldsymbol{x_i})-\hat{f}_n(\boldsymbol{x_i})\right)=-\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\delta_n\left(\hat{f}_n(\boldsymbol{x_i})-\pi_{W_n}f_0(\boldsymbol{x_i})\right)-\delta_n\iota(\boldsymbol{x_i})\right)$$

we have

$$\begin{aligned}
-\mathcal{O}_{\mathbb{P}}\left(\delta_n^2\right) &\leq \inf_{f\in\mathcal{F}_{W_n}}\mathbb{Q}_n(f)-\mathbb{Q}_n(\hat{f}_n)\\
&\leq \mathbb{Q}_n\left(\pi_{W_n}\tilde{f}_n\right)-\mathbb{Q}_n(\hat{f}_n)\\
&\leq \rho_n\left(\pi_{W_n}\tilde{f}_n-f_0\right)^2-\rho_n\left(\hat{f}_n-f_0\right)^2-\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\pi_{W_n}\tilde{f}_n(\boldsymbol{x_i})-\hat{f}_n(\boldsymbol{x_i})\right)\\
&\leq \left(-2\delta_n+\delta_n^2\right)\rho_n\left(\hat{f}_n-f_0\right)^2+2\left(1-\delta_n\right)\delta_n\rho_n\left(\hat{f}_n-f_0\right)\rho_n\left(\pi_{W_n}f_0-f_0\right)\\
&\quad+2\left(1-\delta_n\right)\delta_n\left\langle\hat{f}_n-f_0,\iota\right\rangle_{\rho_n}+\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x_i})-\pi_{W_n}f_0(\boldsymbol{x_i})\right)-\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})\\
&\quad+2\delta_n^2\rho_n\left(\pi_{W_n}f_0-f_0\right)^2+2\delta_n^2\rho_n\left(\pi_{W_n}f_0-f_0\right)+\delta_n^2\\
&\leq \delta_n^2\rho_n\left(\hat{f}_n-f_0\right)^2+2\left(1-\delta_n\right)\delta_n\rho_n\left(\hat{f}_n-f_0\right)\rho_n\left(\pi_{W_n}f_0-f_0\right)\\
&\quad+2\left(1-\delta_n\right)\delta_n\left\langle\hat{f}_n-f_0,\iota\right\rangle_{\rho_n}\\
&\quad+\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x_i})-\pi_{W_n}f_0(\boldsymbol{x_i})\right)-\frac{2}{n}\delta_n\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})+\mathcal{O}_{\mathbb{P}}(\delta_n^2)
\end{aligned}$$

as we know that $2\delta_n^2\rho_n\left(\pi_{W_n}f_0-f_0\right)^2+2\delta_n^2\rho_n\left(\pi_{W_n}f_0-f_0\right)+\delta_n^2=\mathcal{O}_{\mathbb{P}}(\delta_n^2)$ and also $-2\delta_n+\delta_n^2\leq\delta_n^2$.

We can rewrite the last inequality as

$$\begin{aligned}
-\mathcal{O}_{\mathbb{P}}\left(\delta_n\right) &\leq \delta_n\rho_n\left(\hat{f}_n-f_0\right)^2+2\left(1-\delta_n\right)\rho_n\left(\hat{f}_n-f_0\right)\rho_n\left(\pi_{W_n}f_0-f_0\right)\\
&\quad+2\left(1-\delta_n\right)\left\langle\hat{f}_n-f_0,\iota\right\rangle_{\rho_n}+\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x_i})-\pi_{W_n}f_0(\boldsymbol{x_i})\right)-\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})+\mathcal{O}_{\mathbb{P}}(\delta_n)
\end{aligned}\tag{A.2}$$

Our goal now is to show $n\to\infty$ are $\left\langle\hat{f}_n-f_0,\iota\right\rangle_{\rho_n}$ can be expressed as $\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})+o_{\mathbb{P}^*}(n^{-1/2})$, and the other terms are $o_{\mathbb{P}^*}(n^{-1/2})$.

Note that the second condition of this theorem makes the consistency assumption in the Consistency and Rate of Convergence theorems are satisfied. Also, big-O and small-O in $\mathbb{P}$ definitely imply big-O and small-O in $\mathbb{P}^*$, respectively. Hence, by using the Rate of Convergence theorem,

$$\rho_n\left(\hat{f}_n - f_0\right)\rho_n\left(\pi_{W_n}f_0 - f_0\right) = \mathcal{O}_{\mathbb{P}^*}\left(\rho_n\left(\pi_{W_n}f_0 - f_0\right)^2, \rho_n\left(\pi_{W_n}f_0 - f_0\right)\left(\frac{C^*_{n,d,L_n+1,W_n}}{n}\right)^{1/3}\right)$$

The third condition of this theorem implies

$$\begin{aligned}
&\rho_n\left(\pi_{W_n}f_0 - f_0\right)^2 = o\left(o\left(n^{-1/4}\right)^2\right) = o\left(n^{-1/2}\right) \\
&\rho_n\left(\pi_{W_n}f_0 - f_0\right)\left(\frac{C^*_{n,d,L_n+1,W_n}}{n}\right)^{1/3} \\
&= o\left(n^{-1/6}\left(C^*_{n,d,L_n+1,W_n}\right)^{-1/3}\left(C^*_{n,d,L_n+1,W_n}\right)^{1/3}n^{-1/3}\right) = o\left(n^{-1/2}\right)
\end{aligned}$$

which give

$$2\left(1-\delta_n\right)\rho_n\left(\hat{f}_n - f_0\right)\rho_n\left(\pi_{W_n}f_0 - f_0\right) = o_{\mathbb{P}^*}\left(n^{-1/2}\right)$$

Now, we are going to use Remark 4.4.1 and the definition of big-O in probability. By noting that $A_n = C^*_{n,d,L_n+1,W_n}$, $M_n = M^*_{n,L_n+1}$, we have

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\left(\hat{f}_n(\boldsymbol{x_i}) - \pi_{W_n}f_0(\boldsymbol{x_i})\right) &\leq \frac{1}{\sqrt{n}}\sup_{f\in\mathcal{F}_{W_n}}\frac{4}{\sqrt{n}}\left|\sum_{i=1}^{n}\epsilon_i f(\boldsymbol{x_i})\right| \\
&= \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\frac{\left(C^*_{n,d,L_n+1,W_n}\right)^{2/3}\left(M^*_{n,L_n+1}\right)^{1/3}}{n^{1/6}}\right)
\end{aligned}$$

and to have the term in RHS as $o\left(n^{-1/2}\right)$, we require

$$\frac{\left(C^*_{n,d,L_n+1,W_n}\right)^{2/3}\left(M^*_{n,L_n+1}\right)^{1/3}}{n^{1/6}} = \left(\frac{C^*_{n,d,L_n+1,W_n}\sqrt{M^*_{n,L_n+1}}}{n^{1/4}}\right)^{2/3} = o(1)$$

which is satisfied by the second condition.

Hence, (A.2) can be rewritten as

$$
\begin{aligned}
-\mathcal{O}_{\mathbb{P}}(r_n^{-1}) \leq\ & o\left(r_n^{-1}n^{-1/2}\right) + o_{\mathbb{P}^*}\left(n^{-1/2}\right) \\
& + 2\left(1-\delta_n\right)\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n} + o_{\mathbb{P}}(n^{-1/2}) \\
& - \frac{2}{n}\sum_{i=1}^{n}\epsilon_i + \mathcal{O}_{\mathbb{P}}(r_n^{-1})
\end{aligned}
$$

with both (3.2) and the fact that big-O and small-O imply big-O and small-O in probability, respectively, imply

$$
\begin{aligned}
-\left(1-\delta_n\right)\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n} + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i}) &\leq o\left(r_n^{-1}n^{-1/2}\right) + o_{\mathbb{P}^*}\left(n^{-1/2}\right) + o_{\mathbb{P}}(n^{-1/2}) + \mathcal{O}_{\mathbb{P}}(r_n^{-1}) \\
&\leq o_{\mathbb{P}^*}\left(n^{-1/2}\right)
\end{aligned}
$$

If both sides in the last inequality of (A.2) are multiplied by -1 and move $\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n}$ and $\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})$ terms to the LHS and the other terms to the RHS, then we have

$$
\left(1-\delta_n\right)\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i}) \leq o_{\mathbb{P}^*}\left(n^{-1/2}\right)
$$

Therefore, by the Triangle Inequality,

$$
\begin{aligned}
\left|\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})\right| &\leq \left|\left(1-\delta_n\right)\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n} - \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\iota(\boldsymbol{x_i})\right| + \delta_n\left|\left\langle \hat{f}_n - f_0, \iota \right\rangle_{\rho_n}\right| \\
&\leq o_{\mathbb{P}^*}\left(n^{-1/2}\right) + \delta_n\rho_n\left(\hat{f}_n - f_0\right) \\
&= o_{\mathbb{P}^*}\left(n^{-1/2}\right)
\end{aligned}
$$

which gives

$$
\sqrt{n}\left\langle \hat{f}_n - f_0, \iota \right\rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{f}_n(\boldsymbol{x_i}) - f_0(\boldsymbol{x_i})\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i + o_{\mathbb{P}^*}(1)
$$

and thus the conclusion follows from the Central Limit Theorem. □

Proof of Lemma 4.4.1. Suppose that a node of θ, call it β_1, in either the input layer or a

hidden layer, is connected to a node β_2 contained in a layer that is not adjacent to the β_1 layer. Then, we add new nodes β_1'', β_2'',....., β_v'' located in each of the v adjacent layers between the β_1 layer and the β_2 layer that connect β_1 and β_2 adjacently. If initially β_2 has linear term $\beta_{1,output} w_{\beta_1,\beta_2}$, where $\beta_{1,output}$ is the output of β_1, as a part of its linear aggregation input, then $\beta_{1,output} w_{\beta_1,\beta_2} + B$ is both inputs and outputs of node β_1'', β_2'',....., β_v'' with B is taken sufficiently large to ensure $ReLU(\beta_{1,output} w_{\beta_1,\beta_2} + B) = \beta_{1,output} w_{\beta_1,\beta_2} + B$. The node β_2 now has the new $-B$ term adding its old constant input term to ensure that the input from β_v'' remains $\beta_{1,output} w_{\beta_1,\beta_2}$. This shows that any non-adjacent node connections in θ can be transformed into the equivalent adjacent connections.

After transforming that all node connections from the input layer to the last hidden layer into adjacent connecctions, we can add missing previous layer connections by taking zero weights for the input from the related nodes. Then, we can look for the hidden layer with most nodes, take H_n as the number of the nodes in that layer, and add the nodes to other hidden layers to ensure that every hidden layer has H_n nodes. These new nodes have zero inputs and outputs, as all incoming connections to each of these nodes have zero weights, to make sure that the output of the other nodes stay the same. We are then guaranteed to have a multi-layer perceptron with full previous-layer connections and same number of nodes per hidden layer H_n called θ'.

Now, we are going to derive the upper bound condition for H_n in θ'. Suppose that H_1', H_2',...., H_{L_n}' are number of nodes in 1^{st}, 2^{nd},....., and L_n^{th} hidden layers of θ, respectively. Suppose also that θ has been transformed into θ' by constructing adjacent connections subsequently from the input layer to the L_n^{th} hidden layer. As the new nodes in the u^{th} hidden layer are constructed to bridge the non-adjacent connections from all layers before

that layer, the 1^{st}, 2^{nd},....., and L_n^{th} hidden layers in θ' have at most $d + H'_1$, $d + H'_1 + H'_2$,....,

$d + \sum_{u=1}^{L_n} H'_u$ nodes, respectively. We know that $H'_u \leq N_n, \forall 1 \leq u \leq L_n$. Thus we can infer

$H_n \leq d + \sum_{u=1}^{L_n} H'_u \leq d + \left(\max_{1 \leq u \leq L_n} H'_u \right) L_n \leq d + N_n L_n$. □

References

AKPINAR, N.-J., B. KRATZWALD, AND S. FEUERRIEGEL (2019): "Sample Complexity Bounds for Recurrent Neural Networks with Application to Combinatorial Graph Problems," *arXiv preprint arXiv:1901.10289*.

ANTHONY, M., AND P. L. BARTLETT (2009): *Neural network learning: Theoretical foundations*. cambridge university press.

BARTLETT, P. L., N. HARVEY, C. LIAW, AND A. MEHRABIAN (2019): "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks.," *Journal of Machine Learning Research*, 20(63), 1–17.

CHEN, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

CHEN, X., AND X. SHEN (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, pp. 289–314.

FARRELL, M. H., T. LIANG, AND S. MISRA (2019): "Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands," *arXiv preprint arXiv:1809.09953*.

GOODFELLOW, I., Y. BENGIO, AND A. COURVILLE (2016): *Deep learning*. MIT press.

GRENANDER, U. (1981): "Abstract inference," Discussion paper.

HOREL, E., AND K. GIESECKE (2019): "Towards explainable ai: Significance tests for neural networks," *arXiv preprint arXiv:1902.06021*.

LIU, M., J. SHI, Z. LI, C. LI, J. ZHU, AND S. LIU (2016): "Towards better analysis of deep convolutional neural networks," *IEEE transactions on visualization and computer graphics*, 23(1), 91–100.

SHEN, X., ET AL. (1997): "On methods of sieves and penalization," *The Annals of Statistics*, 25(6), 2555–2591.

SHEN, X., C. JIANG, L. SAKHANENKO, AND Q. LU (2019): "Asymptotic Properties of Neural Network Sieve Estimators," *arXiv preprint arXiv:1906.00875*.

SUN, S., W. CHEN, L. WANG, X. LIU, AND T.-Y. LIU (2016): "On the depth of deep neural networks: A theoretical view," in *Thirtieth AAAI Conference on Artificial Intelligence*.

VAART, A. W., AND J. A. WELLNER (1996): *Weak convergence and empirical processes: with applications to statistics*. Springer.

YAROTSKY, D. (2017): "Error bounds for approximations with deep ReLU networks," *Neural Networks*, 94, 103–114.

——— (2018): "Optimal approximation of continuous functions by very deep relu networks," *arXiv preprint arXiv:1802.03620*.

ZHOU, P., AND J. FENG (2018): "Understanding generalization and optimization performance of deep CNNs," *arXiv preprint arXiv:1805.10767*.