Deep Network Approximation for Smooth Functions

Jianfeng Lu * Zuowei Shen † Haizhao Yang ‡ Shijun Zhang §

July 21, 2021

Abstract

This paper establishes optimal approximation error characterization of deep ReLU networks for smooth functions in terms of both width and depth simultaneously. To that end, we first prove that multivariate polynomials can be approximated by deep ReLU networks of width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ with an approximation error $\mathcal{O}(N^{-L})$. Through local Taylor expansions and their deep ReLU network approximations, we show that deep ReLU networks of width $\mathcal{O}(N\ln N)$ and depth $\mathcal{O}(L\ln L)$ can approximate $f \in C^s([0,1]^d)$ with a nearly optimal approximation rate $\mathcal{O}(\|f\|_{C^s([0,1]^d)}N^{-2s/d}L^{-2s/d})$. Our estimate is non-asymptotic in the sense that it is valid for arbitrary width and depth specified by $N \in \mathbb{N}^+$ and $L \in \mathbb{N}^+$, respectively.

Key words. ReLU network, Smooth Function, Polynomial Approximation, Function Composition.

1 Introduction

Deep neural networks have made significant impacts in many fields of computer science and engineering especially for large-scale and high-dimensional learning problems. Well-designed neural network architectures, efficient training algorithms, and high-performance computing technologies have made neural-network-based methods very successful in tremendous real applications. Especially in supervised learning, e.g., image classification and objective detection, the great advantages of neural-network-based methods have been demonstrated over traditional learning methods. Mathematically speaking, supervised learning is essentially a regression problem where the problem of function approximation plays a fundamental role. Understanding the approximation capacity of deep neural networks has become a key question for revealing the power of deep learning. A large number of experiments in real applications have shown the large capacity of deep network approximation from many empirical points of view, motivating

*Department of Mathematics, Department of Physics, and Department of Chemistry, Duke University (jianfeng@math.duke.edu).

†Department of Mathematics, National University of Singapore (matzuows@nus.edu.sg).

‡Department of Mathematics, Purdue University (haizhao@purdue.edu).

§Department of Mathematics, National University of Singapore (zhangshijun@u.nus.edu).

much effort in establishing the theoretical foundation of deep network approximation. One of the fundamental problems is the characterization of the optimal approximation rate of deep neural networks of arbitrary depth and width.

Previously, the quantitative characterization of the approximation power of deep feed-forward neural networks (FNNs) with ReLU activation functions is provided in [27]. For ReLU FNNs with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$, the deep network approximation of $f \in C([0,1])^d$ admits an approximation rate $5\omega_f(8\sqrt{d}N^{-2/d}L^{-2/d})$ in the L^p-norm for $p \in [1,\infty)$, where $\omega_f(\cdot)$ is the modulus of continuity of f. In particular, for the class of Lipschitz continuous functions, the approximation rate is nearly optimal.① The next question is whether the smoothness of functions can improve the approximation rate. In this paper, we investigate the deep network approximation of smaller function space, such as the smooth function space $C^s([0,1]^d)$. Instead of discussing the approximation rate in the L^p-norm for $p \in [1,\infty)$ as in [27], we measure the approximation rate here in the L^∞-norm. As we are only interested in functions in $C^s([0,1]^d)$, the approximation rates in the L^∞-norm implies the ones in the L^p-norm for $p \in [1,\infty)$. To be precise, the main theorem of the present paper, Theorem 1.1 below, shows that ReLU FNNs with width $\mathcal{O}(N \ln N)$ and depth $\mathcal{O}(L \ln L)$ can approximate $f \in C^s([0,1]^d)$ with a nearly optimal approximation rate $\mathcal{O}(\|f\|_{C^s([0,1]^d)} N^{-2s/d} L^{-2s/d})$, where the norm $\|\cdot\|_{C^s([0,1]^d)}$ is defined as

$$\|f\|_{C^s([0,1]^d)} \coloneqq \max\left\{\|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} : \|\boldsymbol{\alpha}\|_1 \le s,\ \boldsymbol{\alpha} \in \mathbb{N}^d\right\}, \quad \text{for any } f \in C^s([0,1]^d).$$

Theorem 1.1 (Main Theorem)**.** *Given a function $f \in C^s([0,1]^d)$ with $s \in \mathbb{N}^+$, for any $N, L \in \mathbb{N}^+$, there exists a ReLU FNN ϕ with width $C_1 d(N+2)\log_2(4N)$ and depth $C_2(L+2)\log_2(2L) + 2d$ such that*

$$\|f - \phi\|_{L^\infty([0,1]^d)} \le C_3 \|f\|_{C^s([0,1]^d)} N^{-2s/d} L^{-2s/d},$$

where $C_1 = 22s^{d+1}3^d$, $C_2 = 18s^2$, and $C_3 = 85(s+1)^d 8^s$.

As we can see from Theorem 1.1, the smoothness improves the approximation efficiency. When functions are sufficiently smooth (e.g., $s \ge d$), since $\mathcal{O}(N^{-2s/d}L^{-2s/d}) \le \mathcal{O}(N^{-2}L^{-2})$, the approximation rate is independent of d. This means that the curse of dimensionality can be reduced for sufficiently smooth functions. The proof of Theorem 1.1 will be presented in Section 2.2 and its tightness will be discussed in Section 2.3. In fact, the logarithm terms in width and depth in Theorem 1.1 can be further reduced if the approximation rate is weaken. Note that

$$\mathcal{O}(N \ln N) = \mathcal{O}(\widetilde{N}) \Longleftrightarrow \mathcal{O}(N) = \mathcal{O}(\widetilde{N}/\ln \widetilde{N}).$$

Applying Theorem 1.1 with $\widetilde{N} = \mathcal{O}(N \log N)$ and $\widetilde{L} = \mathcal{O}(L \log L)$ and the fact that

$$(N/\ln N)^{-2s/d}(L/\ln L)^{-2s/d} \le \mathcal{O}\left(N^{-2(s-\rho)/d} L^{-2(s-\rho)/d}\right)$$

for any $\rho \in (0, s)$, we have the following corollary.

① "nearly optimal" up to a logarithm factor.

Corollary 1.2. *Given a function $f \in C^s([0,1]^d)$ with $s \in \mathbb{N}^+$, for any $N, L \in \mathbb{N}^+$ and $\rho \in (0,s)$, there exist $C_1(s,d)$, $C_2(s,d)$, $C_3(s,d,\rho)$,*[②] *and a ReLU FNN ϕ with width $C_1 N$ and depth $C_2 L$ such that*

$$\|f-\phi\|_{L^\infty([0,1]^d)} \le C_3 \|f\|_{C^s([0,1]^d)} N^{-2(s-\rho)/d} L^{-2(s-\rho)/d}.$$

Theorem 1.1 and Corollary 1.2 characterize the approximation rate in terms of total number of neurons (with an arbitrary distribution in width and depth) and smoothness order of the function to be approximated. In other words, for arbitrary width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$, Theorem 1.1 and Corollary 1.2 provide nearly optimal approximation rates $\mathcal{O}((\frac{N}{\ln N})^{-2s/d}(\frac{L}{\ln L})^{-2s/d})$ and $\mathcal{O}(N^{-2(s-\rho)/d}L^{-2(s-\rho)/d})$ for $\rho \in (0,s)$ (see Theorem 2.3 for the optimality). The only result in this direction we are aware of in literature is Theorem 4.1 of [32]. It shows that ReLU networks with width $2d+10$ and depth L achieve an nearly optimal rate $\mathcal{O}((\frac{L}{\ln L})^{-2s/d})$ for sufficiently large L when approximating functions in the unit ball of $C^s([0,1]^d)$. This result can be considered as a special case of Theorem 1.1 by setting $N = \mathcal{O}(1)$ and L sufficiently large.

The results obtained in [32] and this paper are for $C^s([0,1]^d)$ functions. For Lipschitz continuous functions, it is proved in [31] that the optimal rate for ReLU FNNs with width $2d+10$ and depth $\mathcal{O}(L)$ to approximate Lipschitz continuous functions on $[0,1]^d$ in the L^∞-norm is $\mathcal{O}(L^{-2/d})$. For the purpose of deep network approximation with arbitrary width and depth, the last three authors demonstrated in [27] that the optimal approximation rate for ReLU FNNs with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ to approximate Lipschitz continuous functions on $[0,1]^d$ in the L^p-norm for $p \in [1,\infty)$ is $\mathcal{O}(N^{-2/d}L^{-2/d})$. We remark that, combined with the proof technique of Theorem 2.1 in this work, the norm characterizing error of [27] can be improved to L^∞-norm; it will also remove the log factors in the case of C^1 functions in our results here.

The expressiveness of deep neural networks has been studied extensively from many perspectives, e.g., in terms of combinatorics [22], topology [5], Vapnik-Chervonenkis (VC) dimension [4, 25, 13], fat-shattering dimension [16, 1], information theory [24], classical approximation theory [9, 15, 3, 31, 30, 6, 33, 8, 11, 12, 29, 23, 7, 2, 17, 20], etc. In the early works of approximation theory for neural networks, the universal approximation theorem [9, 14, 15] without approximation rates showed that, given any $\varepsilon > 0$, there exists a sufficiently large neural network approximating a target function in a certain function space within the ε-accuracy. For one-hidden-layer neural networks and sufficiently smooth functions, Barron [3] showed an asymptotic approximation rate $\mathcal{O}(\frac{1}{\sqrt{N}})$ in the L^2-norm, leveraging an idea that is similar to Monte Carlo sampling for high-dimensional integrals. All these related works are summarized in Table 1.

In literature, the approximation rate is often described in terms of the number of parameters of neural networks. Most existing works aims at studying the connection between the number of parameters (weights) and the approximation rates, e.g., smooth functions [19, 18, 30, 10], piecewise smooth functions [24], band-limited functions [21], continuous functions [31]. The key difference between these works and the results of this paper is the variable of characterizing approximation rates. To be precise, results in the papers mentioned above characterize the approximation rates in terms of the number of parameters. To optimize the number of parameters for a given error, these

[②] C_i, for $i = 1,2,3$, can be specified explicitly and we leave the detailed discussion to reader.

Table 1: A summary of existing approximation rates of ReLU FNNs for Lipschitz continuous functions and smooth functions.

paper	function class	width	depth	accuracy	$L^p([0,1]^d)$-norm	tightness	valid for
[30]	polynomial	$\mathcal{O}(1)$	$\mathcal{O}(L)$	$\mathcal{O}(2^{-L})$	$p=\infty$		any $L\in\mathbb{N}^+$
this paper	polynomial	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}(N^{-L})$	$p=\infty$		any $N,L\in\mathbb{N}^+$
[26]	$\mathrm{Lip}([0,1]^d)$	$\mathcal{O}(N)$	3	$\mathcal{O}(N^{-2/d})$	$p\in[1,\infty)$	nearly tight in N	any $N\in\mathbb{N}^+$
[31]	$\mathrm{Lip}([0,1]^d)$	$2d+10$	$\mathcal{O}(L)$	$\mathcal{O}(L^{-2/d})$	$p=\infty$	nearly tight in L	large $L\in\mathbb{N}^+$
[27]	$\mathrm{Lip}([0,1]^d)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}(N^{-2/d}L^{-2/d})$	$p=[1,\infty]$	nearly tight in N and L	any $N,L\in\mathbb{N}^+$
[28]	$\mathrm{Lip}([0,1]^d)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}\big((N^2L^2\ln N)^{-1/d}\big)$	$p=[1,\infty]$	tight in N and L	any $N,L\in\mathbb{N}^+$
[32]	$C^s([0,1]^d)$	$2d+10$	$\mathcal{O}(L)$	$\mathcal{O}((L/\ln L)^{-2s/d})$	$p=\infty$	neatly tight in L	large $L\in\mathbb{N}^+$
this paper	$C^s([0,1]^d)$	$\mathcal{O}(N\ln N)$	$\mathcal{O}(L\ln L)$	$\mathcal{O}(N^{-2s/d}L^{-2s/d})$	$p=\infty$	nearly tight in N and L	any $N,L\in\mathbb{N}^+$
this paper	$C^s([0,1]^d)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{O}((N/\ln N)^{-2s/d}(L/\ln L)^{-2s/d})$	$p=\infty$	nearly tight in N and L	any $N,L\in\mathbb{N}^+$

papers construct very special network architectures, such as very deep but very narrow networks, complicated networks generated by compositing shallow-wide sub-networks and deep-narrow sub-networks, etc, while our approximation rates in Theorem 1.1 and Corollary 1.2 are valid for arbitrary width and depth up to an absolute constant. This gives us much more freedom to design neural networks for a good approximation. In other words, it means the shape of our network architectures is a rectangle with free choice of width and length, which is of more practical interest in real applications and requires innovative constructive proofs.

The approaches characterizing approximation rates in terms of the number of parameters are unable to characterize the approximation rate of FNNs in terms of width and depth simultaneously. Theorem 1.1 and the results in [26, 27] give an explicit characterization of the approximation rate of FNNs in terms of width and depth, in the non-asymptotic regime. Furthermore, applying Theorem 1.1, we have the following corollary.

Corollary 1.3. *Given any $\varepsilon>0$ and a function f in the unit ball of $C^s([0,1]^d)$ with $s\in\mathbb{N}^+$, there exists a ReLU FNN ϕ with $\mathcal{O}(\varepsilon^{-d/(2s)}\ln\frac{1}{\varepsilon})$ parameters such that*

$$\|f-\phi\|_{L^\infty([0,1]^d)}\le\varepsilon.$$

This corollary is followed by setting $N=\mathcal{O}(1)$ and $\varepsilon=\mathcal{O}(L^{-2s/d})$ in Theorem 1.1, which characterizes the approximation rate in terms of the number of parameters. It is essentially equivalent to Theorem 4.1 of [32] by setting $\varepsilon=\mathcal{O}(W^{-2s/d}\ln^{2s/d}W)$, which presents that ReLU networks with W parameters achieve an approximation rate $\mathcal{O}(W^{-2s/d}\ln^{2s/d}W)$ when approximating functions in the unit ball of $C^s([0,1]^d)$. As shown here, we can straightforwardly deduce Corollary 1.3 and Theorem 4.1 of [32] from Theorem 1.1. However, Theorem 1.1 can not be derived from any existing result that characterizes approximation rates in terms of the number of parameters. Therefore, Theorem 1.1 goes beyond existing results on the approximation of deep neural networks.

Finally, in a completely different approach, the authors of [17] establish the approximation capabilities of deep learning models in the form of dynamical systems. This approach focuses on the continuous-time idealization. The key advantage of this viewpoint is that a variety of tools from the continuous-time analysis can be used to analyze the approximation of deep neural networks. Furthermore, approximation results in continuous-time have immediate consequences for its discrete counterpart, which can be viewed as a deep, residual, fully-connected neural network, by a forward Euler discretization in time.

The rest of the present paper is organized as follows. In Section 2, we prove Theorem 1.1 by combining two theorems (Theorems 2.1 and 2.2) that will be proved later. We will also discuss the optimality of Theorem 1.1 in Section 2. Next, Theorem 2.1 will be proved in Section 3 while Theorem 2.2 will be shown in Section 4. Several lemmas supporting Theorem 2.2 will be presented in Section 5. Finally, Section 6 concludes this paper with a short discussion.

2 Approximation of smooth functions

In this section, we will prove the quantitative approximation rate in Theorem 1.1 by construction and discuss its tightness. Notations throughout the proof will be summarized in Section 2.1. The proof of Theorem 1.1 is mainly based on Theorem 2.1 and 2.2, which will be proved in Section 3 and 4, respectively. To show the tightness of Theorem 1.1, we will introduce the VC-dimension in Section 2.3.

2.1 Notations

Now let us summarize the main notations of the present paper as follows.

- Let 1_S be the characteristic function on a set S, i.e., 1_S equals to 1 on S and 0 outside of S.
- Let $\mathcal{B}(\boldsymbol{x}, r) \subseteq \mathbb{R}^d$ be the closed ball with a center $\boldsymbol{x} \subseteq \mathbb{R}^d$ and a radius r.
- Similar to "min" and "max", let $\mathrm{mid}(x_1, x_2, x_3)$ be the middle value of three inputs x_1, x_2, and x_3[③]. For example, $\mathrm{mid}(2,1,3) = 2$ and $\mathrm{mid}(3,2,3) = 3$.
- The set difference of two sets A and B is denoted by $A\backslash B \coloneqq \{x : x \in A,\ x \notin B\}$.
- For any $x \in \mathbb{R}$, let $\lfloor x \rfloor \coloneqq \max\{n : n \le x,\ n \in \mathbb{Z}\}$ and $\lceil x \rceil \coloneqq \min\{n : n \ge x,\ n \in \mathbb{Z}\}$.
- Assume $\boldsymbol{n} \in \mathbb{N}^n$, then $f(\boldsymbol{n}) = \mathcal{O}(g(\boldsymbol{n}))$ means that there exists positive C independent of $\boldsymbol{n}$, f, and g such that $f(\boldsymbol{n}) \le C g(\boldsymbol{n})$ when all entries of $\boldsymbol{n}$ go to $+\infty$.
- The modulus of continuity of a continuous function $f \in C([0,1]^d)$ is defined as

$$\omega_f(r) \coloneqq \sup\left\{|f(\boldsymbol{x}) - f(\boldsymbol{y})| : \|\boldsymbol{x} - \boldsymbol{y}\|_2 \le r,\ \boldsymbol{x}, \boldsymbol{y} \in [0,1]^d\right\}, \quad \text{for any } r \ge 0.$$

- A d-dimensional multi-index is a d-tuple $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_d]^T \in \mathbb{N}^d$. Several related notations are listed below.
  - $\|\boldsymbol{\alpha}\|_1 = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_d|$;
  - $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$, where $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T$;
  - $\boldsymbol{\alpha}! = \alpha_1! \alpha_2! \cdots \alpha_d!$;

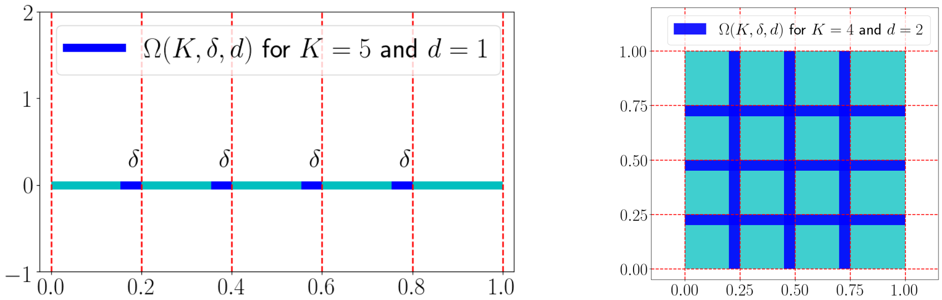
[③] "mid" can be defined via $\mathrm{mid}(x_1, x_2, x_3) = x_1 + x_2 + x_3 - \max(x_1, x_2, x_3) - \min(x_1, x_2, x_3)$, which can be implemented by a ReLU FNN.

– $\partial^{\boldsymbol{\alpha}} = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}$.

- Given $K \in N^+$ and $\delta > 0$ with $\delta < \frac{1}{K}$, define a trifling region $\Omega(K, \delta, d)$ of $[0,1]^d$ as [4]

$$\Omega(K, \delta, d) \coloneqq \bigcup_{i=1}^{d} \Big\{ \boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T : x_i \in \cup_{k=1}^{K-1} \big(\tfrac{k}{K} - \delta, \tfrac{k}{K}\big) \Big\}. \tag{2.1}$$

  In particular, $\Omega(K, \delta, d) = \varnothing$ if $K = 1$. See Figure 1 for two examples of trifling regions.



(a) (b)

Figure 1: Two examples of trifling regions. (a) $K = 5, d = 1$. (b) $K = 4, d = 2$.

- We will use NN as a ReLU neural network for short and use Python-type notations to specify a class of NNs, e.g., $\mathrm{NN}(\mathrm{c}_1;\ \mathrm{c}_2;\ \cdots;\ \mathrm{c}_m)$ is a set of ReLU FNNs satisfying m conditions given by $\{\mathrm{c}_i\}_{1\le i\le m}$, each of which may specify the number of inputs (#input), the total number of nodes in all hidden layers (#node), the number of hidden layers (depth), the number of total parameters (#parameter), and the width in each hidden layer (widthvec), the maximum width of all hidden layers (width), etc. For example, if $\phi \in \mathrm{NN}(\#\mathrm{input} = 2; \mathrm{widthvec} = [100, 100])$, then ϕ satisfies

  – ϕ maps from $\mathbb{R}^2$ to $\mathbb{R}$.

  – ϕ has two hidden layers and the number of nodes in each hidden layer is 100.

- The expression "a network with width N and depth L" means

  – The maximum width of all hidden layers is no more than N.

  – The number of hidden layers is no more than L.

- For $x \in [0, 1)$, suppose its binary representation is $x = \sum_{\ell=1}^{\infty} x_\ell 2^{-\ell}$ with $x_\ell \in \{0, 1\}$, we introduce a special notation $\mathrm{Bin}\, 0.x_1 x_2 \cdots x_L$ to denote the L-term binary representation of x, i.e., $\sum_{\ell=1}^{L} x_\ell 2^{-\ell}$.

[4] The trifling region here is similar to the "don't care" region in our previous paper [27].

2.2 Proof of Theorem 1.1

The introduction of the trifling region $\Omega(K,\delta,d)$ is due to the fact that ReLU FNNs cannot approximate a step function uniformly well (as ReLU activation function is continuous), which is also the reason for the main difficulty of obtaining approximation rates in the $L^\infty([0,1]^d)$-norm in our previous papers [26, 27]. The trifling region is a key technique to simplify the proofs of theories in [26, 27] as well as the proof of Theorem 1.1. First, we present Theorem 2.1 showing that, as long as good uniform approximation by a ReLU FNN can be obtained outside the trifling region, the uniform approximation error can also be well controlled inside the trifling region when the network size is increased. Second, as a simplified version of Theorem 1.1 ignoring the approximation error in the trifling region $\Omega(K,\delta,d)$, Theorem 2.2 shows the existence of a ReLU FNN approximating a target smooth function uniformly well outside the trifling region. Finally, Theorem 2.1 and 2.2 immediately lead to Theorem 1.1. Theorem 2.2 can be applied to improve the theories in [26, 27] to obtain approximation rates in the $L^\infty([0,1]^d)$-norm.

Theorem 2.1. *Given $\varepsilon>0$, $N,L,K\in\mathbb{N}^+$, and $\delta\in(0\frac{1}{3K}]$, assume $f\in C([0,1]^d)$ and $\widetilde{\phi}$ is a ReLU FNN with width N and depth L. If*

$$|f(\boldsymbol{x})-\widetilde{\phi}(\boldsymbol{x})|\le\varepsilon,\quad \text{for any } \boldsymbol{x}\in[0,1]^d\backslash\Omega(K,\delta,d),$$

then there exists a new ReLU FNN ϕ with width $3^d(N+4)$ and depth $L+2d$ such that

$$|f(\boldsymbol{x})-\phi(\boldsymbol{x})|\le\varepsilon+d\cdot\omega_f(\delta),\quad \text{for any } \boldsymbol{x}\in[0,1]^d.$$

Theorem 2.2. *Assume that $f\in C^s([0,1]^d)$ satisfies $\|\partial^{\boldsymbol{\alpha}}f\|_{L^\infty([0,1]^d)}\le 1$ for any $\boldsymbol{\alpha}\in\mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1\le s$. For any $N,L\in\mathbb{N}^+$, there exists a ReLU FNN ϕ with width $21s^{d+1}d(N+2)\log_2(4N)$ and depth $18s^2(L+2)\log_2(2L)$ such that*

$$\|f-\phi\|_{L^\infty([0,1]^d\backslash\Omega(K,\delta,d))}\le 84(s+1)^d8^sN^{-2s/d}L^{-2s/d},$$

where $K=\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor$ and $0<\delta\le\frac{1}{3K}$.

We first prove Theorem 1.1 assuming Theorem 2.1 and 2.2 are true. The proofs of Theorem 2.1 and 2.2 can be found in Section 3 and 4, respectively.

Proof of Theorem 1.1. Define $\bar{f}=\frac{f}{\|f\|_{C^s([0,1]^d)}}$, set $K=\lfloor N^{-2/d}\rfloor\lfloor L^{-1/d}\rfloor^2$, and choose $\delta\in(0,\frac{1}{K})$ such that $\omega_f(\delta)\le N^{-2s/d}L^{-2s/d}$. By Theorem 2.2, there exists a ReLU FNN $\widetilde{\phi}$ with width $21s^{d+1}d(N+2)\log_2(4N)$ and depth $18s^2(L+2)\log_s(2L)$ such that

$$\|\bar{f}-\widetilde{\phi}\|_{L^\infty([0,1]^d\backslash\Omega(K,\delta,d))}\le 84(s+1)^d8^sN^{-2s/d}L^{-2s/d}.$$

By Theorem 2.1, there exists a ReLU FNN $\bar{\phi}$ with width $3^d\big(21s^{d+1}d(N+2)\log_2(4N)+3\big)\le 22s^{d+1}3^dd(N+2)\log_2(4N)$ and depth $18s^2(L+2)\log_s(2L)+2d$ such that

$$\|\bar{f}-\bar{\phi}\|_{L^\infty([0,1]^d)}\le 84(s+1)^d8^sN^{-2s/d}L^{-2s/d}+d\cdot\omega_f(\delta)\le 85(s+1)^d8^sN^{-2s/d}L^{-2s/d}.$$

Finally, set $\phi=\|f\|_{C^s([0,1]^d)}\cdot\bar{\phi}$, then

$$\|f-\phi\|_{L^\infty([0,1]^d)}=\|f\|_{C^s([0,1]^d)}\|\bar{f}-\bar{\phi}\|_{L^\infty([0,1]^d)}\le 85(s+1)^d8^s\|f\|_{C^s([0,1]^d)}N^{-2s/d}L^{-2s/d},$$

which finishes the proof. $\square$

2.3 Optimality of Theorem 1.1

In this section, we will show that the approximation rate in Theorem 1.1 is asymptotically nearly tight. In particular, the approximation rate $\mathcal{O}(N^{-(2s/d+\rho)}L^{-(2s/d+\rho)})$ for any $\rho > 0$ is not attainable, if we use ReLU FNNs with width $\mathcal{O}(N\ln N)$ and depth $\mathcal{O}(L\ln L)$ to approximate functions in $\mathscr{F}_{s,d}$, where $\mathscr{F}_{s,d}$ is the unit ball of $C^s([0,1]^d)$ defined via

$$\mathscr{F}_{s,d} \coloneqq \big\{f \in C^s([0,1]^d) : \|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} \le 1, \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}^d \text{ with } \|\boldsymbol{\alpha}\|_1 \le s\big\}.$$

Theorem 2.3. *Given any $\rho, C_1, C_2, C_3 > 0$ and $s, d \in \mathbb{N}^+$, there exists $f \in \mathscr{F}_{s,d}$ such that, for any $J_0 > 0$, there exist $N, L \in \mathbb{N}^+$ with $NL \ge J_0$ satisfying*

$$\inf_{\phi\in \mathrm{NN}(\text{width}\le C_1 N\ln N;\ \text{depth}\le C_2 L\ln L)} \|\phi - f\|_{L^\infty([0,1]^d)} \ge C_3 N^{-(2s/d+\rho)} L^{-(2s/d+\rho)}.$$

Theorem 2.3 will be proved by contradiction. Assuming Theorem 2.3 is not true, we have the following claim, which can be disproved using the VC dimension upper bound in [13].

Claim 2.4. *There exist $\rho, C_1, C_2, C_3 > 0$ and $s, d \in \mathbb{N}^+$ such that, for any $f \in \mathscr{F}_{s,d}$, there exists $J_0 = J_0(\rho, C_1, C_2, C_3, s, d, f) > 0$ satisfying*

$$\inf_{\phi\in \mathrm{NN}(\text{width}\le C_1 N\ln N;\ \text{depth}\le C_2 L\ln L)} \|\phi - f\|_{L^\infty([0,1]^d)} \le C_3 N^{-(2s/d+\rho)} L^{-(2s/d+\rho)},$$

for all $N, L \in \mathbb{N}^+$ with $NL \ge J_0$.

What remaining is to show that Claim 2.4 is not true.

Disproof of Claim 2.4. Recall that the VC dimension of a class of functions is defined as the cardinality of the largest set of points that this class of functions can shatter. Denote the VC dimension of a function set $\mathcal{F}$ by $\mathrm{VCDim}(\mathcal{F})$. Set $\widetilde{N} = C_1 N\ln N$ and $\widetilde{L} = C_2 L \ln L$. Then by [13], there exists $C_4 > 0$ such that

$$\begin{aligned}&\mathrm{VCDim}\big(\mathrm{NN}(\#\text{input} = d;\ \text{width} \le \widetilde{N}; \text{depth} \le \widetilde{L})\big)\\ &\le C_4(\widetilde{N}\widetilde{L} + d + 2)(\widetilde{N} + 1)\widetilde{L}\ln\big((\widetilde{N}\widetilde{L} + d + 2)(\widetilde{N} + 1)\big) \coloneqq b_u(N, L),\end{aligned} \tag{2.2}$$

which comes from the fact the number of parameter of a ReLU FNN in $\mathrm{NN}(\#\text{input} = d;\ \text{width} \le \widetilde{N}; \text{depth} \le \widetilde{L})$ is less than $(\widetilde{N}\widetilde{L} + d + 2)(\widetilde{N} + 1)$.

Then we will use Claim 2.4 to estimate a lower bound $b_\ell(N, L) = \lfloor (NL)^{\frac{2}{d}+\frac{\rho}{2s}}\rfloor^d$ of

$$\mathrm{VCDim}\big(\mathrm{NN}(\#\text{input} = d;\ \text{width} \le \widetilde{N}; \text{depth} \le \widetilde{L})\big),$$

and this lower bound is asymptotically larger than $b_u(N, L)$, which leads to a contradiction.

More precisely, we will construct $\{f_\beta : \beta \in \mathscr{B}\} \subseteq \mathscr{F}_{s,d}$, which can shatter $b_\ell(N, L) = K^d$ points, where $\mathscr{B}$ is a set defined later and $K = \lfloor (NL)^{\frac{2}{d}+\frac{\rho}{2s}}\rfloor$. Then by Claim 2.4, we will show that there exists a set of ReLU FNNs $\{\phi_\beta : \beta \in \mathscr{B}\}$ with width bounded by $\widetilde{N}$ and depth bounded by $\widetilde{L}$ such that this set can shatter $b_\ell(N, L)$ points. Finally,

$b_\ell(N,L) = K^d = \lfloor (NL)^{\frac{2}{d}+\frac{\rho}{2s}} \rfloor^d$ is asymptotically larger than $b_u(N,L)$, which leads to a contradiction. More details can be found below.

Step 1: Construct $\{f_\beta : \beta \in \mathscr{B}\} \subseteq \mathscr{F}_{s,d}$ that scatters $b_\ell(N,L)$ points.

First, there exists $\widetilde{g} \in C^\infty([0,1]^d)$ such that $\widetilde{g}(0) = 1$ and $\widetilde{g}(\boldsymbol{x}) = 0$ for $\|\boldsymbol{x}\|_2 \geq 1/3$.[5] And we can find a constant $C_5 > 0$ such that $g \coloneqq \widetilde{g}/C_5 \in \mathscr{F}_{s,d}$.

Divide $[0,1]^d$ into K^d non-overlapping sub-cubes $\{Q_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$ as follows:

$$Q_{\boldsymbol{\theta}} \coloneqq \Big\{\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T \in [0,1]^d : x_i \in [\tfrac{\theta_i - 1}{K}, \tfrac{\theta_i}{K}],\ i = 1, 2, \cdots, d\Big\},$$

for any index vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_d]^T \in \{1, 2, \cdots, K\}^d$. Denote the center of $Q_{\boldsymbol{\theta}}$ by $\boldsymbol{x}_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d$. Define

$$\mathscr{B} \coloneqq \Big\{\beta : \beta \text{ is a map from } \{1, 2, \cdots, K\}^d \text{ to } \{-1, 1\}\Big\}.$$

For each $\beta \in \mathscr{B}$, we define, for any $\boldsymbol{x} \in \mathbb{R}^d$,

$$f_\beta(\boldsymbol{x}) \coloneqq \sum_{\boldsymbol{\theta} \in \{1,2,\cdots,K\}^d} K^{-s} \beta(\boldsymbol{\theta}) g_{\boldsymbol{\theta}}(\boldsymbol{x}), \quad \text{where } g_{\boldsymbol{\theta}}(\boldsymbol{x}) = g\big(K \cdot (\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{\theta}})\big).$$

We will show $f_\beta \in \mathscr{F}_{s,d}$ for each $\beta \in \{1, 2, \cdots, K\}^d$. We denote the support of a function h by $\mathrm{supp}(h) \coloneqq \{\boldsymbol{x} : h(\boldsymbol{x}) \neq 0\}$. Then by the definition of g, we have

$$\mathrm{supp}(g_{\boldsymbol{\theta}}) \subseteq \tfrac{2}{3} Q_{\boldsymbol{\theta}}, \quad \text{for any } \boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d,$$

where $\frac{2}{3} Q_{\boldsymbol{\theta}}$ denotes the cube satisfying two conditions: 1) the sidelength is $2/3$ of $Q_{\boldsymbol{\theta}}$'s; 2) the center is the same as $Q_{\boldsymbol{\theta}}$'s.

Now fix $\boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d$ and $\beta \in \mathscr{B}$, for any $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$ and $\alpha \in \mathbb{N}^d$, we have

$$\partial^{\boldsymbol{\alpha}} f_\beta(\boldsymbol{x}) = K^{-s} \beta(\boldsymbol{\theta}) \partial^{\boldsymbol{\alpha}} g_{\boldsymbol{\theta}}(\boldsymbol{x}) = K^{-s} \beta(\boldsymbol{\theta}) K^{\|\alpha\|_1} \partial^{\boldsymbol{\alpha}} g\big(K(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{\theta}})\big),$$

which implies $|\partial^{\boldsymbol{\alpha}} f_\beta(\boldsymbol{x})| = \Big|K^{-(s - \|\alpha\|_1)} \partial^{\boldsymbol{\alpha}} g\big(K(\boldsymbol{x} - \boldsymbol{x}_{\boldsymbol{\theta}})\big)\Big| \leq 1$ if $\|\boldsymbol{\alpha}\|_1 \leq s$. Since $\boldsymbol{\theta}$ is arbitrary and $[0,1]^d = \cup_{\boldsymbol{\theta} \in \{1,2,\cdots,K\}^d} Q_{\boldsymbol{\theta}}$, we have $f_\beta \in \mathscr{F}_{s,d}$ for each $\beta \in \mathscr{B}$. And it is easy to check that $\{f_\beta : \beta \in \mathscr{B}\}$ can shatter $\big\{\boldsymbol{x}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d\big\}$, which has $b_\ell(N,L) = K^d$ elements.

Step 2: Construct $\{\phi_\beta : \beta \in \mathscr{B}\}$ based on $\{f_\beta : \beta \in \mathscr{B}\}$ to scatter $b_\ell(N,L)$ points.

By Claim 2.4, for each $f_\beta \in \{f_\beta : \beta \in \mathscr{B}\}$, there exists $J_\beta > 0$ such that, for all $N, L \in \mathbb{N}$ with $NL \geq J_\beta$, there exists $\phi_\beta \in \mathrm{NN}(\text{width} \leq \widetilde{N}; \text{depth} \leq \widetilde{L})$

$$|f_\beta(\boldsymbol{x}) - \phi_\beta(\boldsymbol{x})| \leq C_3 (NL)^{-s(\frac{2}{d} + \frac{\rho}{s})}, \quad \text{for any } \boldsymbol{x} \in [0,1]^d.$$

Set $J_1 = \max\{J_\beta : \beta \in \mathscr{B}\}$. Note that there exists $J_2 > 0$ such that, for $N, L \in \mathbb{N}^+$ with $NL \geq J_2$,

$$\tfrac{K^{-s}}{C_5} = \tfrac{1}{C_5} \lfloor (NL)^{\frac{2}{d} + \frac{\rho}{2s}} \rfloor^{-s} > C_3 (NL)^{-s(\frac{2}{d} + \frac{\rho}{s})}.$$

Now fix $\beta \in \mathscr{B}$ and $\boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d$, for $N, L \in \mathbb{N}^+$ with $NL \geq \max\{J_1, J_2\}$, we have

$$|f_\beta(\boldsymbol{x}_{\boldsymbol{\theta}})| = K^{-s} g_{\boldsymbol{\theta}}(\boldsymbol{x}_{\boldsymbol{\theta}}) = \tfrac{K^{-s}}{C_5} > C_3 (NL)^{-s(\frac{2}{d} + \frac{\rho}{s})} \geq |f_\beta(\boldsymbol{x}_{\boldsymbol{\theta}}) - \phi_\beta(\boldsymbol{x}_{\boldsymbol{\theta}})|.$$

[5] For example, we can set $\widetilde{g}(\boldsymbol{x}) = C \exp(\frac{1}{\|3\boldsymbol{x}\|_2^2 - 1})$ if $\|\boldsymbol{x}\|_2 < 1/3$ and $\widetilde{g}(\boldsymbol{x}) = 0$ if $\|\boldsymbol{x}\|_2 \geq 1/3$, where C is a proper constant such that $\widetilde{g}(0) = 1$.

In other words, for any $\beta \in \mathscr{B}$ and $\boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d$, $f_\beta(\boldsymbol{x_\theta})$ and $\phi_\beta(\boldsymbol{x_\theta})$ have the same sign. Then $\{\phi_\beta : \beta \in \mathscr{B}\}$ shatters $\big\{\boldsymbol{x_\theta} : \boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d\big\}$ since $\{f_\beta : \beta \in \mathscr{B}\}$ shatters $\big\{\boldsymbol{x_\theta} : \boldsymbol{\theta} \in \{1, 2, \cdots, K\}^d\big\}$ as discussed in Step 1. Hence,

$$\mathrm{VCDim}\big(\{\phi_\beta : \beta \in \mathscr{B}\}\big) \geq K^d = b_\ell(N, L), \tag{2.3}$$

for $N, L \in \mathbb{N}^+$ with $NL \geq \max\{J_1, J_2\}$.

Step 3: Contradiction.

By Equation (2.2) and (2.3), for any $N, L \in \mathbb{N}$ with $NL \geq \max\{J_1, J_2\}$, we have

$$b_\ell(N, L) \leq \mathrm{VCDim}\big(\{\phi_\beta : \beta \in \mathscr{B}\}\big) \leq \mathrm{VCDim}\big(\mathrm{NN}(\text{width} \leq \widetilde{N};\ \text{depth} \leq \widetilde{L})\big) \leq b_u(N, L),$$

implying that

$$\begin{aligned}
\lfloor (NL)^{2/d + \rho/(2\alpha)} \rfloor^d &\leq C_4(\widetilde{L}\widetilde{N} + d + 2)(\widetilde{N} + 1)\widetilde{L} \ln\big((\widetilde{L}\widetilde{N} + d + 2)(\widetilde{N} + 1)\big) \\
&= \mathcal{O}\big(\widetilde{N}^2\widetilde{L}^2 \ln(\widetilde{N}^2\widetilde{L})\big) \\
&= \mathcal{O}\Big((C_1 N \ln N)^2 (C_2 L \ln L)^2 \ln\big((C_1 N \ln N)^2 C_2 L \ln L\big)\Big),
\end{aligned}$$

which is a contradiction for sufficiently large $N, L \in \mathbb{N}$. So we finish the proof. $\square$

We would like to remark that the approximation rate $\mathcal{O}(N^{-(2s/d + \rho_1)} L^{-(2s/d + \rho_2)})$ for $\rho_1, \rho_2 \geq 0$ with $\rho_1 + \rho_2 > 0$ is not achievable either. The argument follows similar ideas as in the proof above.

3 Proof of Theorem 2.1

Intuitively speaking, Theorem 2.1 shows that: if a ReLU FNN g approximates f well except for a trifling region, then we can extend g to approximate f well on the whole domain. For example, if g approximates a one-dimensional continuous function f well except for a region in $\mathbb{R}$ with a sufficiently small measure δ, then $\mathrm{mid}\big(g(x+\delta), g(x), g(x-\delta)\big)$ can approximate f well on the whole domain, where $\mathrm{mid}(\cdot, \cdot, \cdot)$ is a function returning the middle value of three inputs and can be implemented via a ReLU FNN as shown in Lemma 3.1. This key idea is called the horizontal shift (translation) of g in this paper.

Lemma 3.1. *There exists a ReLU FNN ϕ with width* 14 *and depth* 2 *such that*

$$\mathrm{mid}(x_1, x_2, x_3) = \phi(x_1, x_2, x_3).$$

Proof. Let σ be the ReLU activation function, i.e., $\sigma(x) = \max\{0, x\}$. Recall the fact

$$x = \sigma(x) - \sigma(-x) \quad \text{and} \quad |x| = \sigma(x) + \sigma(-x), \quad \text{for any } x \in \mathbb{R}.$$

Therefore,

$$\max(x_1, x_2) = \tfrac{x_1 + x_2 + |x_1 - x_2|}{2} = \tfrac{1}{2}\sigma(x_1 + x_2) - \tfrac{1}{2}\sigma(-x_1 - x_2) + \tfrac{1}{2}\sigma(x_1 - x_2) + \tfrac{1}{2}\sigma(x_2 - x_1).$$

So there exists a ReLU FNN ψ_1 with width 4 and depth 1 such that $\psi_1(x_1,x_2) = \max(x_1,x_2)$ for any $x_1,x_2 \in \mathbb{R}$. So for any $x_1,x_2,x_3 \in \mathbb{R}$,

$$\max(x_1,x_2,x_3) = \max\big(\max(x_1,x_2),x_3\big) = \psi_1\big(\psi_1(x_1,x_2),\sigma(x_3)-\sigma(-x_3)\big) \coloneqq \phi_1(x_1,x_2,x_3).$$

So ϕ_1 can be implemented by a ReLU FNN with width 6 and depth 2. Similarly, we can construct a ReLU FNN ϕ_2 with width 6 and depth 2 such that

$$\phi_2(x_1,x_2,x_3) = \min(x_1,x_2,x_3), \quad \text{for any } x_1,x_2,x_3 \in \mathbb{R}.$$

Notice that

$$\begin{aligned}\mathrm{mid}(x_1,x_2,x_3) &= x_1+x_2+x_3-\max(x_1,x_2,x_3)-\min(x_1,x_2,x_3)\\ &= \sigma(x_1+x_2+x_3)-\sigma(-x_1-x_2-x_3)-\phi_1(x_1,x_2,x_3)-\phi_2(x_1,x_2,x_3).\end{aligned}$$

Hence, $\mathrm{mid}(x_1,x_2,x_3)$ can be implemented by a ReLU FNN ϕ with width 14 and depth 2, which means we finish the proof. □

The next lemma shows a simple but useful property of the $\mathrm{mid}(x_1,x_2,x_3)$ function that helps to exclude poor approximation in the trifling region.

Lemma 3.2. *For any $\varepsilon > 0$, if at least two of $\{x_1,x_2,x_3\}$ are in $\mathcal{B}(y,\varepsilon)$, then $\mathrm{mid}(x_1,x_2,x_3) \in \mathcal{B}(y,\varepsilon)$.*

Proof. Without loss of generality, we may assume $x_1,x_2 \in \mathcal{B}(y,\varepsilon)$ and $x_1 \le x_2$. Then the proof can be divided into three cases.

1. If $x_3 < x_1$, then $\mathrm{mid}(x_1,x_2,x_3) = x_1 \in \mathcal{B}(y,\varepsilon)$.
2. If $x_1 \le x_3 \le x_2$, then $\mathrm{mid}(x_1,x_2,x_3) = x_3 \in \mathcal{B}(y,\varepsilon)$ since $y-\varepsilon \le x_1 \le x_3 \le x_2 \le y+\varepsilon$.
3. If $x_2 < x_3$, then $\mathrm{mid}(x_1,x_2,x_3) = x_2 \in \mathcal{B}(y,\varepsilon)$.

So we finish the proof. □

Next, given a function g approximating f well on [0,1] except for a trifling region, Lemma 3.3 below shows how to use the $\mathrm{mid}(x_1,x_2,x_3)$ function to construct a new function ϕ uniformly approximating f well on $[0,1]$, leveraging the useful property of $\mathrm{mid}(x_1,x_2,x_3)$ in Lemma 3.2.

Lemma 3.3. *Given $\epsilon > 0$, $K \in \mathbb{N}^+$, and $\delta > 0$ with $\delta \le \frac{1}{3K}$, assume g is defined on $\mathbb{R}$ and $f,g \in C([0,1])$ with*

$$|f(x)-g(x)| \le \varepsilon, \quad \text{for any } x \in [0,1]\backslash\Omega(K,\delta,1).$$

Then

$$|\phi(x)-f(x)| \le \varepsilon + \omega_f(\delta), \quad \text{for any } x \in [0,1],$$

where

$$\phi(x) \coloneqq \mathrm{mid}\big(g(x-\delta),g(x),g(x+\delta)\big), \quad \text{for any } x \in \mathbb{R}.$$

Proof. Divide $[0,1]$ into K parts $Q_k = [\frac{k}{K}, \frac{k+1}{K}]$ for $k = 0, 1, \cdots, K-1$. For each k, we write

$$Q_k = Q_{k,1} \cup Q_{k,2} \cup Q_{k,3} \cup Q_{k,4},$$

where $Q_{k,1} = [\frac{k}{K}, \frac{k}{K} + \delta]$, $Q_{k,2} = [\frac{k}{K} + \delta, \frac{k+1}{K} - 2\delta]$, $Q_{k,3} = [\frac{k+1}{K} - 2\delta, \frac{k+1}{K} - \delta]$, and $Q_{k,4} = [\frac{k+1}{K} - \delta, \frac{k+1}{K}]$.

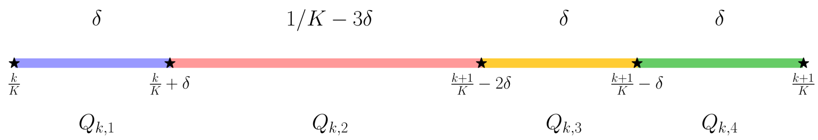


Figure 2: Illustrations of $Q_{k,i}$ for $i = 1, 2, 3, 4$.

Notice that $Q_{k+1,4} \subseteq [0,1] \backslash \Omega(K, \delta, 1)$ and $Q_{k,i} \subseteq [0,1] \backslash \Omega(K, \delta, 1)$ for $k = 0, 1, \cdots, k-1$, $i = 1, 2, 3$. For any $k \in \{0, 1, \cdots, K-1\}$, we consider the following four cases.

Case 1: $x \in Q_{k,1}$.

If $x \in Q_{k,1}$, then $x \in [0,1] \backslash \Omega(K, \delta, 1)$ and $x + \delta \in Q_{k,2} \cup Q_{k,3} \subseteq [0,1] \backslash \Omega(K, \delta, 1)$. It follows that

$$g(x) \in \mathcal{B}\big(f(x), \varepsilon\big) \subseteq \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big)$$

and

$$g(x + \delta) \in \mathcal{B}\big(f(x + \delta), \varepsilon\big) \subseteq \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big).$$

By Lemma 3.2, we get

$$\mathrm{mid}\big(g(x - \delta), g(x), g(x + \delta)\big) \in \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big).$$

Case 2: $x \in Q_{k,2}$.

If $x \in Q_{k,2}$, then $x - \delta, x, x + \delta \in [0,1] \backslash \Omega(K, \delta, 1)$. It follows that

$$g(x - \delta), g(x), g(x + \delta) \in \mathcal{B}\big(f(x), \varepsilon\big) \subseteq \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big),$$

which implies by Lemma 3.2 that

$$\mathrm{mid}\big(g(x - \delta), g(x), g(x + \delta)\big) \in \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big).$$

Case 3: $x \in Q_{k,3}$.

If $x \in Q_{k,3}$, then $x \in [0,1] \backslash \Omega(K, \delta, 1)$ and $x - \delta \in Q_{k,1} \cup Q_{k,2} \subseteq [0,1] \backslash \Omega(K, \delta, 1)$. It follows that

$$g(x) \in \mathcal{B}\big(f(x), \varepsilon\big) \subseteq \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big)$$

and

$$g(x - \delta) \in \mathcal{B}\big(f(x - \delta), \varepsilon\big) \subseteq \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big).$$

By Lemma 3.2, we get

$$\mathrm{mid}\big(g(x - \delta), g(x), g(x + \delta)\big) \in \mathcal{B}\big(f(x), \varepsilon + \omega_f(\delta)\big).$$

Case 4: $x \in Q_{k,4}$.

If $x \in Q_{k,4}$, we can divide this case into two sub-cases.

- If $k \in \{0,1,\cdots,K-2\}$, then $x-\delta \in Q_{k,3} \in [0,1]\backslash\Omega(K,\delta,1)$ and $x+\delta \in Q_{k+1,1} \subseteq [0,1]\backslash\Omega(K,\delta,1)$. It follows that
$$g(x-\delta) \in \mathcal{B}\big(f(x-\delta),\varepsilon\big) \subseteq \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big)$$
and
$$g(x+\delta) \in \mathcal{B}\big(f(x+\delta),\varepsilon\big) \subseteq \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big).$$
By Lemma 3.2, we get
$$\mathrm{mid}\big(g(x-\delta),g(x),g(x+\delta)\big) \in \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big).$$

- If $k = K-1$, then $x \in Q_{K-1,4} \subseteq [0,1]\backslash\Omega(K,\delta,1)$ and $x-\delta \in Q_{k,3} \subseteq [0,1]\backslash\Omega(K,\delta,1)$. It follows that
$$g(x) \in \mathcal{B}\big(f(x),\varepsilon\big) \subseteq \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big)$$
and
$$g(x-\delta) \in \mathcal{B}\big(f(x-\delta),\varepsilon\big) \subseteq \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big).$$
By Lemma 3.2, we get
$$\mathrm{mid}\big(g(x-\delta),g(x),g(x+\delta)\big) \in \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big).$$

Since $[0,1] = \cup_{k=0}^{K-1}\Big(\cup_{i=1}^{4} Q(k,i)\Big)$, we have
$$\mathrm{mid}\big(g(x-\delta),g(x),g(x+\delta)\big) \in \mathcal{B}\big(f(x),\varepsilon+\omega_f(\delta)\big), \quad \text{for any } x \in [0,1].$$
Notice that $\phi(x) = \mathrm{mid}\big(g(x-\delta),g(x),g(x+\delta)\big)$, it holds that
$$|\phi(x)-f(x)| \le \varepsilon+\omega_f(\delta), \quad \text{for any } x \in [0,1].$$
So we finish the proof. □

The next lemma below is an analog of Lemma 3.3.

Lemma 3.4. *Given $\varepsilon > 0$, $K \in \mathbb{N}^+$, and $\delta \in (0,\frac{1}{3K}]$, assume $f,g \in C([0,1]^d)$ with*
$$|f(\boldsymbol{x})-g(\boldsymbol{x})| \le \varepsilon, \quad \text{for any } \boldsymbol{x} \in [0,1]^d\backslash\Omega(K,\delta,d).$$
Let $\phi_0 = g$ and $\{\boldsymbol{e}_i\}_{i=1}^{d}$ be the standard basis in $\mathbb{R}^d$. By induction, we define
$$\phi_{i+1}(\boldsymbol{x}) \coloneqq \mathrm{mid}\big(\phi_i(\boldsymbol{x}-\delta\boldsymbol{e}_{i+1}),\phi_i(\boldsymbol{x}),\phi_i(\boldsymbol{x}+\delta\boldsymbol{e}_{i+1})\big), \quad \text{for } i=0,1,\cdots,d-1.$$
Let $\phi \coloneqq \phi_d$, then
$$|f(\boldsymbol{x})-\phi(\boldsymbol{x})| \le \varepsilon+d\cdot\omega_f(\delta), \quad \text{for any } \boldsymbol{x} \in [0,1]^d.$$

Proof. For $\ell = 0, 1, \cdots, d$, we denote

$$E_\ell \coloneqq \left\{ \boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T : x_i \in [0,1] \text{ for } i \le \ell,\ x_j \in [0,1] \backslash \Omega(K, \delta, 1) \text{ for } j > \ell \right\}.$$

Notice that $E_0 = [0,1]^d \backslash \Omega(K, \delta, d)$ and $E_d = [0,1]^d$. See Figure 3 for the illustration of E_ℓ.

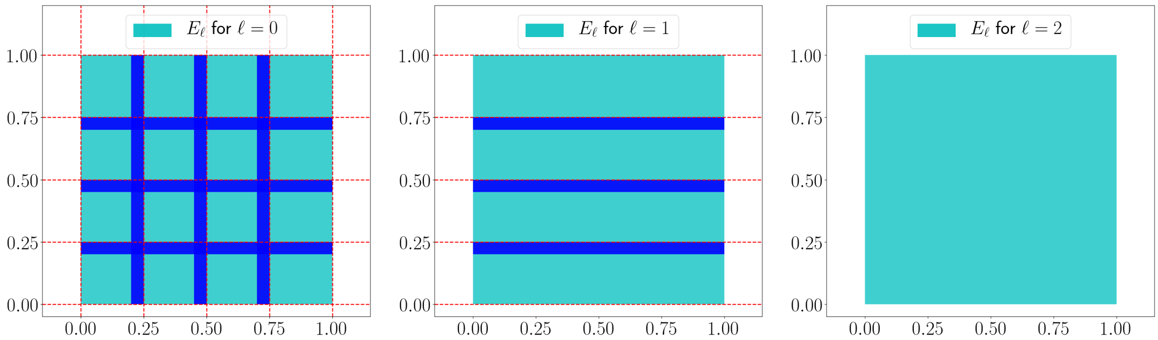


Figure 3: Illustrations of E_ℓ for $\ell = 0, 1, 2$ and $K = 4$.

We would like to construct $\phi_0, \phi_1, \cdots, \phi_d$ by induction such that, for each $\ell \in \{0, 1, \cdots, d\}$,

$$\phi_\ell(\boldsymbol{x}) \in \mathcal{B}\big(f(\boldsymbol{x}), \varepsilon + \ell \cdot \omega_f(\delta)\big), \quad \text{for any } \boldsymbol{x} \in E_\ell. \tag{3.1}$$

Let us first consider the case $\ell = 0$. Notice that $\phi_0 = g$ and $E_0 = [0,1]^d \backslash \Omega(K, \delta, d)$ for any $\boldsymbol{\theta} \in \{0, 1, \cdots, d\}^d$. Then we have

$$\phi_0(\boldsymbol{x}) \in \mathcal{B}\big(f(\boldsymbol{x}), \varepsilon\big), \quad \text{for any } \boldsymbol{x} \in E_0.$$

That is, Equation (3.1) is true for $\ell = 0$.

Now assume Equation (3.1) is true for $\ell = i$. We will prove that it also holds for $\ell = i + 1$. For any $\boldsymbol{x}^{[i]} \coloneqq [x_1, \cdots, x_i, x_{i+2}, \cdots, x_d]^T \in \mathbb{R}^{d-1}$, we set

$$\psi_{\boldsymbol{x}^{[i]}}(t) \coloneqq \phi_i(x_1, \cdots, x_i, t, x_{i+2}, \cdots, x_d), \quad \text{for any } t \in \mathbb{R},$$

and

$$f_{\boldsymbol{x}^{[i]}}(t) \coloneqq f(x_1, \cdots, x_i, t, x_{i+2}, \cdots, x_d), \quad \text{for any } t \in \mathbb{R}.$$

Since Equation (3.1) holds for $\ell = i$, by fixing $x_1, \cdots, x_i \in [0,1]$ and $x_{i+2}, \cdots, x_d \in [0,1] \backslash \Omega(K, \delta, 1)$, we have

$$\phi_i(x_1, \cdots, x_i, t, x_{i+2}, \cdots, x_d) \in \mathcal{B}\big(f(x_1, \cdots, x_i, t, x_{i+2}, \cdots, x_d), \varepsilon + i \cdot \omega_f(\delta)\big),$$

for any $t \in [0,1] \backslash \Omega(K, \delta, 1)$. It holds that

$$\psi_{\boldsymbol{x}^{[i]}}(t) \in \mathcal{B}\big(f_{\boldsymbol{x}^{[i]}}(t), \varepsilon + i \cdot \omega_f(\delta)\big), \quad \text{for any } t \in [0,1] \backslash \Omega(K, \delta, 1).$$

Then by Lemma 3.3, we get

$$\operatorname{mid}\big(\psi_{\boldsymbol{x}^{[i]}}(t - \delta), \psi_{\boldsymbol{x}^{[i]}}(t), \psi_{\boldsymbol{x}^{[i]}}(t + \delta)\big) \in \mathcal{B}\big(f_{\boldsymbol{x}^{[i]}}(t), \varepsilon + (i+1)\omega_f(\delta)\big), \quad \text{for any } t \in [0,1].$$

That is, for any $x_{i+1} = t \in [0,1]$,

$$\begin{aligned}&\operatorname{mid}\Big(\phi_i(x_1,\cdots,x_i,x_{i+1}-\delta,x_{i+2},\cdots,x_d),\phi_i(x_1,\cdots,x_d),\phi_i(x_1,\cdots,x_i,x_{i+1}+\delta,x_{i+2},\cdots,x_d)\Big)\\ &\in \mathcal{B}\Big(f(x_1,\cdots,x_d),\varepsilon+(i+1)\omega_f(\delta)\Big).\end{aligned}$$

Since $x_1,\cdots,x_i \in [0,1]$ and $x_{i+2},\cdots,x_d \in [0,1]\backslash\Omega(K,\delta,1)$ are arbitrary, then for any $\boldsymbol{x} \in E_{i+1}$,

$$\operatorname{mid}\big(\phi_i(\boldsymbol{x}-\delta\boldsymbol{e}_{i+1}),\phi_i(\boldsymbol{x}),\phi_i(\boldsymbol{x}+\delta\boldsymbol{e}_{i+1})\big) \in \mathcal{B}\big(f(\boldsymbol{x}),\varepsilon+(i+1)\omega_f(\delta)\big),$$

which implies

$$\phi_{i+1}(\boldsymbol{x}) \in \mathcal{B}\big(f(\boldsymbol{x}),\varepsilon+(i+1)\omega_f(\delta)\big), \quad \text{for any } \boldsymbol{x} \in E_{i+1}.$$

So we show that Equation (3.1) is true for $\ell = i+1$.

By the principle of induction, we have

$$\phi(\boldsymbol{x}) \coloneqq \phi_d(\boldsymbol{x}) \in \mathcal{B}\big(f(\boldsymbol{x}),\varepsilon+d\cdot\omega_f(\delta)\big), \quad \text{for any } \boldsymbol{x} \in E_d = [0,1]^d.$$

Therefore,

$$|\phi(\boldsymbol{x}) - f(\boldsymbol{x})| \le \varepsilon + d\cdot\omega_f(\delta), \quad \text{for any } \boldsymbol{x} \in [0,1]^d,$$

which means we finish the proof. □

Now we are ready to prove Theorem 2.1.

Proof of Theorem 2.1. Set $\phi_0 = \widetilde{\phi}$ and define ϕ_i for $i = 1,2,\cdots,d-1$ by induction as follows:

$$\phi_{i+1}(\boldsymbol{x}) \coloneqq \operatorname{mid}\big(\phi_i(\boldsymbol{x}-\delta\boldsymbol{e}_{i+1}),\phi_i(\boldsymbol{x}),\phi_i(\boldsymbol{x}+\delta\boldsymbol{e}_{i+1})\big), \quad \text{for } i = 0,1,\cdots,d-1.$$

Notice that $\phi_0 = \widetilde{\phi}$ is a ReLU FNN with width N and depth L and $\operatorname{mid}(x_1,x_2,x_3)$ can be implemented by a ReLU FNN with width 14 and depth 2. Hence, by the above induction formula, ϕ_d can be implemented with a ReLU FNN with width $3^d\max\{N,5\} \le 3^d(N+4)$ and depth $L+2d$. Finally, let $\phi \coloneqq \phi_d$. Then by Lemma 3.4, we have

$$|f(\boldsymbol{x}) - \phi(\boldsymbol{x})| \le \varepsilon + d\cdot\omega_f(\delta), \quad \text{for any } \boldsymbol{x} \in [0,1]^d.$$

So we finish the proof. □

4 Proof of Theorem 2.2

In this section, we prove Theorem 2.2, a weaker version of the main theorem of this paper (Theorem 1.1) targeting a ReLU FNN constructed to approximate a smooth function outside the trifling region. The main idea is to construct ReLU FNNs through Taylor expansions of smooth functions. We first discuss the sketch of the proof in Section 4.1 and give the detailed proof in Section 4.2.

4.1 Sketch of the proof of Theorem 2.2

Let $K = \mathcal{O}(N^{2/d}L^{2/d})$. For any $\boldsymbol{\theta} \in \{0,1,\cdots,K-1\}^d$ and $\boldsymbol{x} \in \{\boldsymbol{z} : \frac{\theta_i}{K} \le z_i \le \frac{\theta_i+1}{K},\ i = 1,2,\cdots,d\}$, there exists $\xi_{\boldsymbol{x}} \in (0,1)$ such that

$$f(\boldsymbol{x}) = \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\theta}/K)}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} + \sum_{\|\boldsymbol{\alpha}\|_1 = s} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\theta}/K + \xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} \coloneqq \mathscr{T}_1 + \mathscr{T}_2,^{⑥}$$

where $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x} - \frac{\boldsymbol{\theta}}{K}$. It is clear that the magnitude of $\mathscr{T}_2$ is bounded by $\mathcal{O}(K^{-s}) = \mathcal{O}(N^{-2s/d}L^{-2s/d})$. So we only need to construct a ReLU FNN $\phi \in \mathcal{NN}\big(\text{width} \le \mathcal{O}(N);\ \text{depth} \le \mathcal{O}(L)\big)$ to approximate

$$\mathscr{T}_1 = \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\theta}/K)}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}}$$

with an error $\mathcal{O}(N^{-2s/d}L^{-2s/d})$. To approximate $\mathscr{T}_1$ well by ReLU FNNs, we need three key steps as follows.

- Construct a ReLU FNN $P_{\boldsymbol{\alpha}}$ to approximate the polynomial $\boldsymbol{h}^{\boldsymbol{\alpha}}$ for each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \le s-1$.

- Construct a ReLU FNN $\boldsymbol{\psi}$ to approximate a step function that reduces the function approximation problem to a point fitting problem at fixed grid points. For example, a ReLU FNN mapping $\boldsymbol{x}$ to $\boldsymbol{\theta}/K$ if $x_i \in [\theta_i/K, (\theta_i+1)/K)$ for $i = 1,2,\cdots,d$ and $\boldsymbol{\theta} \in \{0,1,\cdots,K-1\}^d$.

- Construct a ReLU FNN $\phi_{\boldsymbol{\alpha}}$ to approximate $\partial^{\boldsymbol{\alpha}} f$ via solving the point fitting problem in the last step, i.e., $\phi_{\boldsymbol{\alpha}}$ fits $\partial^{\boldsymbol{\alpha}} f$ on given grid points for each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \le s-1$.

We will establish three propositions corresponding to these three steps above. Before showing this construction, we first summarize several propositions as follows. They will be applied to support the construction of the desired ReLU FNNs. Their proofs will be available in the next section.

First, we construct a ReLU FNN $P_{\boldsymbol{\alpha}}$ to approximate $\boldsymbol{h}^{\boldsymbol{\alpha}}$ according to Proposition 4.1 below, a general proposition for approximating multivariable polynomials.

Proposition 4.1. *Assume $P(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 = k \ge 2$. For any $N, L \in \mathbb{N}^+$, there exists a ReLU FNN ϕ with width $9(N+1)+k-2$ and depth $7k(k-1)L$ such that*

$$|\phi(\boldsymbol{x}) - P(\boldsymbol{x})| \le 9(k-1)(N+1)^{-7kL}, \quad \text{for any } \boldsymbol{x} \in [0,1]^d.$$

Proposition 4.1 shows that ReLU FNNs with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ is able to approximate polynomials with the rate $\mathcal{O}(N)^{-\mathcal{O}(L)}$. This reveals the power of depth in ReLU FNNs for approximating polynomials, from function compositions. The starting point of a good approximation of functions is to approximate polynomials with high accuracy. In classical approximation theory, approximation power of any numerical

⑥Notice that $\sum_{\|\boldsymbol{\alpha}\|_1=s}$ is short for $\sum_{\|\boldsymbol{\alpha}\|_1=s,\, \boldsymbol{\alpha}\in\mathbb{N}^d}$. For simplicity, we will use the same notation throughout the present paper.

scheme depends on the degree of polynomials that can be locally reproduced. Being able to approximate polynomials with high accuracy of deep ReLU FNNs plays a vital role in the proof of Theorem 1.1. It is interesting to study whether there is any other function space with reasonable size, besides polynomial space, having an exponential rate $\mathcal{O}(N)^{-\mathcal{O}(L)}$ when approximated by ReLU FNNs. Obviously, the space of smooth function is too big due to the optimality of Theorem 1.1 as shown in Theorem 2.3.

Proposition 4.1 can be generalized to the case of polynomials defined on an arbitrary hypercube $[a,b]^d$. Let us give an example for the polynomial xy below. Its proof will be provided later in Section 5.

Lemma 4.2. *For any $N, L \in \mathbb{N}^+$ and $a, b \in \mathbb{R}$ with $a < b$, there exists a ReLU FNN ϕ with width $9N + 1$ and depth L such that*

$$|\phi(x,y) - xy| \le 6(b-a)^2 N^{-L}, \quad \text{for any } x, y \in [a,b].$$

Second, we construct a step function $\boldsymbol{\psi}$ mapping $\boldsymbol{x} \in \{\boldsymbol{z} : \frac{\theta_i}{K} \le z_i < \frac{\theta_i+1}{K},\ i = 1, 2, \cdots, d\}$ to $\frac{\boldsymbol{\theta}}{K}$. We only need to approximate one-dimensional step functions, because in the multidimensional case we can simply set $\boldsymbol{\psi}(\boldsymbol{x}) = [\psi(x_1), \psi(x_2), \cdots, \psi(x_d)]^T$, where ψ is a one-dimensional step function. In particular, we shall construct ReLU FNNs with width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ to approximate step functions with $\mathcal{O}(K) = \mathcal{O}(N^{2/d}L^{2/d})$ "steps" as in Proposition 4.3 below.

Proposition 4.3. *For any $N, L, d \in \mathbb{N}^+$ and $\delta > 0$ with $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$ and $\delta \le \frac{1}{3K}$, there exists a one-dimensional ReLU FNN ϕ with width $4N + 5$ and depth $4L + 4$ such that*

$$\phi(x) = \frac{k}{K}, \quad \text{if } x \in [\tfrac{k}{K}, \tfrac{k+1}{K} - \delta \cdot 1_{\{k<K-1\}}] \text{ for } k = 0, 1, \cdots, K-1.$$

Finally, we construct a ReLU FNN $\phi_{\boldsymbol{\alpha}}$ to approximate $\partial^{\boldsymbol{\alpha}} f$ via solving a point fitting problem, i.e., we only need $\phi_{\boldsymbol{\alpha}}$ to approximate $\partial^{\boldsymbol{\alpha}} f$ well at grid points $\{\frac{\boldsymbol{\theta}}{K}\}$ as follows

$$\left|\phi_{\boldsymbol{\alpha}}(\tfrac{\boldsymbol{\theta}}{K}) - \partial^{\boldsymbol{\alpha}} f(\tfrac{\boldsymbol{\theta}}{K})\right| \le \mathcal{O}(N^{-2s/d} L^{-2s/d}), \quad \text{for any } \boldsymbol{\theta} \in \{0, 1, \cdots, K-1\}^d.$$

We can construct ReLU FNNs with width $\mathcal{O}(sN \ln N)$ and depth $\mathcal{O}(L \ln L)$ to fit $\mathcal{O}(N^2 L^2)$ points with an error $\mathcal{O}(N^{-2s} L^{-2s})$ by Proposition 4.4 below.

Proposition 4.4. *Given any $N, L, s \in \mathbb{N}^+$ and $\xi_i \in [0,1]$ for $i = 0, 1, \cdots, N^2L^2 - 1$, there exists a ReLU FNN ϕ with width $8s(2N+3)\log_2(4N)$ and depth $(5L+8)\log_2(2L)$ such that*

1. *$|\phi(i) - \xi_i| \le N^{-2s} L^{-2s}$, for $i = 0, 1, \cdots, N^2L^2 - 1$;*
2. *$0 \le \phi(t) \le 1$, for any $t \in \mathbb{R}$.*

The proofs of Proposition 4.1, 4.3, and 4.4 can be found in Section 5.1, 5.2, and 5.3, respectively. Finally, let us summarize the main ideas of proving Theorem 1.1 in Table 2.

Table 2: A list of ReLU FNNs, their sizes, approximation targets, and approximation errors. The construction of the final network $\phi(\boldsymbol{x})$ is based on a sequence of sub-networks listed before $\phi(\boldsymbol{x})$. Recall that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{\psi}(\boldsymbol{x})$.

Target function	ReLU FNN	Width	Depth	Approximation error
Step function	$\boldsymbol{\psi}(\boldsymbol{x})$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	No error out of $\Omega(K,\delta,d)$
x_1x_2	$\widetilde{\phi}(x_1,x_2)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathscr{E}_1 = \mathcal{O}\big((N+1)^{-2s(L+1)}\big)$
$\boldsymbol{h}^{\boldsymbol{\alpha}}$	$P_{\boldsymbol{\alpha}}(\boldsymbol{h})$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathscr{E}_2 = \mathcal{O}\big((N+1)^{-7s(L+1)}\big)$
$\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))$	$\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))$	$\mathcal{O}(N\ln N)$	$\mathcal{O}(L\ln L)$	$\mathscr{E}_3 = \mathcal{O}\big(N^{-2s}L^{-2s}\big)$
$\sum_{\\|\boldsymbol{\alpha}\\|\le s-1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}}$	$\sum_{\\|\boldsymbol{\alpha}\\|\le s-1} \widetilde{\phi}\Big(\frac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big)$	$\mathcal{O}(N\ln N)$	$\mathcal{O}(L\ln L)$	$\mathcal{O}(\mathscr{E}_1+\mathscr{E}_2+\mathscr{E}_3)$
$f(\boldsymbol{x})$	$\phi(\boldsymbol{x}) \coloneqq \sum_{\\|\boldsymbol{\alpha}\\|\le s-1} \widetilde{\phi}\Big(\frac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{x}-\boldsymbol{\psi}(\boldsymbol{x}))\Big)$	$\mathcal{O}(N\ln N)$	$\mathcal{O}(L\ln L)$	$\mathcal{O}(\\|\boldsymbol{h}\\|_2^{-s}+\mathscr{E}_1+\mathscr{E}_2+\mathscr{E}_3)$ $\le \mathcal{O}(K^{-s}) = \mathcal{O}(N^{-2s/d}L^{-2s/d})$

4.2 Constructive proof

According to the key ideas of proving Theorem 2.2 we summarized in the previous sub-section, we are ready to present the detailed proof.

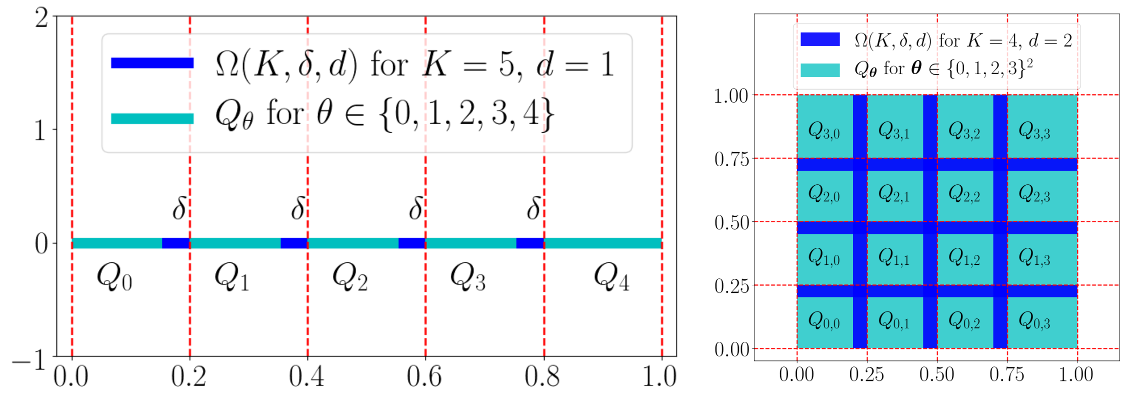
Proof of Theorem 2.2. The detailed proof can be divided into three steps as follows.

Step 1: Basic setting.

Let $\Omega(K,\delta,d)$ partition $[0,1]^d$ into K^d cubes $Q_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \{0,1,\cdots,K-1\}^d$. In particular, for each $\boldsymbol{\theta} = [\theta_1,\theta_2,\cdots,\theta_d]^T \in \{0,1,\cdots,K-1\}^d$, we define

$$Q_{\boldsymbol{\theta}} = \Big\{\boldsymbol{x} = [x_1,x_2,\cdots,x_d]^T : x_i \in \big[\tfrac{\theta_i}{K}, \tfrac{\theta_i+1}{K} - \delta\cdot 1_{\{\theta_i<K-1\}}\big],\ i = 1,2,\cdots,d\Big\}.$$

It is clear that $[0,1]^d = \Omega(K,\delta,d) \bigcup \big(\cup_{\boldsymbol{\theta}\in\{0,1,\cdots,K-1\}^d} Q_{\boldsymbol{\theta}}\big)$. See Figure 4 for the illustration of $Q_{\boldsymbol{\theta}}$.



(a) (b)

Figure 4: Illustrations of $Q_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \{0,1,\cdots,K-1\}^d$. (a) $K=5,\ d=1$. (b) $K=4,\ d=2$..

By Proposition 4.3, there exists a ReLU FNN ψ with width $4N+5$ and depth $4L+4$ such that

$$\psi(x) = \tfrac{k}{K}, \quad \text{if } x \in \big[\tfrac{k}{K}, \tfrac{k+1}{K} - \delta\cdot 1_{\{k<K-1\}}\big] \text{ for } k = 0,1,\cdots,K-1.$$

Then for each $\boldsymbol{\theta} \in \{0, 1, \cdots, K-1\}^d$, $\psi(x_i) = \frac{\theta_i}{K}$ if $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$ for $i = 1, 2, \cdots, d$.

Define

$$\boldsymbol{\psi}(\boldsymbol{x}) \coloneqq [\psi(x_1), \psi(x_2), \cdots, \psi(x_d)]^T, \quad \text{for any } \boldsymbol{x} \in [0,1]^d,$$

then

$$\boldsymbol{\psi}(\boldsymbol{x}) = \tfrac{\boldsymbol{\theta}}{K} \quad \text{if } \boldsymbol{x} \in Q_{\boldsymbol{\theta}}, \quad \text{for } \boldsymbol{\theta} \in \{0, 1, \cdots, K-1\}^d.$$

Now we fix a $\boldsymbol{\theta} \in \{0, 1, \cdots, K-1\}^d$ in the proof below. For any $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$, by the Taylor expansion, there exists a $\xi_{\boldsymbol{x}} \in (0,1)$ such that

$$f(\boldsymbol{x}) = \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} + \sum_{\|\boldsymbol{\alpha}\|_1 = s} \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}} \boldsymbol{h})}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}}, \quad \text{where } \boldsymbol{h} = \boldsymbol{x} - \boldsymbol{\psi}(\boldsymbol{x}).$$

Step 2: The construction of the target ReLU FNN.

By Lemma 4.2, there exists $\widetilde{\phi} \in \mathrm{NN}(\text{width} \le 9N + 10;\ \text{depth} \le 2sL + 2s)$ such that

$$|\widetilde{\phi}(x_1, x_2) - x_1 x_2| \le 216(N+1)^{-2s(L+1)} \coloneqq \mathscr{E}_1, \quad \text{for any } x_1, x_2 \in [-3, 3]. \tag{4.1}$$

If $2 \le \|\boldsymbol{\alpha}\|_1 \le s - 1$, by Proposition 4.1, there exist ReLU FNNs $P_{\boldsymbol{\alpha}}$ with width $9(N+1) + \|\boldsymbol{\alpha}\|_1 - 2 \le 9N + s + 6$ and depth $7s(\|\boldsymbol{\alpha}\|_1 - 1)(L+1) \le 7s^2(L+1)$ such that

$$|P_\alpha(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}| \le 9(\|\boldsymbol{\alpha}\|_1 - 1)(N+1)^{-7s(L+1)} \le 9s(N+1)^{-7s(L+1)}, \quad \text{for any } \boldsymbol{x} \in [0,1]^d.$$

And it is trivial to construct ReLU FNNs $P_{\boldsymbol{\alpha}}$ to approximate $\boldsymbol{x}^{\boldsymbol{\alpha}}$ when $\|\boldsymbol{\alpha}\|_1 \le 1$. Hence, for each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \le s-1$, there always exists $P_{\boldsymbol{\alpha}} \in \mathrm{NN}(\text{width} \le 9N + s + 6;\ \text{depth} \le 7s^2(L+1))$ such that

$$|P_\alpha(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}| \le 9s(N+1)^{-7s(L+1)} \coloneqq \mathscr{E}_2, \quad \text{for any } \boldsymbol{x} \in [0,1]^d. \tag{4.2}$$

For each $i = 0, 1, \cdots, K^d - 1$, define

$$\boldsymbol{\eta}(i) = [\eta_1, \eta_2, \cdots, \eta_d]^T \in \{0, 1, \cdots, K-1\}^d$$

such that $\sum_{j=1}^{d} \eta_j K^{j-1} = i$. We will drop the input i in $\boldsymbol{\eta}(i)$ later for simplicity. For each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \le s - 1$, define

$$\xi_{\boldsymbol{\alpha}, i} = \big(\partial^{\boldsymbol{\alpha}} f(\tfrac{\boldsymbol{\eta}}{K}) + 1\big)/2.$$

Notice that $K^d = \big(\lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor\big)^d \le N^2 L^2$ and $\xi_{\boldsymbol{\alpha}, i} \in [0,1]$ for $i = 0, 1, \cdots, K^d - 1$. By Proposition 4.4, there exists $\widetilde{\phi}_{\boldsymbol{\alpha}}$ in

$$\mathrm{NN}\big(\text{width} \le 8s(2N+3)\log_2(4N);\ \text{depth} \le (5L+8)\log_2(2L)\big)$$

such that

$$|\widetilde{\phi}_{\boldsymbol{\alpha}}(i) - \xi_{\boldsymbol{\alpha}, i}| \le N^{-2s} L^{-2s}, \quad \text{for } i = 0, 1, \cdots, K^d - 1 \text{ and } \|\boldsymbol{\alpha}\|_1 \le s - 1.$$

Define

$$\phi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \coloneqq 2\widetilde{\phi}_{\boldsymbol{\alpha}}\Big(\sum_{j=1}^{d} x_j K^j\Big) - 1, \quad \text{for any } \boldsymbol{x} = [x_1, x_2, \cdots, x_d]^d \in \mathbb{R}^d.$$

For each $\|\boldsymbol{\alpha}\|_1 \le s-1$, it is clear that $\phi_{\boldsymbol{\alpha}}$ is also in

$$\mathrm{NN}\big(\text{width} \le 8s(2N+3)\log_2(4N);\ \text{depth} \le (5L+8)\log_2(2L)\big).$$

Then for each $\boldsymbol{\eta} = [\eta_1, \eta_2, \cdots, \eta_d]^T \in \{0, 1, \cdots, K-1\}^d$ corresponding to $i = \sum_{j=1}^d \eta_j K^{j-1}$, each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \le s-1$, we have

$$\big|\phi_{\boldsymbol{\alpha}}(\tfrac{\boldsymbol{\eta}}{K}) - \partial^{\boldsymbol{\alpha}} f(\tfrac{\boldsymbol{\eta}}{K})\big| = \Big|2\widetilde{\phi}_{\boldsymbol{\alpha}}\Big(\sum_{j=1}^d \eta_j K^{j-1}\Big) - 1 - (2\xi_{\boldsymbol{\alpha},i} - 1)\Big| = 2|\widetilde{\phi}_{\boldsymbol{\alpha}}(i) - \xi_{\boldsymbol{\alpha},i}| \le 2N^{-2s}L^{-2s}.$$

It follows from $\boldsymbol{\psi}(\boldsymbol{x}) = \frac{\boldsymbol{\theta}}{K}$ for $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$ that

$$\big|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))\big| = \big|\phi_{\boldsymbol{\alpha}}(\tfrac{\boldsymbol{\theta}}{K}) - \partial^{\boldsymbol{\alpha}} f(\tfrac{\boldsymbol{\theta}}{K})\big| \le 2N^{-2s}L^{-2s} \coloneqq \mathscr{E}_3. \tag{4.3}$$

Now we are ready to construct the target ReLU FNN ϕ. Define

$$\phi(\boldsymbol{x}) \coloneqq \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}\big(\boldsymbol{x} - \boldsymbol{\psi}(\boldsymbol{x})\big)\Big), \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^d. \tag{4.4}$$

Step 3: Approximation error estimation.

Now let us estimate the error for any $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$. See Table 2 for a summary of the approximations errors. It is easy to check that $|f(\boldsymbol{x}) - \phi(\boldsymbol{x})|$ is bounded by

$$\begin{aligned}
&\Big|\sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} + \sum_{\|\boldsymbol{\alpha}\|_1 = s} \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} - \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \widetilde{\phi}\Big(\phi_{\boldsymbol{\alpha}}\big(\boldsymbol{\psi}(\boldsymbol{x})\big), P_{\boldsymbol{\alpha}}\big(\boldsymbol{x} - \boldsymbol{\psi}(\boldsymbol{x})\big)\Big)\Big| \\
&\le \sum_{\|\boldsymbol{\alpha}\|_1 = s} \Big|\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}}\Big| + \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big|\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} - \widetilde{\phi}\Big(\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})), P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big)\Big| \coloneqq \mathscr{I}_1 + \mathscr{I}_2.
\end{aligned}$$

Recall the fact $\sum_{\|\boldsymbol{\alpha}\|=s} 1 = (s+1)^{d-1}$ and $\sum_{\|\boldsymbol{\alpha}\| \le s-1} 1 = \sum_{i=0}^{s-1}(i+1)^{d-1} \le s^d$. For the first part $\mathscr{I}_1$, we have

$$\mathscr{I}_1 = \sum_{\|\boldsymbol{\alpha}\|_1 = s} \Big|\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}}\Big| \le \sum_{\|\boldsymbol{\alpha}\|_1 = s} \Big|\tfrac{1}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}}\Big| \le (s+1)^{d-1}K^{-s}.$$

Now let us estimate the second part $\mathscr{I}_2$ as follows.

$$\begin{aligned}
\mathscr{I}_2 &= \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big|\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} - \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big)\Big| \\
&\le \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big|\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} - \widetilde{\phi}\Big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big)\Big| \\
&\quad + \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big|\widetilde{\phi}\Big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \widetilde{\phi}\Big(\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})), P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big)\Big| \\
&\coloneqq \mathscr{I}_{2,1} + \mathscr{I}_{2,2}.
\end{aligned}$$

By Equation (4.2), $\mathscr{E}_2 \le 2$, and $\boldsymbol{x}^{\boldsymbol{\alpha}} \in [0,1]$ for any $\boldsymbol{x} \in [0,1]^d$, we have $P_{\boldsymbol{\alpha}}(\boldsymbol{x}) \in [-2,3] \subseteq [-3,3]$, for any $\boldsymbol{x} \in [0,1]^d$ and $\|\boldsymbol{\alpha}\|_1 \le s-1$. Together with Equation (4.1), we have, for any $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$,

$$\begin{aligned}
\mathscr{I}_{2,1} &= \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big| \tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} - \widetilde{\phi}\big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\big) \Big| \\
&\le \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big(\Big| \tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} - \tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big| + \Big| \tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) - \widetilde{\phi}\big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\big) \Big| \Big) \\
&\le \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big(\tfrac{1}{\boldsymbol{\alpha}!} \big| \boldsymbol{h}^{\boldsymbol{\alpha}} - P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \big| + \mathscr{E}_1 \Big) \le \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} (\mathscr{E}_2 + \mathscr{E}_1) \le s^d (\mathscr{E}_1 + \mathscr{E}_2).
\end{aligned}$$

In order to estimate $\mathscr{I}_{2,2}$, we need the following fact: for any $x_1, \bar{x}_1, x_2 \in [-3,3]$,

$$|\widetilde{\phi}(x_1,x_2) - \widetilde{\phi}(\bar{x}_1,x_2)| \le |\widetilde{\phi}(x_1,x_2) - x_1x_2| + |\widetilde{\phi}(\bar{x}_1,x_2) - \bar{x}_1x_2| + |x_1x_2 - \bar{x}_1x_2| \le 2\mathscr{E}_1 + 3|x_1 - \bar{x}_1|.$$

For each $\boldsymbol{\alpha} \in \mathbb{R}^d$ with $\|\boldsymbol{\alpha}\|_1 \le s-1$ and $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$, since $\mathscr{E}_3 \in [0,2]$ and $\frac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \in [-1,1]$ in Equation (4.3), we have $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) \in [-3,3]$. Together with $P_{\boldsymbol{\alpha}}(\boldsymbol{x}) \in [-3,3]$, we have, for any $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$,

$$\begin{aligned}
\mathscr{I}_{2,2} &= \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big| \widetilde{\phi}\big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\big) - \widetilde{\phi}\big(\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})), P_{\boldsymbol{\alpha}}(\boldsymbol{h})\big) \Big| \\
&\le \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} \Big(2\mathscr{E}_1 + 3 \Big| \tfrac{\partial^{\boldsymbol{\alpha}} f(\psi(\boldsymbol{x}))}{\boldsymbol{\alpha}!} - \phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) \Big| \Big) \le \sum_{\|\boldsymbol{\alpha}\|_1 \le s-1} (2\mathscr{E}_1 + 3\mathscr{E}_3) \le s^d (2\mathscr{E}_1 + 3\mathscr{E}_3).
\end{aligned}$$

Therefore, for any $\boldsymbol{x} \in Q_{\boldsymbol{\theta}}$,

$$\begin{aligned}
|f(\boldsymbol{x}) - \phi(\boldsymbol{x})| \le \mathscr{I}_1 + \mathscr{I}_2 &\le \mathscr{I}_1 + \mathscr{I}_{2,1} + \mathscr{I}_{2,2} \\
&\le (s+1)^{d-1} K^{-s} + s^d(\mathscr{E}_1 + \mathscr{E}_2) + s^d(2\mathscr{E}_1 + 3\mathscr{E}_3) \\
&\le (s+1)^d (K^{-s} + 3\mathscr{E}_1 + \mathscr{E}_2 + 3\mathscr{E}_3).
\end{aligned}$$

Since $\boldsymbol{\theta} \in \{0,1,\cdots,K-1\}^d$ is arbitrary and the fact $[0,1]^d \backslash \Omega(K,\delta,d) \subseteq \cup_{\boldsymbol{\theta} \in \{0,1,\cdots,K-1\}^d} Q_{\boldsymbol{\theta}}$, we have

$$|f(\boldsymbol{x}) - \phi(\boldsymbol{x})| \le (s+1)^d (K^{-s} + 3\mathscr{E}_1 + \mathscr{E}_2 + 3\mathscr{E}_3), \quad \text{for any } \boldsymbol{x} \in [0,1]^d \backslash \Omega(K,\delta,d).$$

Recall that $(N+1)^{-7s(L+1)} \le (N+1)^{-2s(L+1)} \le (N+1)^{-2s} 2^{-2sL} \le N^{-2s} L^{-2s}$ and $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor \ge \frac{N^{2/d} L^{2/d}}{8}$. Then we have

$$\begin{aligned}
&(s+1)^d (K^{-s} + 3\mathscr{E}_1 + \mathscr{E}_2 + 3\mathscr{E}_3) \\
&= (s+1)^d \Big(K^{-s} + 648(N+1)^{-2s(L+1)} + 9s(N+1)^{-7s(L+1)} + 6N^{-2s} L^{-2s} \Big) \\
&\le (s+1)^d \Big(8^s N^{-2s/d} L^{-2s/d} + (654 + 9s) N^{-2s} L^{-2s} \Big) \\
&\le (s+1)^d (8^s + 654 + 9s) N^{-2s/d} L^{-2s/d} \le 84(s+1)^d 8^s N^{-2s/d} L^{-2s/d}.
\end{aligned}$$

What remaining is to estimate the width and depth of ϕ. Recall that $\boldsymbol{\psi} \in \mathrm{NN}\big(\text{width} \le d(4N+5);\ \text{depth} \le 4(L+1)\big)$, $\widetilde{\phi} \in \mathrm{NN}\big(\text{width} \le 9N+10;\ \text{depth} \le 2s(L+1)\big)$, $P_{\boldsymbol{\alpha}} \in \mathrm{NN}\big(\text{width} \le 9N+s+6;\ \text{depth} \le 7s^2(L+1)\big)$, and $\phi_{\boldsymbol{\alpha}} \in \mathrm{NN}\big(\text{width} \le 8s(2N+3)\log_2(4N);\ \text{depth} \le (5L+8)\log_2(2L)\big)$ for $\boldsymbol{\alpha} \in \mathbb{N}$ with $\|\boldsymbol{\alpha}\|_1 \le s-1$. By Equation (4.4), ϕ can be implemented by a ReLU FNN with width $21s^{d+1}d(N+2)\log_2(4N)$ and depth $18s^2(L+2)\log_2(2L)$ as desired. So we finish the proof. $\square$

5 Proofs of Propositions in Section 4.1

In this section, we will prove all propositions in Section 4.1.

5.1 Proof of Proposition 4.1 for polynomial approximation

To prove Proposition 4.1, we will construct ReLU FNNs to approximate polynomials following the four steps below.

- $f(x) = x^2$. We approximate $f(x) = x^2$ by the combinations and compositions of "teeth functions".
- $f(x,y) = xy$. To approximate $f(x,y) = xy$, we use the result of the previous step and the fact $xy = 2\big((\frac{x+y}{2})^2 - (\frac{x}{2})^2 - (\frac{y}{2})^2\big)$.
- $f(x_1, x_2, \cdots, x_d) = x_1 x_2 \cdots x_d$. We approximate $f(x_1, x_2, \cdots, x_d) = x_1 x_2 \cdots x_d$ for any d via induction based on the result of the previous step.
- General multivariable polynomials. Any one-term polynomial of degree k can be written as $C z_1 z_2 \cdots z_k$, where C is a constant, then use the result of the previous step.

The idea of using "teeth functions" (see Figure 5) was first raised in [30] for approximating x^2 using FNNs with width 6 and depth $\mathcal{O}(L)$ and achieving an error $\mathcal{O}(2^{-L})$; our construction is different to and more general than that in [30], working for ReLU FNNs of width $\mathcal{O}(N)$ and depth $\mathcal{O}(L)$ for any N and L, and achieving an error $\mathcal{O}(N^{-L})$. As discussed above below Proposition 4.1, this $\mathcal{O}(N)^{-\mathcal{O}(L)}$ approximation rate of polynomial functions shows the power of depth in ReLU FNNs via function composition.

First, let us show how to construct ReLU FNNs to approximate $f(x) = x^2$.

Lemma 5.1. *For any $N, L \in \mathbb{N}^+$, there exists a ReLU FNN ϕ with width $3N$ and depth L such that*
$$|\phi(x) - x^2| \le N^{-L}, \quad \text{for any } x \in [0,1].$$

Proof. Define a set of teeth functions $T_i : [0,1] \to [0,1]$ by induction as follows. Let
$$T_1(x) = \begin{cases} 2x, & x \le \frac{1}{2}, \\ 2(1-x), & x > \frac{1}{2}, \end{cases}$$
and
$$T_i = T_{i-1} \circ T_1, \quad \text{for } i = 2, 3, \cdots.$$
It is easy to check that T_i has 2^{i-1} teeth and
$$T_{m+n} = T_m \circ T_n, \quad \text{for any } m, n \in \mathbb{N}^+.$$
See Figure 5 for more details of T_i.

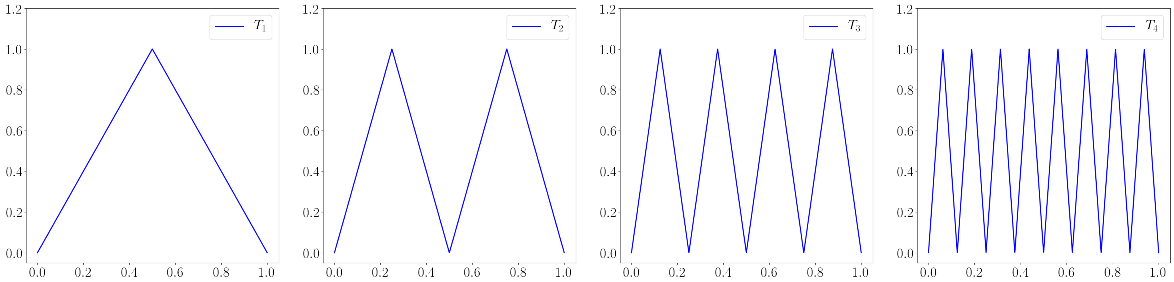


Figure 5: Illustrations of teeth functions T_1, T_2, T_3, and T_4.

Define piecewise linear functions $f_s : [0,1] \to [0,1]$ for $s \in \mathbb{N}^+$ satisfying the following two requirements (see Figure 6 for several examples of f_s).

- $f_s\left(\frac{j}{2^s}\right) = \left(\frac{j}{2^s}\right)^2$ for $j = 0, 1, 2, \cdots, 2^s$.
- $f_s(x)$ is linear between any two adjacent points of $\{\frac{j}{2^s} : j = 0, 1, 2, \cdots, 2^s\}$.

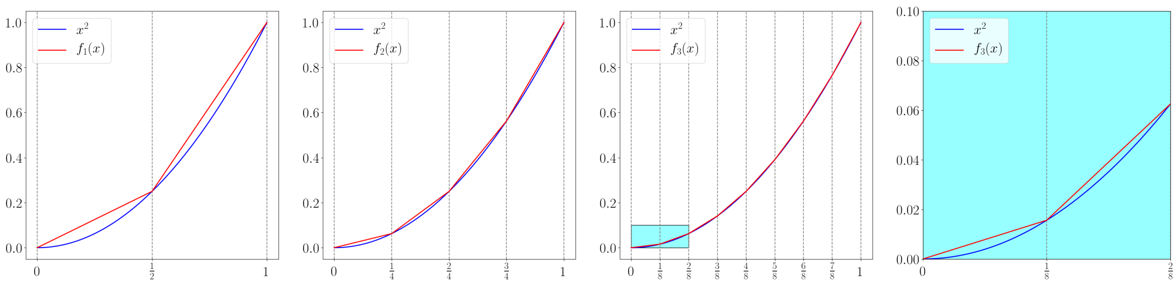


Figure 6: Illustrations of f_1, f_2, and f_3.

It follows from the fact $\frac{(x-h)^2+(x+h)^2}{2} - x^2 = h^2$ that

$$|x^2 - f_s(x)| \le 2^{-2(s+1)}, \quad \text{for any } x \in [0,1] \text{ and } s \in \mathbb{N}^+, \tag{5.1}$$

and

$$f_{i-1}(x) - f_i(x) = \frac{T_i(x)}{2^{2i}}, \quad \text{for any } x \in [0,1] \text{ and } i = 2, 3, \cdots.$$

Then

$$f_s(x) = f_1(x) + \sum_{i=2}^{s}(f_i - f_{i-1}) = x - (x - f_1(x)) - \sum_{i=2}^{s} \frac{T_i(x)}{2^{2i}} = x - \sum_{i=1}^{s} \frac{T_i(x)}{2^{2i}},$$

for any $x \in [0,1]$ and $s \in \mathbb{N}^+$.

Given $N \in \mathbb{N}^+$, there exists a unique $k \in \mathbb{N}^+$ such that $(k-1)2^{k-1} + 1 \le N \le k2^k$. For this k, we can construct a ReLU FNN ϕ as shown in Figure 7 to approximate f_s. Notice that T_i can be implemented by a one-hidden-layer ReLU FNN with width 2^i. Hence, ϕ in Figure 7 has width $k2^k + 1 \le 3N$ ⑦ and depth $2L$.

In fact, ϕ in Figure 7 can be interpreted as a ReLU FNN with width $3N$ and depth L since half of the hidden layers have the identify function as their activation

⑦This inequality is clear for $k = 1, 2, 3, 4$. In the case $k \ge 5$, we have $k2^k + 1 \le \frac{k2^k+1}{N} N \le \frac{(k+1)2^k}{(k-1)2^{k-1}} N \le 2\frac{k+1}{k-1} N \le 3N$.

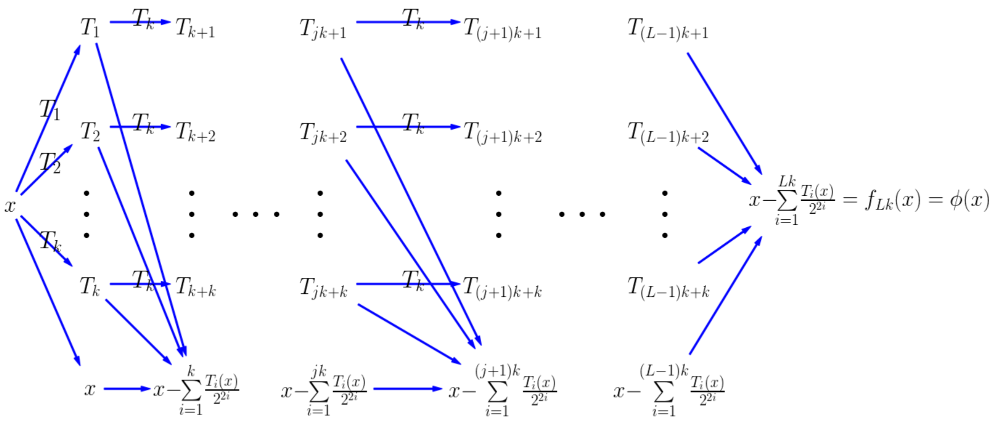


Figure 7: An illustration of the target ReLU FNN for approximating x^2. We drop the ReLU activation function in this figure since $T_i(x)$ is always positive for all $i \in \mathbb{N}^+$ and $x \in [0,1]$. Each arrow with T_k means that there is a ReLU FNN approximating T_k and mapping the function from the starting point of the arrow to generate a new function at the end point of the arrow. Arrows without T_k means a multiplication with a scalar contributing to one component of the linear combination in the bottom part of the network sketch.

functions. If all activation functions in a certain hidden layer are identity, the depth can be reduced by one by combining adjacent two linear transforms into one. For example, suppose $\boldsymbol{W}_1 \in \mathbb{R}^{N_1 \times N_2}$, $\boldsymbol{W}_2 \in \mathbb{R}^{N_2 \times N_3}$, and σ is an identity map that can be applied to vectors or matrices elementwisely, then $\boldsymbol{W}_1 \sigma(\boldsymbol{W}_2 \boldsymbol{x}) = \boldsymbol{W}_3 \boldsymbol{x}$ for any $\boldsymbol{x} \in \mathbb{R}^{N_3}$, where $\boldsymbol{W}_3 = \boldsymbol{W}_1 \cdot \boldsymbol{W}_2 \in \mathbb{R}^{N_1 \times N_3}$.

What remaining is to estimate the approximation error of $\phi(x) \approx x^2$. By Equation (5.1), for any $x \in [0,1]$, we have

$$|x^2 - \phi(x)| \le |x^2 - f_{Lk}| \le 2^{-2(Lk+1)} \le 2^{-2Lk} \le N^{-L},$$

where the last inequality comes from $N \le k 2^k \le 2^{2k}$. So we finish the proof. $\square$

We have constructed a ReLU FNN to approximate $f(x) = x^2$. By the fact $xy = 2\big((\frac{x+y}{2})^2 - (\frac{x}{2})^2 - (\frac{y}{2})^2\big)$, it is easy to construct a new ReLU FNN to approximate $f(x,y) = xy$ as follows.

Lemma 5.2. *For any $N, L \in \mathbb{N}^+$, there exists a ReLU FNN ϕ with width $9N$ and depth L such that*

$$|\phi(x,y) - xy| \le 6N^{-L}, \quad \text{for any } x, y \in [0,1].$$

Proof. By Lemma 5.1, there exists a ReLU FNN ψ with width $3N$ and depth L such that

$$|x^2 - \psi(x)| \le N^{-L}, \quad \text{for any } x \in [0,1].$$

Together with the fact

$$xy = 2\big((\tfrac{x+y}{2})^2 - (\tfrac{x}{2})^2 - (\tfrac{y}{2})^2\big), \quad \text{for any } x, y \in \mathbb{R},$$

we construct the target function ϕ as

$$\phi(x,y) \coloneqq 2\big(\psi(\tfrac{x+y}{2}) - \psi(\tfrac{x}{2}) - \psi(\tfrac{y}{2})\big), \quad \text{for any } x,y \in \mathbb{R}.$$

It follows that

$$\begin{aligned}|xy - \phi(x,y)| &= \Big|2\big((\tfrac{x+y}{2})^2 - (\tfrac{x}{2})^2 - (\tfrac{y}{2})^2\big) - 2\big(\psi(\tfrac{x+y}{2}) - \psi(\tfrac{x}{2}) - \psi(\tfrac{y}{2})\big)\Big| \\ &\le 2\,\Big|(\tfrac{x+y}{2})^2 - \psi(\tfrac{x+y}{2})\Big| + 2\,\Big|(\tfrac{x}{2})^2 - \psi(\tfrac{x}{2})\Big| + 2\,\Big|(\tfrac{y}{2})^2 - \psi(\tfrac{y}{2})\Big| \le 6N^{-L}.\end{aligned}$$

It is easy to check that ϕ is a network with width $9N$ and depth L. Therefore, we have finished the proof. □

Now let us prove Lemma 4.2 that shows how to construct a ReLU FNN to approximate $f(x,y) = xy$ on $[a,b]^2$ with arbitrary $a < b$, i.e., a rescaled version of Lemma 5.2.

Proof of Lemma 4.2. By Lemma 5.2, there exists a ReLU FNN ψ with width $9N$ and depth L such that

$$|\psi(\widetilde{x},\widetilde{y}) - \widetilde{x}\widetilde{y}| \le 6N^{-L}, \quad \text{for any } \widetilde{x},\widetilde{y} \in [0,1].$$

Set $x = a + (b-a)\widetilde{x}$ and $y = a + (b-a)\widetilde{y}$ for any $\widetilde{x},\widetilde{y} \in [0,1]$, we have

$$\Big|\psi(\tfrac{x-a}{b-a}, \tfrac{y-a}{b-a}) - \tfrac{x-a}{b-a}\tfrac{y-a}{b-a}\Big| \le 6N^{-L}, \quad \text{for any } x,y \in [a,b].$$

It follows that

$$\Big|(b-a)^2\psi(\tfrac{x-a}{b-a}, \tfrac{y-a}{b-a}) + a(x+y) - a^2 - xy\Big| \le 6(b-a)^2N^{-L}, \quad \text{for any } x,y \in [a,b].$$

Define

$$\phi(x,y) \coloneqq (b-a)^2\psi(\tfrac{x-a}{b-a}, \tfrac{y-a}{b-a}) + a(x+y) - a^2, \quad \text{for any } x,y \in \mathbb{R}.$$

Hence,

$$\big|\phi(x,y) - xy\big| \le 6(b-a)^2N^{-L}, \quad \text{for any } x,y \in [a,b].$$

Moreover, ϕ can be easily implemented by a ReLU FNN with width $9N+1$ and depth L. The result is proved. □

The next lemma constructs a ReLU FNN to approximate a multivariable function $f(x_1,x_2,\cdots,x_k) = x_1x_2\cdots x_k$ on $[0,1]^k$.

Lemma 5.3. *For any $N,L \in \mathbb{N}^+$, there exists a ReLU FNN ϕ with width $9(N+1)+k-2$ and depth $7k(k-1)L$ such that*

$$|\phi(\boldsymbol{x}) - x_1x_2\cdots x_k| \le 9(k-1)(N+1)^{-7kL}, \quad \text{for any } \boldsymbol{x} = [x_1,x_2,\cdots,x_k]^T \in [0,1]^k,\ k \ge 2.$$

Proof. By Lemma 4.2, there exists a ReLU FNN ϕ_1 with width $9(N+1)+1$ and depth $7kL$ such that

$$|\phi_1(x,y) - xy| \le 6(1.2)^2(N+1)^{-7kL} \le 9(N+1)^{-7kL}, \quad \text{for any } x,y \in [-0.1,1.1]. \tag{5.2}$$

Next, we construct $\phi_i : [0,1]^{i+1} \to [0,1]$ by induction for $i = 1,2,\cdots,k-1$ such that

- ϕ_i is a ReLU FNN with width $9(N+1)+i-1$ and depth $7kiL$ for each $i \in \{1,2,\cdots,k-1\}$.
- The following inequality holds for any $i \in \{1,2,\cdots,k-1\}$ and $x_1,x_2,\cdots,x_{i+1} \in [0,1]$

$$|\phi_i(x_1,\cdots,x_{i+1}) - x_1x_2\cdots x_{i+1}| \le 9i(N+1)^{-7kL}. \tag{5.3}$$

Now let us show the induction process in more details as follows.

1. When $i=1$, it is obvious that the two required conditions are true: 1) $9(N+1)+i-1 = 9(N+1)$ and $iL = L$ if $i=1$; 2) Equation (5.2) implies Equation (5.3) for $i=1$.
2. Now assume ϕ_i has been defined, then define

$$\phi_{i+1}(x_1,\cdots,x_{i+2}) \coloneqq \phi_1\big(\phi_i(x_1,\cdots,x_{i+1}),x_{i+2}\big), \quad \text{for any } x_1,\cdots,x_{i+2} \in \mathbb{R}.$$

Notice that the width and depth of ϕ_i are $9(N+1)+i-1$ and $7kiL$, respectively. Then ϕ_{i+2} can be implemented via a ReLU FNN with width $9(N+1)+i-1+1 = 9(N+1)+i$ and depth $7kiL + 7kL = 7k(i+1)L$.

By the hypothesis of induction, we have

$$|\phi_i(x_1,\cdots,x_{i+1}) - x_1x_2\cdots x_{i+1}| \le 9i(N+1)^{-7kL}.$$

Recall the fact $9i(N+1)^{-7kL} \le 9k2^{-7k} \le 9k\frac{1}{90k} = 0.1$ for any $N,L,k \in \mathbb{N}^+$ and $i \in \{1,2,\cdots,k-1\}$. It follows that

$$\phi_i(x_1,\cdots,x_{i+1}) \in [-0.1,1.1], \quad \text{for any } x_1,\cdots,x_{i+1} \in [0,1].$$

Therefore, for any $x_1,x_2,\cdots,x_{i+2} \in [0,1]$,

$$\begin{aligned}
&|\phi_{i+1}(x_1,\cdots,x_{i+2}) - x_1x_2\cdots x_{i+2}| = \big|\phi_1\big(\phi_i(x_1,\cdots,x_{i+1}),x_{i+2}\big) - x_1x_2\cdots x_{i+2}\big| \\
&\le \big|\phi_1\big(\phi_i(x_1,\cdots,x_{i+1}),x_{i+2}\big) - \phi_i(x_1,\cdots,x_{i+1})x_{i+2}\big| + \big|\phi_i(x_1,\cdots,x_{i+1})x_{i+2} - x_1x_2\cdots x_{i+2}\big| \\
&\le 9(N+1)^{-7kL} + 9i(N+1)^{-7kL} = 9(i+1)(N+1)^{-7kL}.
\end{aligned}$$

Now let $\phi \coloneqq \phi_{k-1}$, by the principle of induction, we have

$$|\phi(x_1,\cdots,x_k) - x_1x_2\cdots x_k| \le 9(k-1)(N+1)^{-7kL}, \quad \text{for any } x_1,x_2,\cdots,x_k \in [0,1].$$

So ϕ is the desired ReLU FNN with width $9(N+1)+k-2$ and depth $7k(k-1)L$. □

Now we are ready to prove Proposition 4.1 for approximating general multivariable polynomials via ReLU FNNs.

Proof of Proposition 4.1. Denote $\boldsymbol{\alpha} = [\alpha_1,\alpha_2,\cdots,\alpha_d]^T$ and let $[z_1,z_2,\cdots,z_k]^T$ be the vector such that

$$z_\ell = x_j, \quad \text{if } \sum_{i=1}^{j-1}\alpha_i < \ell \le \sum_{i=1}^{j}\alpha_i, \quad \text{for } j = 1,2,\cdots,d.$$

That is,

$$[z_1, z_2, \cdots, z_k]^T = [\overbrace{x_1, \cdots, x_1}^{\alpha_1 \text{ times}}, \overbrace{x_2, \cdots, x_2}^{\alpha_2 \text{ times}}, \cdots, \overbrace{x_d, \cdots, x_d}^{\alpha_d \text{ times}}]^T \in \mathbb{R}^k.$$

Then we have $P(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}} = z_1 z_2 \cdots z_k$.

We construct the target ReLU FNN in two steps. First, there exists a linear map $\boldsymbol{\phi}_1$ that duplicates inputs in $\boldsymbol{x}$ to form a new vector $[z_1, z_2, \cdots, z_k]^T$. Second, by Lemma 5.3, there exists such a ReLU FNN ϕ_2 with width $9(N+1)+k-2$ and depth $7k(k-1)L$ such that ϕ_2 maps $[z_1, z_2, \cdots, z_k]^T$ to $P(\boldsymbol{x}) = z_1 z_2 \cdots z_k$ within the target accuracy. Hence, we can construct our final target ReLU FNN via $\phi_2 \circ \boldsymbol{\phi}_1(\boldsymbol{x}) = \phi(\boldsymbol{x})$. By incorporating the linear map in $\boldsymbol{\phi}_1$ into the first linear map of ϕ, we can treat ϕ as a ReLU FNN with width $9(N+1)+k-2$ and depth $7k(k-1)L$ with a desired approximation accuracy. So, we finish the proof. □

5.2 Proof of Proposition 4.3 for step function approximation

To prove Proposition 4.3 in this sub-section, we will discuss how to pointwisely approximate step functions by ReLU FNNs except for a trifling region. Before proving Proposition 4.3, let us first introduce a basic lemma about fitting $\mathcal{O}(N_1 N_2)$ samples using a two-hidden-layer ReLU FNN with $\mathcal{O}(N_1 + N_2)$ neurons.

Lemma 5.4. *For any $N_1, N_2 \in \mathbb{N}^+$, given $N_1(N_2+1)+1$ samples $(x_i, y_i) \in \mathbb{R}^2$ with $x_0 < x_1 < \cdots < x_{N_1(N_2+1)}$ and $y_i \geq 0$ for $i = 0, 1, \cdots, N_1(N_2+1)$, there exists $\phi \in \mathrm{NN}(\#\mathrm{input} = 1; \mathrm{widthvec} = [2N_1, 2N_2+1])$ satisfying the following conditions.*

1. *$\phi(x_i) = y_i$ for $i = 0, 1, \cdots, N_1(N_2+1)$;*
2. *ϕ is linear on each interval $[x_{i-1}, x_i]$ for $i \notin \{(N_2+1)j : j = 1, 2, \cdots, N_1\}$.*

The above lemma is Proposition 2.1 of [27] and the reader is referred to [27] for its proof. Essentially, this lemma shows the equivalence of one-hidden-layer ReLU FNNs of size $\mathcal{O}(N^2)$ and two-hidden-layer ones of size $\mathcal{O}(N)$ to fit $\mathcal{O}(N^2)$ samples.

The next lemma below shows that special shallow and wide ReLU FNNs can be represented by deep and narrow ones. This lemma was proposed as Proposition 2.2 in [27].

Lemma 5.5. *Given any $N, L \in \mathbb{N}^+$, for arbitrary $\phi_1 \in \mathrm{NN}(\#\mathrm{input} = 1; \mathrm{widthvec} = [N, NL])$, there exists $\phi_2 \in \mathrm{NN}(\#\mathrm{input} = 1; \mathrm{width} \leq 2N+4; \mathrm{depth} \leq L+2)$ such that $\phi_1(x) = \phi_2(x)$ for any $x \in \mathbb{R}$.*

Now, let us present the detailed proof of Proposition 4.3.

Proof of Proposition 4.3. We divide the proof into two cases: $d = 1$ and $d \geq 2$.

Case 1: $d = 1$.

In this case $K = N^2 L^2$, and we denote $M = N^2 L$. Then we consider the sample set

$$\{(\tfrac{m}{M}, m) : m = 0, 1, \cdots, M-1\} \cup \{(\tfrac{m+1}{M} - \delta, m) : m = 0, 1, \cdots, M-2\} \cup \{(1, M-1), (2, 0)\}.$$

Its cardinality is $2M+1 = N\cdot\big((2NL-1)+1\big)+1$. By Lemma 5.4 with $N_1 = N$ and $N_2 = 2NL-1$, there exist $\phi_1 \in \mathrm{NN}(\text{widthvec} = [2N, 2(2NL-1)+1]) = \mathrm{NN}(\text{widthvec} = [2N, 4NL-1])$ such that

- $\phi_1(\frac{M-1}{M}) = \phi_1(1) = M-1$ and $\phi_1(\frac{m}{M}) = \phi_1(\frac{m+1}{M}-\delta) = m$ for $m = 0,1,\cdots,M-2$;
- ϕ_1 is linear on $[\frac{M-1}{M}, 1]$ and each interval $[\frac{m}{M}, \frac{m+1}{M}-\delta]$ for $m = 0,1,\cdots,M-2$.

Then

$$\phi_1(x) = m, \quad \text{if } x \in [\tfrac{m}{M}, \tfrac{m+1}{M} - \delta\cdot 1_{\{m<M-1\}}], \quad \text{for } m = 0,1,\cdots,M-1. \tag{5.4}$$

Now consider the sample set

$$\{(\tfrac{\ell}{ML}, \ell) : \ell = 0,1,\cdots,L-1\} \cup \{(\tfrac{\ell+1}{ML}-\delta, \ell) : \ell = 0,1,\cdots,L-2\} \cup \{(\tfrac{1}{M}, L-1), (2,0)\}.$$

Its cardinality is $2L+1 = 1\cdot\big((2L-1)+1\big)+1$. By Lemma 5.4 with $N_1 = 1$ and $N_2 = 2L-1$, there exists $\phi_2 \in \mathrm{NN}(\text{widthvec} = [2, 2(2L-1)+1]) = \mathrm{NN}(\text{widthvec} = [2, 4L-1])$ such that

- $\phi_2(\frac{L-1}{ML}) = \phi_2(\frac{1}{M}) = L-1$ and $\phi_2(\frac{\ell}{ML}) = \phi_2(\frac{\ell+1}{ML}-\delta) = \ell$ for $\ell = 0,1,\cdots,L-2$;
- ϕ_2 is linear on $[\frac{L-1}{ML}, \frac{1}{M}]$ and each interval $[\frac{\ell}{ML}, \frac{\ell+1}{ML}-\delta]$ for $\ell = 0,1,\cdots,L-2$.

It follows that, for $m = 0,1,\cdots,M-1$, $\ell = 0,1,\cdots,L-1$,

$$\phi_2\big(x - \tfrac{1}{M}\phi_1(x)\big) = \phi_2(x - \tfrac{m}{M}) = \ell, \quad \text{if } x \in [\tfrac{mL+\ell}{ML}, \tfrac{mL+\ell+1}{ML} - \delta\cdot 1_{\{\ell<L-1\}}]. \tag{5.5}$$

Define

$$\phi(x) \coloneqq \frac{L\phi_1(x) + \phi_2\big(x - \frac{1}{M}\phi_1(x)\big)}{ML}, \quad \text{for any } x \in \mathbb{R}.$$

Notice that each $k \in \{0,1,\cdots,ML-1\} = \{0,1,\cdots,K-1\}$ can be uniquely represented by $k = mL+\ell$ for $m \in \{0,1,\cdots,M-1\}$ and $\ell \in \{0,1,\cdots,L-1\}$. By Equation (5.4) and (5.5), if $x \in [\frac{k}{ML}, \frac{k+1}{ML} - \delta\cdot 1_{\{k<ML-1\}}] = [\frac{k}{K}, \frac{k+1}{K} - \delta\cdot 1_{\{k<K-1\}}]$ and $k = mL+\ell$ for $m \in \{0,1,\cdots,M-1\}$, $\ell \in \{0,1,\cdots,L-1\}$, we have

$$\phi(x) = \frac{L\phi_1(x) + \phi_2\big(x - \frac{1}{M}\phi_1(x)\big)}{ML} = \frac{Lm + \phi_2(x - \frac{m}{M})}{ML} = \frac{Lm+\ell}{ML} = \frac{k}{N^2L^2} = \frac{k}{K}.$$

By Lemma 5.5,

$$\phi_1 \in \mathrm{NN}(\text{widthvec} = [2N, 4NL-1]) \subseteq \mathrm{NN}(\text{width} \le 4N+4;\ \text{depth} \le 2L+2)$$

and

$$\phi_2 \in \mathrm{NN}(\text{widthvec} = [2, 4L-1]) \subseteq \mathrm{NN}(\text{width} \le 8;\ \text{depth} \le 2L+2).$$

Hence, ϕ can be implemented by a ReLU FNN with width $4N+5$ and depth $4L+4$. So we finish the proof.

Case 2: $d \ge 2$.

Now we consider the case when $d \ge 2$. For the sample set

$$\big\{(\tfrac{k}{K}, \tfrac{k}{K}) : k = 0,1,\cdots,K-1\big\} \cup \big\{(\tfrac{k+1}{K}-\delta, \tfrac{k}{K}) : k = 0,1,\cdots,K-2\big\} \cup \big\{(1, \tfrac{K-1}{K}), (2,1)\big\},$$

whose cardinality is $2K+1 = \lfloor N^{1/d}\rfloor\big((2\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor - 1)+1\big)+1$. By Lemma 5.4 with $N_1 = \lfloor N^{1/d}\rfloor$ and $N_2 = 2\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor - 1$, there exists ϕ in

$$\begin{aligned} &\mathrm{NN}(\text{widthvec} = [2\lfloor N^{1/d}\rfloor, 2(2\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor - 1)+1]) \\ \subseteq\ &\mathrm{NN}(\text{widthvec} = [2\lfloor N^{1/d}\rfloor, 4\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor - 1]) \end{aligned}$$

such that

- $\phi(2) = 1$, $\phi(\frac{K-1}{K}) = \phi(1) = \frac{K-1}{K}$, and $\phi(\frac{k}{K}) = \phi(\frac{k+1}{K} - \delta) = \frac{k}{K}$ for $k = 0, 1, \cdots, K-2$;
- ϕ is linear on $[\frac{K-1}{K}, 1]$ and each interval $[\frac{k}{K}, \frac{k+1}{K} - \delta]$ for $k = 0, 1, \cdots, K-2$.

Then

$$\phi(x) = \tfrac{k}{K}, \quad \text{if } x \in [\tfrac{k}{K}, \tfrac{k+1}{K} - \delta\cdot 1_{\{k<K-1\}}], \quad \text{for } k = 0, 1, \cdots, K-1.$$

By Lemma 5.5,

$$\begin{aligned} \phi &\in \mathrm{NN}(\text{widthvec} = [2\lfloor N^{1/d}\rfloor, 4\lfloor N^{1/d}\rfloor\lfloor L^{2/d}\rfloor - 1]) \\ &\subseteq \mathrm{NN}(\text{width} \le 4\lfloor N^{1/d}\rfloor + 4;\ \text{depth} \le 2\lfloor L^{2/d}\rfloor + 2) \\ &\subseteq \mathrm{NN}(\text{width} \le 4N+5;\ \text{depth} \le 4L+4). \end{aligned}$$

This establishes the Proposition. □

5.3 Proof of Proposition 4.4 for point fitting

In this sub-section, we will discuss how to use ReLU FNNs to fit a collection of points in $\mathbb{R}^2$.[8] It is trivial to fit n points via one-hidden-layer ReLU FNNs with $\mathcal{O}(n)$ parameters. However, to prove Proposition 4.4, we need to fit $\mathcal{O}(n)$ points with much less parameters, which is the main difficulty of our proof. Our proof below is mainly based on the "bit extraction" technique and the composition architecture of neural networks.

Let us first introduce a basic lemma based on the "bit extraction" technique, which is in fact Lemma 2.6 of [27].

Lemma 5.6. *For any $N, L \in \mathbb{N}^+$, any $\theta_{m,\ell} \in \{0,1\}$ for $m = 0, 1, \cdots, M-1$, $\ell = 0, 1, \cdots, L-1$, where $M = N^2L$, there exists a ReLU FNN ϕ with width $4N+5$ and depth $3L+4$ such that $\phi(m,\ell) = \sum_{j=0}^{\ell}\theta_{m,j}$, for $m = 0, 1, \cdots, M-1$, $\ell = 0, 1, \cdots, L-1$.*

Next, let us introduce Lemma 5.7, a variant of Lemma 5.6 for a different mapping for the "bit extraction". Its proof is based on Lemma 5.4, 5.5, and 5.6.

Lemma 5.7. *For any $N, L \in \mathbb{N}^+$ and any $\theta_i \in \{0,1\}$ for $i = 0, 1, \cdots, N^2L^2 - 1$, there exists a ReLU FNN ϕ with width $8N+10$ and depth $5L+6$ such that $\phi(i) = \theta_i$, for $i = 0, 1, \cdots, N^2L^2 - 1$.*

---

[8] Fitting a collection of points $\{(x_i, y_i)\}$ in $\mathbb{R}^2$ means that the target ReLU FNN takes the value y_i at the location x_i.

Proof. The case $L = 1$ is simple. We assume $L \geq 2$ below.

Denote $M = N^2 L$, for each $i \in \{0, 1, \cdots, N^2L^2 - 1\}$, there exists a unique representation $i = mL + \ell$ for $m = 0, 1, \cdots, M - 1$ an $L = 0, 1, \cdots, L - 1$. So we define, for $m = 0, 1, \cdots, M - 1$ and $\ell = 0, 1, \cdots, L - 1$,

$$a_{m,\ell} \coloneqq \theta_i, \quad \text{where } i = mL + \ell.$$

Then we set $b_{m,0} = 0$ for $m = 0, 1, \cdots, M - 1$ and $b_{m,\ell} = a_{m,\ell-1}$ for $m = 0, 1, \cdots, M - 1$ and $\ell = 1, \cdots, L - 1$.

By Lemma 5.6, there exist $\phi_1, \phi_2 \in \mathrm{NN}(\text{width} \leq 4N + 5;\ \text{depth} \leq 3L + 4)$ such that

$$\phi_1(m, \ell) = \sum_{j=1}^{\ell} a_{m,j} \quad \text{and} \quad \phi_2(m, \ell) = \sum_{j=1}^{\ell} b_{m,j},$$

for $m = 0, 1, \cdots, M - 1$ and $\ell = 0, 1, \cdots, L - 1$. We consider the sample set

$$\{(mL, m) : m = 0, 1, \cdots, M\} \cup \Big\{\big((m+1)L - 1, m\big) : m = 0, 1, \cdots, M - 1\Big\} \subseteq \mathbb{R}^2.$$

Its cardinality is $2M + 1 = N \cdot \big((2NL - 1) + 1\big) + 1$. By Lemma 5.4 with $N_1 = N$ and $N_2 = 2NL - 1$, there exists $\psi \in \mathrm{NN}(\#\text{input} = 1; \text{widthvec} = [2N, 2(2NL - 1) + 1]) = \mathrm{NN}(\#\text{input} = 1; \text{widthvec} = [2N, 4NL - 1])$ such that

- $\psi(ML) = M$ and $\psi(mL) = \psi\big((m+1)L - 1\big) = m$ for $m = 0, 1, \cdots, M - 1$;
- ψ is linear on each interval $[mL, (m+1)L - 1]$ for $m = 0, 1, \cdots, M - 1$.

It follows that

$$\psi(i) = m \quad \text{where } i = mL + \ell, \quad \text{for } m = 0, 1, \cdots, M - 1 \text{ and } \ell = 0, 1, \cdots, L - 1.$$

Define

$$\phi(x) \coloneqq \phi_1\big(\psi(x), x - L\psi(x)\big) - \phi_2\big(\psi(x), x - L\psi(x)\big), \quad \text{for any } x \in \mathbb{R}.$$

For $i = 0, 1, \cdots, N^2L^2 - 1$, represent $i = mL + \ell$ for $m = 0, 1, \cdots, M - 1$ and $\ell = 0, 1, \cdots, L - 1$. We have

$$\begin{aligned}
\phi(i) &= \phi_1\big(\psi(i), i - L\psi(i)\big) - \phi_2\big(\psi(i), i - L\psi(i)\big) \\
&= \phi_1(m, \ell) - \phi_2(m, \ell) \\
&= \sum_{j=1}^{\ell} a_{m,j} - \sum_{j=1}^{\ell} b_{m,j} = a_{m,\ell} = \theta_i.
\end{aligned}$$

What remaining is to estimate the width and depth of ϕ. Notice that

$$\phi_1, \phi_2 \in \mathrm{NN}(\text{width} \leq 4N + 5;\ \text{depth} \leq 3L + 4).$$

And by Lemma 5.5,

$$\psi \in \mathrm{NN}(\text{widthvec} = [2N, 4NL - 1]) \subseteq \mathrm{NN}(\text{width} \leq 4N + 4;\ \text{depth} \leq 2L + 2).$$

Hence, by the definition of ϕ, ϕ can be implemented by a ReLU FNN with width $8N + 10$ and depth $5L + 6$. $\square$

With Lemma 5.7 in hand, we are now ready to prove Proposition 4.4.

Proof of Proposition 4.4. Denote $J = \lceil 2s\log_2(NL+1)\rceil$. For each $\xi_i \in [0,1]$, there exist $\xi_{i,1}, \xi_{i,2}, \cdots, \xi_{i,J} \in \{0,1\}$ such that

$$\left|\xi_i - \mathrm{Bin}\, 0.\xi_{i,1}\xi_{i,2}\cdots\xi_{i,J}\right| \le 2^{-J}, \quad \text{for } i = 0,1,\cdots,N^2L^2-1.$$

By Lemma 5.7, there exist $\phi_1, \phi_2, \cdots, \phi_J \in \mathrm{NN}(\text{width} \le 8N+10;\ \text{depth} \le 5L+6)$ such that

$$\phi_j(i) = \xi_{i,j}, \quad \text{for } i = 0,1,\cdots,N^2L^2-1,\ j = 1,2,\cdots,J.$$

Define

$$\widetilde{\phi}(x) \coloneqq \sum_{j=1}^{J} 2^{-j}\phi_j(x), \quad \text{for any } x \in \mathbb{R}.$$

It follows that, for $i = 0,1,\cdots,N^2L^2-1$,

$$|\widetilde{\phi}(i) - \xi_i| = \Big|\sum_{j=1}^{J} 2^{-j}\phi_j(i) - \xi_i\Big| = \Big|\sum_{j=1}^{J} 2^{-j}\xi_{i,j} - \xi_i\Big| = \left|\mathrm{Bin}\, 0.\xi_{i,1}\xi_{i,2}\cdots\xi_{i,J} - \xi_i\right| \le 2^{-J}.$$

Notice that

$$2^{-J} = 2^{-\lceil 2s\log_2(NL+1)\rceil} \le 2^{-2s\log_2(NL+1)} = (NL+1)^{-2s} \le N^{-2s}L^{-2s}.$$

Now let us estimate the width and depth of $\widetilde{\phi}$. Recall that

$$\begin{aligned} J = \lceil 2s\log_2(NL+1)\rceil &\le 2s\big(1 + \log_2(NL+1)\big) \le 2s\big(1 + \log_2(2N) + \log_2 L\big) \\ &\le 2s\big(1+\log_2(2N)\big)\big(1+\log_2 L\big) \le 2s\log_2(4N)\log_2(2L), \end{aligned}$$

and $\phi_j \in \mathrm{NN}(\text{width} \le 8N+10;\ \text{depth} \le 5L+6)$. Then $\widetilde{\phi} = \sum_{j=1}^{J} 2^{-j}\phi_j$ can be implemented by a ReLU FNN with width $2s(8N+10)\log_2(4N) + 2 \le 8s(2N+3)\log_2(4N)$ and depth $(5L+6)\log_2(2L)$.

Finally, we define

$$\phi(x) = \min\big\{\max\{0, \widetilde{\phi}(x)\}, 1\big\}, \quad \text{for any } x \in \mathbb{R}.$$

Then $0 \le \phi(x) \le 1$ for any $x \in \mathbb{R}$ and ϕ can be implemented by a ReLU FNN with width $8s(2N+3)\log_2(4N)$ and depth $(5L+6)\log_2(2L) + 2 \le (5L+8)\log_2(2L)$. Notice that

$$\widetilde{\phi}(i) = \sum_{j=1}^{J} 2^{-j}\phi_j(i) = \sum_{j=1}^{J} 2^{-j}\xi_{i,j} \in [0,1], \quad \text{for } i = 0,1,\cdots,N^2L^2-1.$$

It follows that

$$|\phi(i) - \xi_i| = \Big|\min\big\{\max\{0,\widetilde{\phi}(i)\},1\big\} - \xi_i\Big| = |\widetilde{\phi}(i) - \xi_i| \le N^{-2s}L^{-2s}, \quad \text{for } i = 0,1,\cdots,N^2L^2-1.$$

The proof is complete. $\square$

6 Conclusions

This paper has established a nearly optimal approximation rate of ReLU FNNs in terms of both width and depth to approximate smooth functions. It is shown that ReLU FNNs with width $\mathcal{O}(N \ln N)$ and depth $\mathcal{O}(L \ln L)$ can approximate functions in the unit ball of $C^s([0,1]^d)$ with approximation rate $\mathcal{O}(N^{-2s/d} L^{-2s/d})$. Through VC dimension, it is also proved that this approximation rate is asymptotically nearly tight for the closed unit ball of smooth function class $C^s([0,1]^d)$.

We would like to remark that our analysis is for the fully connected feed-forward neural networks with the ReLU activation function. It would be an interesting direction to generalize our results to neural networks with other architectures (e.g., convolutional neural networks and ResNet) and activation functions (e.g., tanh and sigmoid functions). These will be left as future work.

Acknowledgments

The work of J. Lu is supported in part by the National Science Foundation via grants DMS-1415939 and CCF-1934964. Z. Shen is supported by Tan Chin Tuan Centennial Professorship. H. Yang H. Yang was partially supported by the National Science Foundation under award DMS-1945029.

References

[1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

[2] C. Bao, Q. Li, Z. Shen, C. Tai, L. Wu, and X. Xiang. Approximation analysis of convolutional neural networks. 2019.

[3] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

[4] P. Bartlett, V. Maiorov, and R. Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, 10:2159–2173, 1998.

[5] M. Bianchini and F. Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565, Aug 2014.

[6] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

[7] M. Chen, H. Jiang, W. Liao, and T. Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8174–8184. Curran Associates, Inc., 2019.

[8] C. K. Chui, S.-B. Lin, and D.-X. Zhou. Construction of neural networks for realization of localized deep learning. *Frontiers in Applied Mathematics and Statistics*, 4:14, 2018.

[9] G. Cybenko. Approximation by superpositions of a sigmoidal function. *MCSS*, 2:303–314, 1989.

[10] W. E and Q. Wang. Exponential convergence of the deep neural network approximation for analytic functions. *CoRR*, abs/1807.00297, 2018.

[11] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks. *arXiv e-prints*, page arXiv:1905.01208, May 2019.

[12] I. Gühring, G. Kutyniok, and P. Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *arXiv e-prints*, page arXiv:1902.07896, Feb 2019.

[13] N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. In S. Kale and O. Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1064–1068, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

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

[15] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[16] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, June 1994.

[17] Q. Li, T. Lin, and Z. Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of European Mathematical Society*, to appear.

[18] S. Liang and R. Srikant. Why deep neural networks? *CoRR*, abs/1610.04161, 2016.

[19] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6231–6239. Curran Associates, Inc., 2017.

[20] H. Montanelli and H. Yang. Error bounds for deep ReLU networks using the Kolmogorov–Arnold superposition theorem. *Neural Networks*, 129:1–6, 2020.

[21] H. Montanelli, H. Yang, and Q. Du. Deep ReLU networks overcome the curse of dimensionality for bandlimited functions. *arXiv e-prints*, page arXiv:1903.00735, Mar. 2019.

[22] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2924–2932. Curran Associates, Inc., 2014.

[23] R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.

[24] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

[25] A. Sakurai. Tight bounds for the VC-dimension of piecewise polynomial networks. In *Advances in Neural Information Processing Systems*, pages 323–329. Neural information processing systems foundation, 1999.

[26] Z. Shen, H. Yang, and S. Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

[27] Z. Shen, H. Yang, and S. Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

[28] Z. Shen, H. Yang, and S. Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, to appear.

[29] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

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

[31] D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018.

[32] D. Yarotsky and A. Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13005–13015. Curran Associates, Inc., 2020.

[33] D.-X. Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.