

Deep Network Approximation for Smooth Functions

Jianfeng Lu^{*} Zuowei Shen[†] Haizhao Yang[‡] Shijun Zhang[§]

January 9, 2020

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 (current institute) and National University of Singapore (when the project was started) (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, a quantitative characterization of the approximation power of deep feed-forward neural networks (FNNs) with ReLU activation functions is provided in [19]. 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 a 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 [19], 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

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

Theorem 1.1 (Main Theorem). *Give 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*

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

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., $s \geq d$), since $\mathcal{O}(N^{-2s/d}L^{-2s/d}) \leq \mathcal{O}(N^{-2}L^{-2})$, 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 weakened. Note that

$$\mathcal{O}(N \ln N) = \mathcal{O}(\tilde{N}) \iff \mathcal{O}(N) = \mathcal{O}(\tilde{N}/\ln \tilde{N}).$$

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

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

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

^①“nearly optimal” up to a logarithm factor.

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

$$69 \quad \|f - \phi\|_{L^\infty([0,1]^d)} \leq C_3 \|f\|_{C^s([0,1]^d)} N^{-2(s-\rho)/d} L^{-2(s-\rho)/d}.$$

70 Theorem 1.1 and the results in [20, 19] provide non-asymptotic analysis of approx-
71 imation of FNNs, while many others in the literature take asymptotic approaches, i.e.,
72 the network size has to be sufficiently large. The expressiveness of deep neural networks
73 has been studied extensively from many perspectives, e.g., in terms of combinatorics [16],
74 topology [4], Vapnik-Chervonenkis (VC) dimension [3, 18, 9], fat-shattering dimension
75 [12, 1], information theory [17], classical approximation theory [7, 11, 2, 22, 21, 5, 23, 6],
76 etc. In the early works of approximation theory for neural networks, the universal ap-
77 proximation theorem [7, 10, 11] without approximation rates showed that, given any
78 $\varepsilon > 0$, there exists a sufficiently large neural network approximating a target function
79 in a certain function space within the ε -accuracy. For one-hidden-layer neural networks
80 and sufficiently smooth functions, Barron [2] showed an asymptotic approximation rate
81 $\mathcal{O}(\frac{1}{\sqrt{N}})$ in the L^2 -norm, leveraging an idea that is similar to Monte Carlo sampling for
82 high-dimensional integrals. The approximation theoretic results have been extended to
83 deep neural networks in recent years, specifying the approximation rate in terms of the
84 network size for various kinds of functions, e.g., smooth functions [14, 13, 21, 8], piecewise
85 smooth functions [17], band-limited functions [15], continuous functions [22]. However,
86 these approaches are unable to characterize the approximation rate of FNNs in terms of
87 width and depth simultaneously. Instead, Theorem 1.1 and the results in [20, 19] give
88 explicit characterization of the approximation rate of FNNs in terms of width and depth,
89 in the non-asymptotic regime.

90 In Theorem 1.1, we estimate the accuracy in terms of width and depth. This can be
91 translated in terms of the number of parameters of ReLU FNNs versus approximation
92 accuracy. Applying Theorem 1.1, we have following corollary.

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

$$95 \quad \|f - \phi\|_{L^\infty([0,1]^d)} \leq \varepsilon.$$

96 This corollary is followed by setting $N = \mathcal{O}(1)$ and $\varepsilon = \mathcal{O}(L^{-2s/d})$ in Theorem 1.1.
97 Compared with [21], which provides an estimate of number of parameters of $\mathcal{O}(\varepsilon^{-d/s} \ln \frac{1}{\varepsilon})$,
98 our Corollary 1.3 gives a quadratic improvement. Our approximation rate of the ReLU
99 FNNs is nearly optimal. In fact, our Theorem 2.3 gives a nearly optimal, matching lower
100 bound of the approximation rate, based on the nearly optimal estimation of the VC
101 dimension of neural networks for classification in [9].

102 The results obtained in this paper are for $C^s([0,1]^d)$ functions, for Lipschitz func-
103 tions, it is proved in [22] that the optimal rate for ReLU FNNs with width $2d + 10$ and
104 depth $\mathcal{O}(L)$ to approximate Lipschitz continuous functions on $[0,1]^d$ in the L^∞ -norm
105 is $\mathcal{O}(L^{-2/d})$. For the purpose of deep network approximation with arbitrary width and

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

depth, the last three authors demonstrated in [19] 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 [19] can be improved to L^∞ -norm; it will also remove the log factors in the case of C^1 functions in our results here. All these related works are summarized in Table 1.

Table 1: A summary of existing approximation rates of ReLU FNNs for Lipschitz continuous functions and smooth functions. N , L , and W represent the width, the depth, and the number of parameters of a ReLU FNN, respectively.

paper	function class	width	depth	#parameter	accuracy	$L^p([0, 1]^d)$ -norm	tightness	valid for
[21] this paper	polynomial polynomial	$\mathcal{O}(N)$	$\mathcal{O}(\ln \frac{1}{\varepsilon})$ $\mathcal{O}(L)$	$\mathcal{O}(\ln \frac{1}{\varepsilon})$	ε $\mathcal{O}(N^{-L})$	$p = \infty$ $p = \infty$		small $\varepsilon > 0$ any $N, L \in \mathbb{N}^+$
[20] [22] [19]	$\text{Lip}([0, 1]^d)$ $\text{Lip}([0, 1]^d)$ $\text{Lip}([0, 1]^d)$	$\mathcal{O}(N)$ $2d + 10$ $\mathcal{O}(N)$	3 $\mathcal{O}(L)$ $\mathcal{O}(L)$		$\mathcal{O}(N^{-2/d})$ $\mathcal{O}(L^{-2/d})$ $\mathcal{O}(N^{-2/d}L^{-2/d})$	$p \in [1, \infty)$ $p = \infty$ $p = [1, \infty)$	nearly tight in N nearly tight in L nearly tight in N and L	any $N \in \mathbb{N}^+$ large $L \in \mathbb{N}^+$ any $N, L \in \mathbb{N}^+$
[21] this paper this paper	$C^s([0, 1]^d)$ $C^s([0, 1]^d)$ $C^s([0, 1]^d)$	$\mathcal{O}(1)$ $\mathcal{O}(N \ln N)$	$\mathcal{O}(\ln \frac{1}{\varepsilon})$ $\mathcal{O}(\varepsilon^{-d/(2s)} \ln \frac{1}{\varepsilon})$ $\mathcal{O}(L \ln L)$	$\mathcal{O}(\varepsilon^{-d/s} \ln \frac{1}{\varepsilon})$ $\mathcal{O}(\varepsilon^{-d/(2s)} \ln \frac{1}{\varepsilon})$	ε ε $\mathcal{O}(N^{-2s/d}L^{-2s/d})$	$p = \infty$ $p = \infty$ $p = \infty$	not tight in ε nearly tight in ε nearly tight in N and L	any $\varepsilon > 0$ any $\varepsilon > 0$ any $N, L \in \mathbb{N}^+$

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 $\mathcal{B}(\mathbf{x}, r) \subseteq \mathbb{R}^d$ be the closed ball with a center $\mathbf{x} \in \mathbb{R}^d$ and a radius r .
- Similar to “min” and “max”, let $\text{mid}(x_1, x_2, x_3)$ be the middle value of three inputs x_1 , x_2 , and x_3 ^③. For example, $\text{mid}(2, 1, 3) = 2$ and $\text{mid}(3, 2, 3) = 3$.

^③“mid” can be defined via $\text{mid}(x_1, x_2, x_3) = x_1 + x_2 + x_3 - \max(x_1, x_2, x_3) - \min(x_1, x_2, x_3)$, which can be implemented by a ReLU FNN.

- The set difference of two sets A and B is denoted by $A \setminus B := \{x : x \in A, x \notin B\}$.
- For any $x \in \mathbb{R}$, let $\lfloor x \rfloor := \max\{n : n \leq x, n \in \mathbb{Z}\}$ and $\lceil x \rceil := \min\{n : n \geq x, n \in \mathbb{Z}\}$.
- Assume $\mathbf{n} \in \mathbb{N}^n$, then $f(\mathbf{n}) = \mathcal{O}(g(\mathbf{n}))$ means that there exists positive C independent of \mathbf{n} , f , and g such that $f(\mathbf{n}) \leq Cg(\mathbf{n})$ when all entries of \mathbf{n} go to $+\infty$.
- The modulus of continuity of a continuous function $f \in C([0, 1]^d)$ is defined as

$$\omega_f(r) := \sup \{|f(\mathbf{x}) - f(\mathbf{y})| : \|\mathbf{x} - \mathbf{y}\|_2 \leq r, \mathbf{x}, \mathbf{y} \in [0, 1]^d\}, \quad \text{for any } r \geq 0.$$

- A d -dimensional multi-index is a d -tuple $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_d]^T \in \mathbb{N}^d$. Several related notations are listed below.

$$- \|\boldsymbol{\alpha}\|_1 = |\alpha_1| + |\alpha_2| + \dots + |\alpha_d|;$$

$$- \mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}, \text{ where } \mathbf{x} = [x_1, x_2, \dots, x_d]^T;$$

$$- \boldsymbol{\alpha}! = \alpha_1! \alpha_2! \dots \alpha_d!;$$

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

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

$$\Omega(K, \delta, d) := \bigcup_{i=1}^d \left\{ \mathbf{x} = [x_1, x_2, \dots, x_d]^T : x_i \in \bigcup_{k=1}^{K-1} \left(\frac{k}{K} - \delta, \frac{k}{K} \right) \right\}. \quad (2.1)$$

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

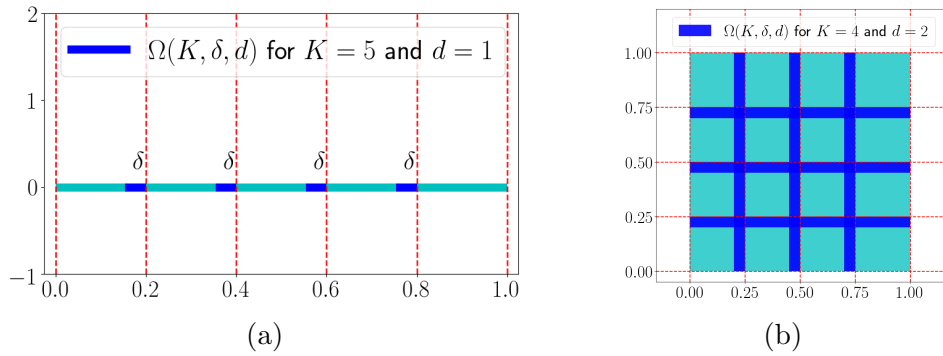


Figure 1: Two examples of trifling regions. (a) $K = 5, d = 1$. (b) $K = 4, d = 2$.

147

- We will use NN as a ReLU neural network for short and use Python-type notations to specify a class of NNs, e.g., $\text{NN}(c_1; c_2; \dots; c_m)$ is a set of ReLU FNNs satisfying m conditions given by $\{c_i\}_{1 \leq i \leq m}$, 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 $\phi \in \text{NN}(\text{\#input} = 2; \text{widthvec} = [100, 100])$, then ϕ satisfies

^④The trifling region here is similar to the “don’t care” region in our previous paper [19].

- 156 – ϕ maps from \mathbb{R}^2 to \mathbb{R} .
- 157 – ϕ has two hidden layers and the number of nodes in each hidden layer is 100.
- 158 • The expression “a network with width N and depth L ” means
 - 159 – The maximum width of all hidden layers is no more than N .
 - 160 – The number of hidden layers is no more than L .
- 161 • For $x \in [0, 1]$, suppose its binary representation is $x = \sum_{\ell=1}^{\infty} x_{\ell} 2^{-\ell}$ with $x_{\ell} \in \{0, 1\}$,
 - 162 we introduce a special notation $\text{Bin}0.x_1x_2\cdots x_L$ to denote the L -term binary repre-
 - 163 sentation of x , i.e., $\sum_{\ell=1}^L x_{\ell} 2^{-\ell}$.

164 2.2 Proof of Theorem 1.1

165 The introduction of the trifling region $\Omega(K, \delta, d)$ is due to the fact that ReLU FNNs
 166 cannot approximate a step function uniformly well (as ReLU activation function is con-
 167 tinuous), which is also the reason for the main difficulty of obtaining approximation
 168 rates in the $L^{\infty}([0, 1]^d)$ -norm in our previous papers [20, 19]. The trifling region is a key
 169 technique to simplify the proofs of theories in [20, 19] as well as the proof of Theorem 1.1.
 170 First, we present Theorem 2.1 showing that, as long as good uniform approximation by a
 171 ReLU FNN can be obtained outside the trifling region, the uniform approximation error
 172 can also be well controlled inside the trifling region when the network size is increased.
 173 Second, as a simplified version of Theorem 1.1 ignoring the approximation error in the
 174 trifling region $\Omega(K, \delta, d)$, Theorem 2.2 shows the existence of a ReLU FNN approximat-
 175 ing a target smooth function uniformly well outside the trifling region. Finally, Theorem
 176 2.1 and 2.2 immediately lead to Theorem 1.1. Theorem 2.2 can be applied to improve
 177 the theories in [20, 19] to obtain approximation rates in the $L^{\infty}([0, 1]^d)$ -norm.

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

$$180 \quad |f(\mathbf{x}) - \tilde{\phi}(\mathbf{x})| \leq \varepsilon, \quad \text{for any } \mathbf{x} \in [0, 1]^d \setminus \Omega(K, \delta, d),$$

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

$$182 \quad |f(\mathbf{x}) - \phi(\mathbf{x})| \leq \varepsilon + d \cdot \omega_f(\delta), \quad \text{for any } \mathbf{x} \in [0, 1]^d.$$

183 **Theorem 2.2.** *Assume that $f \in C^s([0, 1]^d)$ satisfies $\|\partial^{\alpha} f\|_{L^{\infty}([0, 1]^d)} \leq 1$ for any $\alpha \in \mathbb{N}^d$
 184 with $\|\alpha\|_1 \leq s$. For any $N, L \in \mathbb{N}^+$, there exists a ReLU FNN ϕ with width $21s^{d+1}d(N +$
 185 $2) \log_2(4N)$ and depth $18s^2(L + 2) \log_2(2L)$ such that*

$$186 \quad \|f - \phi\|_{L^{\infty}([0, 1]^d \setminus \Omega(K, \delta, d))} \leq 84(s + 1)^d 8^s N^{-2s/d} L^{-2s/d},$$

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

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

190 *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 $\delta \in$
 191 $(0, \frac{1}{K})$ such that $\omega_f(\delta) \leq N^{-2s/d} L^{-2s/d}$. By Theorem 2.2, there exists a ReLU FNN $\tilde{\phi}$
 192 with width $21s^{d+1}d(N+2)\log_2(4N)$ and depth $18s^2(L+2)\log_s(2L)$ such that

$$193 \quad \|\bar{f} - \tilde{\phi}\|_{L^\infty([0,1]^d \setminus \Omega(K,\delta,d))} \leq 84(s+1)^d 8^s N^{-2s/d} L^{-2s/d}.$$

194 By Theorem 2.1, there exists a ReLU FNN $\bar{\phi}$ with width $3^d(21s^{d+1}d(N+2)\log_2(4N)+3) \leq$
 195 $22s^{d+1}3^d d(N+2)\log_2(4N)$ and depth $18s^2(L+2)\log_s(2L) + 2d$ such that

$$196 \quad \|\bar{f} - \bar{\phi}\|_{L^\infty([0,1]^d)} \leq 84(s+1)^d 8^s N^{-2s/d} L^{-2s/d} + d \cdot \omega_f(\delta) \leq 85(s+1)^d 8^s N^{-2s/d} L^{-2s/d}.$$

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

$$198 \quad \|f - \phi\|_{L^\infty([0,1]^d)} = \|f\|_{C^s([0,1]^d)} \|\bar{f} - \bar{\phi}\|_{L^\infty([0,1]^d)} \leq 85(s+1)^d 8^s \|f\|_{C^s([0,1]^d)} N^{-2s/d} L^{-2s/d},$$

199 which finishes the proof. \square

200 2.3 Optimality of Theorem 1.1

201 In this section, we will show that the approximation rate in Theorem 1.1 is asymp-
 202 totically nearly tight. In particular, the approximation rate $\mathcal{O}(N^{-(2s/d+\rho)} L^{-(2s/d+\rho)})$ for
 203 any $\rho > 0$ is not attainable, if we use ReLU FNNs with width $\mathcal{O}(N \ln N)$ and depth
 204 $\mathcal{O}(L \ln L)$ to approximate functions in $\mathcal{F}_{s,d}$, where $\mathcal{F}_{s,d}$ is the unit ball of $C^s([0,1]^d)$
 205 defined via

$$206 \quad \mathcal{F}_{s,d} := \{f \in C^s([0,1]^d) : \|\partial^\alpha f\|_{L^\infty([0,1]^d)} \leq 1, \text{ for all } \alpha \in \mathbb{N}^d \text{ with } \|\alpha\|_1 \leq s\}.$$

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

$$209 \quad \inf_{\phi \in \text{NN}(\text{width} \leq C_1 N \ln N; \text{depth} \leq C_2 L \ln L)} \|\phi - f\|_{L^\infty([0,1]^d)} \geq C_3 N^{-(2s/d+\rho)} L^{-(2s/d+\rho)}.$$

210 Theorem 2.3 will be proved by contradiction. Assuming Theorem 2.3 is not true, we
 211 have the following claim, which can be disproved using the VC dimension upper bound
 212 in [9].

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

$$215 \quad \inf_{\phi \in \text{NN}(\text{width} \leq C_1 N \ln N; \text{depth} \leq C_2 L \ln L)} \|\phi - f\|_{L^\infty([0,1]^d)} \leq C_3 N^{-(2s/d+\rho)} L^{-(2s/d+\rho)},$$

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

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

218 *Disproof of Claim 2.4.* Recall that the VC dimension of a class of functions is defined
 219 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 $\text{VCDim}(\mathcal{F})$. Set $\tilde{N} = C_1 N \ln N$ and $\tilde{L} = C_2 L \ln L$. Then by [9], there exists $C_4 > 0$ such that

$$\begin{aligned} & \text{VCDim}(\text{NN}(\# \text{input} = d; \text{width} \leq \tilde{N}; \text{depth} \leq \tilde{L})) \\ & \leq C_4(\tilde{N}\tilde{L} + d + 2)(\tilde{N} + 1)\tilde{L} \ln((\tilde{N}\tilde{L} + d + 2)(\tilde{N} + 1)) := b_u(N, L), \end{aligned} \quad (2.2)$$

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

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

$$\text{VCDim}(\text{NN}(\# \text{input} = d; \text{width} \leq \tilde{N}; \text{depth} \leq \tilde{L})),$$

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

More precisely, we will construct $\{f_\beta : \beta \in \mathcal{B}\} \subseteq \mathcal{F}_{s,d}$, which can shatter $b_\ell(N, L) = K^d$ points, where \mathcal{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 $\{\phi_\beta : \beta \in \mathcal{B}\}$ with width bounded by \tilde{N} and depth bounded by \tilde{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 $\{f_\beta : \beta \in \mathcal{B}\} \subseteq \mathcal{F}_{s,d}$ that scatters $b_\ell(N, L)$ points.

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

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

$$Q_\theta := \{\mathbf{x} = [x_1, x_2, \dots, x_d]^T \in [0, 1]^d : x_i \in [\frac{\theta_i - 1}{K}, \frac{\theta_i}{K}], \ i = 1, 2, \dots, d\},$$

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

$$\mathcal{B} := \{\beta : \beta \text{ is a map from } \{1, 2, \dots, K\}^d \text{ to } \{-1, 1\}\}.$$

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

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

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

$$\text{supp}(g_\theta) \subseteq \frac{2}{3}Q_\theta, \quad \text{for any } \theta \in \{1, 2, \dots, K\}^d,$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\text{VCDim}(\{\phi_\beta : \beta \in \mathcal{B}\}) \geq K^d = b_\ell(N, L), \quad (2.3)$$

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

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

implying that

$$\begin{aligned} [(NL)^{2/d+\rho/(2\alpha)}]^d &\leq C_4(\tilde{L}\tilde{N} + d + 2)(\tilde{N} + 1)\tilde{L} \ln((\tilde{L}\tilde{N} + d + 2)(\tilde{N} + 1)) \\ &= \mathcal{O}(\tilde{N}^2 \tilde{L}^2 \ln(\tilde{N}^2 \tilde{L})) \\ &= \mathcal{O}\left((C_1 N \ln N)^2 (C_2 L \ln L)^2 \ln((C_1 N \ln N)^2 C_2 L \ln L)\right), \end{aligned}$$

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 $\text{mid}(g(x+\delta), g(x), g(x-\delta))$ can approximate f well on the whole domain, where $\text{mid}(\cdot, \cdot, \cdot)$ 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*

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

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,

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

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 $x_1, x_2, x_3 \in \mathbb{R}$,

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

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

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

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

The next lemma shows a simple but useful property of the $\text{mid}(x_1, x_2, x_3)$ 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 $\text{mid}(x_1, x_2, x_3) \in \mathcal{B}(y, \varepsilon)$.*

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

1. If $x_3 < x_1$, then $\text{mid}(x_1, x_2, x_3) = x_1 \in \mathcal{B}(y, \varepsilon)$.

309 2. If $x_1 \leq x_3 \leq x_2$, then $\text{mid}(x_1, x_2, x_3) = x_3 \in \mathcal{B}(y, \varepsilon)$ since $y - \varepsilon \leq x_1 \leq x_3 \leq x_2 \leq y + \varepsilon$.
 310 3. If $x_2 < x_3$, then $\text{mid}(x_1, x_2, x_3) = x_2 \in \mathcal{B}(y, \varepsilon)$.
 311 So we finish the proof. \square

312 Next, given a function g approximating f well on $[0, 1]$ except for a trifling region,
 313 Lemma 3.3 below shows how to use the $\text{mid}(x_1, x_2, x_3)$ function to construct a new
 314 function ϕ uniformly approximating f well on $[0, 1]$, leveraging the useful property of
 315 $\text{mid}(x_1, x_2, x_3)$ in Lemma 3.2.

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

$$318 \quad |f(x) - g(x)| \leq \varepsilon, \quad \text{for any } x \in [0, 1] \setminus \Omega(K, \delta, 1).$$

319 Then

$$320 \quad |\phi(x) - f(x)| \leq \varepsilon + \omega_f(\delta), \quad \text{for any } x \in [0, 1],$$

321 where

$$322 \quad \phi(x) := \text{mid}(g(x - \delta), g(x), g(x + \delta)), \quad \text{for any } x \in \mathbb{R}.$$

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

$$324 \quad Q_k = Q_{k,1} \cup Q_{k,2} \cup Q_{k,3} \cup Q_{k,4},$$

325 where $Q_{k,1} = [\frac{k}{K}, \frac{k}{K} + \delta]$, $Q_{k,1} = [\frac{k}{K} + \delta, \frac{k+1}{K} - 2\delta]$, $Q_{k,1} = [\frac{k+1}{K} - 2\delta, \frac{k+1}{K} - \delta]$, and $Q_{k,1} =$
 326 $[\frac{k+1}{K} - \delta, \frac{k+1}{K}]$.

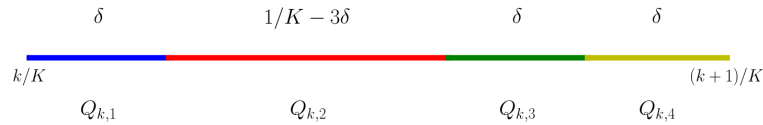


Figure 2: Illustrations of $Q_{k,i}$ for $i = 1, 2, 3, 4$.

327 Notice that $Q_{k+1,4} \subseteq [0, 1] \setminus \Omega(K, \delta, 1)$ and $Q_{k,i} \subseteq [0, 1] \setminus \Omega(K, \delta, 1)$ for $k = 0, 1, \dots, k -$
 328 1 , $i = 1, 2, 3$. For any $k \in \{0, 1, \dots, K-1\}$, we consider the following four cases.

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

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

$$332 \quad g(x) \in \mathcal{B}(f(x), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta))$$

333 and

$$334 \quad g(x + \delta) \in \mathcal{B}(f(x + \delta), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

335 By Lemma 3.2, we get

$$336 \quad \text{mid}(g(x - \delta), g(x), g(x + \delta)) \in \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

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

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

$$339 \quad g(x - \delta), g(x), g(x + \delta) \in \mathcal{B}(f(x), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)),$$

340 which implies by Lemma 3.2 that

$$341 \quad \text{mid}(g(x - \delta), g(x), g(x + \delta)) \in \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

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

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

$$345 \quad g(x) \in \mathcal{B}(f(x), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta))$$

346 and

$$347 \quad g(x - \delta) \in \mathcal{B}(f(x - \delta), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

348 By Lemma 3.2, we get

$$349 \quad \text{mid}(g(x - \delta), g(x), g(x + \delta)) \in \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

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

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

- 352 • If $k \in \{0, 1, \dots, K - 2\}$, then $x - \delta \in Q_{k,3} \in [0, 1] \setminus \Omega(K, \delta, 1)$ and $x + \delta \in Q_{k+1,1} \subseteq$
353 $[0, 1] \setminus \Omega(K, \delta, 1)$. It follows that

$$354 \quad g(x - \delta) \in \mathcal{B}(f(x - \delta), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta))$$

355 and

$$356 \quad g(x + \delta) \in \mathcal{B}(f(x + \delta), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

357 By Lemma 3.2, we get

$$358 \quad \text{mid}(g(x - \delta), g(x), g(x + \delta)) \in \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

- 359 • If $k = K - 1$, then $x \in Q_{k+1,4} \subseteq [0, 1] \setminus \Omega(K, \delta, 1)$ and $x - \delta \in Q_{k,3} \subseteq [0, 1] \setminus \Omega(K, \delta, 1)$.
360 It follows that

$$361 \quad g(x) \in \mathcal{B}(f(x), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta))$$

362 and

$$363 \quad g(x - \delta) \in \mathcal{B}(f(x - \delta), \varepsilon) \subseteq \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

364 By Lemma 3.2, we get

$$365 \quad \text{mid}(g(x - \delta), g(x), g(x + \delta)) \in \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)).$$

366 Since $[0, 1] = \cup_{k=0}^{K-1} \left(\cup_{i=1}^4 Q(k, i) \right)$, we have

$$367 \quad \text{mid}(g(x - \delta), g(x), g(x + \delta)) \in \mathcal{B}(f(x), \varepsilon + \omega_f(\delta)), \quad \text{for any } x \in [0, 1].$$

368 Notice that $\phi(x) = \text{mid}(g(x - \delta), g(x), g(x + \delta))$, it holds that

$$369 \quad |\phi(x) - f(x)| \leq \varepsilon + \omega_f(\delta), \quad \text{for any } x \in [0, 1].$$

370 So we finish the proof. □

371 The next lemma below is an analog of Lemma 3.3.

372 **Lemma 3.4.** *Given $\varepsilon > 0$, $K \in \mathbb{N}^+$, and $\delta > 0$ with $\delta \leq \frac{1}{3K}$, assume $f, g \in C([0, 1]^d)$ with*

$$373 \quad |f(\mathbf{x}) - g(\mathbf{x})| \leq \varepsilon, \quad \text{for any } \mathbf{x} \in [0, 1]^d \setminus \Omega(K, \delta, d).$$

374 Let $\phi_0 = g$ and $\{\mathbf{e}_i\}_{i=1}^d$ be the standard basis in \mathbb{R}^d . By induction, we define

$$375 \quad \phi_{i+1}(\mathbf{x}) := \text{mid}(\phi_i(\mathbf{x} - \delta \mathbf{e}_{i+1}), \phi_i(\mathbf{x}), \phi_i(\mathbf{x} + \delta \mathbf{e}_{i+1})), \quad \text{for } i = 0, 1, \dots, d-1.$$

376 Let $\phi := \phi_d$, then

$$377 \quad |f(\mathbf{x}) - \phi(\mathbf{x})| \leq \varepsilon + d \cdot \omega_f(\delta), \quad \text{for any } \mathbf{x} \in [0, 1]^d.$$

378 *Proof.* For $\ell = 0, 1, \dots, d$, we denote

$$379 \quad E_\ell := \{\mathbf{x} = [x_1, x_2, \dots, x_d]^T : x_i \in [0, 1] \text{ for } i \leq \ell, x_j \in [0, 1] \setminus \Omega(K, \delta, 1) \text{ for } j > \ell\}.$$

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

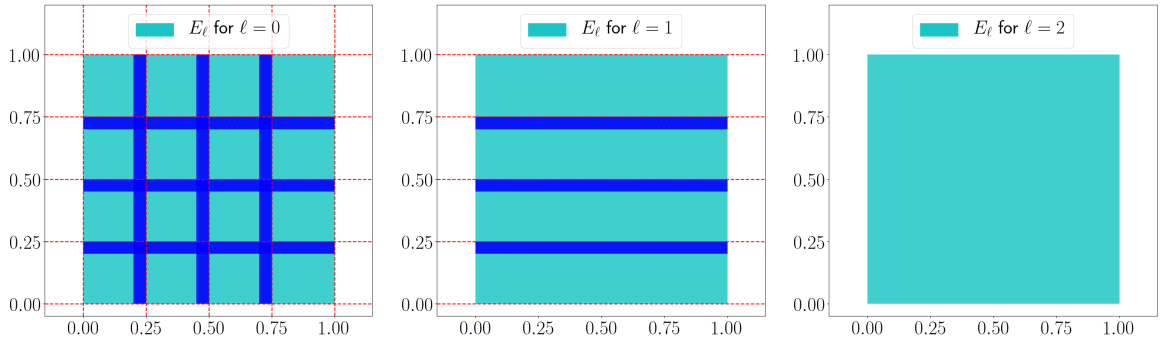


Figure 3: Illustrations of E_ℓ for $\ell = 0, 1, 2$ and $K = 4$.

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

$$383 \quad \phi_\ell(\mathbf{x}) \in \mathcal{B}(f(\mathbf{x}), \varepsilon + \ell \cdot \omega_f(\delta)), \quad \text{for any } \mathbf{x} \in E_\ell. \quad (3.1)$$

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

$$387 \quad \phi_0(\mathbf{x}) \in \mathcal{B}(f(\mathbf{x}), \varepsilon), \quad \text{for any } \mathbf{x} \in E_0.$$

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

389 Now assume Equation (3.1) is true for $\ell = i$. We will prove that it also holds for
 390 $\ell = i + 1$. For any $\mathbf{x}^{[i]} := [x_1, \dots, x_i, x_{i+2}, \dots, x_d]^T \in \mathbb{R}^{d-1}$, we set

$$391 \quad \psi_{\mathbf{x}^{[i]}}(t) := \phi_i(x_1, \dots, x_i, t, x_{i+2}, \dots, x_d), \quad \text{for any } t \in \mathbb{R},$$

392 and

$$393 \quad f_{\mathbf{x}^{[i]}}(t) := f(x_1, \dots, x_i, t, x_{i+2}, \dots, x_d), \quad \text{for any } t \in \mathbb{R}.$$

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

$$396 \quad \phi_i(x_1, \dots, x_i, t, x_{i+2}, \dots, x_d) \in \mathcal{B}(f(x_1, \dots, x_i, t, x_{i+2}, \dots, x_d), \varepsilon + i \cdot \omega_f(\delta)),$$

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

$$398 \quad \psi_{\mathbf{x}^{[i]}}(t) \in \mathcal{B}(f_{\mathbf{x}^{[i]}}(t), \varepsilon + i \cdot \omega_f(\delta)), \quad \text{for any } t \in [0, 1] \setminus \Omega(K, \delta, 1).$$

399 Then by Lemma 3.3, we get

$$400 \quad \text{mid}(\psi_{\mathbf{x}^{[i]}}(t - \delta), \psi_{\mathbf{x}^{[i]}}(t), \psi_{\mathbf{x}^{[i]}}(t + \delta)) \in \mathcal{B}(f_{\mathbf{x}^{[i]}}(t), \varepsilon + (i + 1)\omega_f(\delta)), \quad \text{for any } t \in [0, 1].$$

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

$$402 \quad \begin{aligned} & \text{mid}(\phi_i(x_1, \dots, x_i, x_{i+1} - \delta, x_{i+2}, \dots, x_d), \phi_i(x_1, \dots, x_i, x_{i+1}, x_{i+2}, \dots, x_d), \\ & \quad \phi_i(x_1, \dots, x_i, x_{i+1} + \delta, x_{i+2}, \dots, x_d)) \\ & \in \mathcal{B}(f(x_1, \dots, x_d), \varepsilon + (i + 1)\omega_f(\delta)). \end{aligned}$$

403 Since $x_1, \dots, x_i \in [0, 1]$ and $x_{i+2}, \dots, x_d \in [0, 1] \setminus \Omega(K, \delta, 1)$ are arbitrary, then for any $\mathbf{x} \in$
 404 E_{i+1} ,

$$405 \quad \text{mid}(\phi_i(\mathbf{x} - \delta \mathbf{e}_{i+1}), \phi_i(\mathbf{x}), \phi_i(\mathbf{x} + \delta \mathbf{e}_{i+1})) \in \mathcal{B}(f(\mathbf{x}), \varepsilon + (i + 1)\omega_f(\delta)),$$

406 which implies

$$407 \quad \phi_{i+1}(\mathbf{x}) \in \mathcal{B}(f(\mathbf{x}), \varepsilon + (i + 1)\omega_f(\delta)), \quad \text{for any } \mathbf{x} \in E_{i+1}.$$

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

409 By the principle of induction, we have

$$410 \quad \phi(\mathbf{x}) := \phi_d(\mathbf{x}) \in \mathcal{B}(f(\mathbf{x}), \varepsilon + d \cdot \omega_f(\delta)), \quad \text{for any } \mathbf{x} \in E_d = [0, 1]^d.$$

411 Therefore,

$$412 \quad |\phi(\mathbf{x}) - f(\mathbf{x})| \leq \varepsilon + d \cdot \omega_f(\delta), \quad \text{for any } \mathbf{x} \in [0, 1]^d,$$

413 which means we finish the proof. □

414 Now we are ready to prove Theorem 2.1.

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

$$416 \quad \phi_{i+1}(\mathbf{x}) := \text{mid}(\phi_i(\mathbf{x} - \delta \mathbf{e}_{i+1}), \phi_i(\mathbf{x}), \phi_i(\mathbf{x} + \delta \mathbf{e}_{i+1})), \quad \text{for } i = 0, 1, \dots, d-1.$$

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

$$421 \quad |f(\mathbf{x}) - \phi(\mathbf{x})| \leq \varepsilon + d \cdot \omega_f(\delta), \quad \text{for any } \mathbf{x} \in [0, 1]^d.$$

422 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 $\boldsymbol{\theta} \in \{0, 1, \dots, K-1\}^d$ and $\mathbf{x} \in \{\mathbf{z} : \frac{\theta_i}{K} \leq z_i \leq \frac{\theta_i+1}{K}, i = 1, 2, \dots, d\}$, there exists $\xi_{\mathbf{x}} \in (0, 1)$ such that

$$f(\mathbf{x}) = \sum_{\|\boldsymbol{\alpha}\|_1 \leq s-1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\theta}/K)}{\boldsymbol{\alpha}!} \mathbf{h}^{\boldsymbol{\alpha}} + \sum_{\|\boldsymbol{\alpha}\|_1 = s} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\theta}/K + \xi_{\mathbf{x}} \mathbf{h})}{\boldsymbol{\alpha}!} \mathbf{h}^{\boldsymbol{\alpha}} := \mathcal{T}_1 + \mathcal{T}_2, \quad \textcircled{6}$$

where $\mathbf{h}(\mathbf{x}) = \mathbf{x} - \frac{\boldsymbol{\theta}}{K}$. It is clear that the magnitude of \mathcal{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 $\phi \in \text{NN}(\text{width} \leq \mathcal{O}(N); \text{depth} \leq \mathcal{O}(L))$ to approximate

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

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

- Construct a ReLU FNN $P_{\boldsymbol{\alpha}}$ to approximate the polynomial $\mathbf{h}^{\boldsymbol{\alpha}}$ for each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \leq s-1$.
- 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 \mathbf{x} to $\boldsymbol{\theta}/K$ if $x_i \in [\theta_i/K, (\theta_i+1)/K)$ for $i = 1, 2, \dots, d$ and $\boldsymbol{\theta} \in \{0, 1, \dots, K-1\}^d$.
- 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 $\|\boldsymbol{\alpha}\|_1 \leq s-1$.

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 $\mathbf{h}^{\boldsymbol{\alpha}}$ according to Proposition 4.1 below, a general proposition for approximating multivariable polynomials.

Proposition 4.1. *Assume $P(\mathbf{x}) = \mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d}$ for $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 = k \geq 2$. 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*

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

^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.

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 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*

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

Second, we construct a step function ψ mapping $\mathbf{x} \in \{\mathbf{z} : \frac{\theta_i}{K} \leq z_i < \frac{\theta_{i+1}}{K}, i = 1, 2, \dots, d\}$ to $\frac{\theta}{K}$. We only need to approximate one-dimensional step functions, because in the multidimensional case we can simply set $\psi(\mathbf{x}) = [\psi(x_1), \psi(x_2), \dots, \psi(x_d)]^T$, 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 $\delta \leq \frac{1}{3K}$, there exists a one-dimensional ReLU FNN ϕ with width $4N + 5$ and depth $4L + 4$ such that*

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

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

$$|\phi_\alpha(\frac{\theta}{K}) - \partial^\alpha f(\frac{\theta}{K})| \leq \mathcal{O}(N^{-2s/d} L^{-2s/d}), \quad \text{for any } \theta \in \{0, 1, \dots, K-1\}^d.$$

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 $i = 0, 1, \dots, N^2 L^2 - 1$, there exists a ReLU FNN ϕ with width $8s(2N + 3) \log_2(4N)$ and depth $(5L + 8) \log_2(2L)$ such that*

$$1. |\phi(i) - \xi_i| \leq N^{-2s} L^{-2s}, \text{ for } i = 0, 1, \dots, N^2 L^2 - 1;$$

$$2. 0 \leq \phi(t) \leq 1, \text{ for any } t \in \mathbb{R}.$$

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 $\phi(\mathbf{x})$ is based on a sequence of sub-networks listed before $\phi(\mathbf{x})$. Recall that $\mathbf{h}(\mathbf{x}) = \mathbf{x} - \psi(\mathbf{x})$.

Target function	ReLU FNN	Width	Depth	Approximation error
Step function	$\psi(\mathbf{x})$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	No error out of $\Omega(K, \delta, d)$
$x_1 x_2$	$\tilde{\phi}(x_1, x_2)$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{E}_1 = \mathcal{O}((N+1)^{-2s(L+1)})$
\mathbf{h}^α	$P_\alpha(\mathbf{h})$	$\mathcal{O}(N)$	$\mathcal{O}(L)$	$\mathcal{E}_2 = \mathcal{O}((N+1)^{-7s(L+1)})$
$\partial^\alpha f(\psi(\mathbf{x}))$	$\phi_\alpha(\psi(\mathbf{x}))$	$\mathcal{O}(N \ln N)$	$\mathcal{O}(L \ln L)$	$\mathcal{E}_3 = \mathcal{O}(N^{-2s} L^{-2s})$
$\sum_{\ \alpha\ \leq s-1} \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha$	$\sum_{\ \alpha\ \leq s-1} \tilde{\phi}\left(\frac{\phi_\alpha(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right)$	$\mathcal{O}(N \ln N)$	$\mathcal{O}(L \ln L)$	$\mathcal{O}(\mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3)$
$f(\mathbf{x})$	$\phi(\mathbf{x}) := \sum_{\ \alpha\ \leq s-1} \tilde{\phi}\left(\frac{\phi_\alpha(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{x} - \psi(\mathbf{x}))\right)$	$\mathcal{O}(N \ln N)$	$\mathcal{O}(L \ln L)$	$\mathcal{O}(\ \mathbf{h}\ _2^{-s} + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3) \leq \mathcal{O}(K^{-s}) = \mathcal{O}(N^{-2s/d} L^{-2s/d})$

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_θ for $\theta \in \{0, 1, \dots, K-1\}^d$. In particular, for each $\theta = [\theta_1, \theta_2, \dots, \theta_d]^T \in \{0, 1, \dots, K-1\}^d$, we define

$$Q_\theta = \left\{ \mathbf{x} = [x_1, x_2, \dots, x_d]^T : x_i \in \left[\frac{\theta_i}{K}, \frac{\theta_i+1}{K} - \delta \cdot 1_{\{\theta_i < K-1\}} \right], i = 1, 2, \dots, d \right\}.$$

It is clear that $[0, 1]^d = \Omega(K, \delta, d) \cup \left(\cup_{\theta \in \{0, 1, \dots, K-1\}^d} Q_\theta \right)$. See Figure 4 for the illustration of Q_θ .

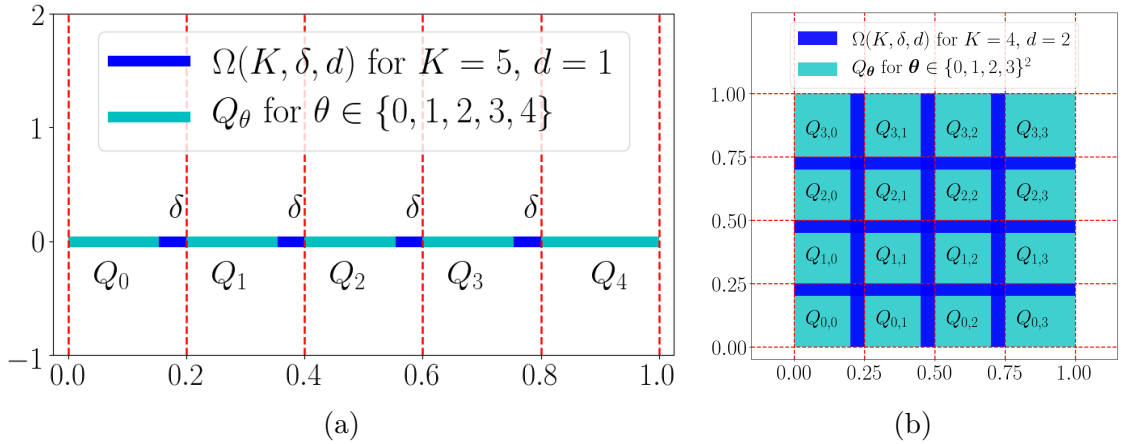


Figure 4: Illustrations of Q_θ for $\theta \in \{0, 1, \dots, K-1\}^d$. (a) $K=5, d=1$. (b) $K=4, d=2$.

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

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

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

512 Define

513
$$\boldsymbol{\psi}(\mathbf{x}) := [\psi(x_1), \psi(x_2), \dots, \psi(x_d)]^T, \quad \text{for any } \mathbf{x} \in [0, 1]^d,$$

514 then

515
$$\boldsymbol{\psi}(\mathbf{x}) = \frac{\boldsymbol{\theta}}{K} \quad \text{if } \mathbf{x} \in Q_{\boldsymbol{\theta}}, \quad \text{for } \boldsymbol{\theta} \in \{0, 1, \dots, K-1\}^d.$$

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

518
$$f(\mathbf{x}) = \sum_{\|\boldsymbol{\alpha}\|_1 \leq s-1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\mathbf{x}))}{\boldsymbol{\alpha}!} \mathbf{h}^{\boldsymbol{\alpha}} + \sum_{\|\boldsymbol{\alpha}\|_1 = s} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\mathbf{x}) + \xi_{\mathbf{x}} \mathbf{h})}{\boldsymbol{\alpha}!} \mathbf{h}^{\boldsymbol{\alpha}}, \quad \text{where } \mathbf{h} = \mathbf{x} - \boldsymbol{\psi}(\mathbf{x}).$$

519 **Step 2:** The construction of the target ReLU FNN.

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

521
$$|\widetilde{\phi}(x_1, x_2) - x_1 x_2| \leq 216(N+1)^{-2s(L+1)} := \mathcal{E}_1, \quad \text{for any } x_1, x_2 \in [-3, 3]. \quad (4.1)$$

522 If $2 \leq \|\boldsymbol{\alpha}\|_1 \leq s-1$, by Proposition 4.1, there exist ReLU FNNs $P_{\boldsymbol{\alpha}}$ with width
523 $9(N+1) + \|\boldsymbol{\alpha}\|_1 - 2 \leq 9N + s + 6$ and depth $7s(\|\boldsymbol{\alpha}\|_1 - 1)(L+1) \leq 7s^2(L+1)$ such that

524
$$|P_{\boldsymbol{\alpha}}(\mathbf{x}) - \mathbf{x}^{\boldsymbol{\alpha}}| \leq 9(\|\boldsymbol{\alpha}\|_1 - 1)(N+1)^{-7s(L+1)} \leq 9s(N+1)^{-7s(L+1)}, \quad \text{for any } \mathbf{x} \in [0, 1]^d.$$

525 And it is trivial to construct ReLU FNNs $P_{\boldsymbol{\alpha}}$ to approximate $\mathbf{x}^{\boldsymbol{\alpha}}$ when $\|\boldsymbol{\alpha}\|_1 \leq 1$. Hence,
526 for each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \leq s-1$, there always exists $P_{\boldsymbol{\alpha}} \in \text{NN}(\text{width} \leq 9N + s + 6; \text{depth} \leq$
527 $7s^2(L+1))$ such that

528
$$|P_{\boldsymbol{\alpha}}(\mathbf{x}) - \mathbf{x}^{\boldsymbol{\alpha}}| \leq 9s(N+1)^{-7s(L+1)} := \mathcal{E}_2, \quad \text{for any } \mathbf{x} \in [0, 1]^d. \quad (4.2)$$

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

530
$$\boldsymbol{\eta}(i) = [\eta_1, \eta_2, \dots, \eta_d]^T \in \{0, 1, \dots, K-1\}^d$$

531 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
532 $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \leq s-1$, define

533
$$\xi_{\boldsymbol{\alpha}, i} = (\partial^{\boldsymbol{\alpha}} f(\frac{\boldsymbol{\eta}}{K}) + 1)/2.$$

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

536
$$\text{NN}(\text{width} \leq 8s(2N+3)\log_2(4N); \text{depth} \leq (5L+8)\log_2(2L))$$

537 such that

538
$$|\widetilde{\phi}_{\boldsymbol{\alpha}}(i) - \xi_{\boldsymbol{\alpha}, i}| \leq N^{-2s} L^{-2s}, \quad \text{for } i = 0, 1, \dots, K^d - 1 \text{ and } \|\boldsymbol{\alpha}\|_1 \leq s-1.$$

539 Define

540
$$\phi_{\boldsymbol{\alpha}}(\mathbf{x}) := 2\widetilde{\phi}_{\boldsymbol{\alpha}}\left(\sum_{j=1}^d x_j K^j\right) - 1, \quad \text{for any } \mathbf{x} = [x_1, x_2, \dots, x_d]^d \in \mathbb{R}^d.$$

For each $\|\alpha\|_1 \leq s-1$, it is clear that ϕ_α is also in

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

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

$$\left| \phi_\alpha\left(\frac{\eta}{K}\right) - \partial^\alpha f\left(\frac{\eta}{K}\right) \right| = \left| 2\tilde{\phi}_\alpha\left(\sum_{j=1}^d \eta_j K^{j-1}\right) - 1 - (2\xi_{\alpha,i} - 1) \right| = 2|\tilde{\phi}_\alpha(i) - \xi_{\alpha,i}| \leq 2N^{-2s}L^{-2s}.$$

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

$$\left| \phi_\alpha(\psi(\mathbf{x})) - \partial^\alpha f(\psi(\mathbf{x})) \right| = \left| \phi_\alpha\left(\frac{\theta}{K}\right) - \partial^\alpha f\left(\frac{\theta}{K}\right) \right| \leq 2N^{-2s}L^{-2s} := \mathcal{E}_3. \quad (4.3)$$

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

$$\phi(\mathbf{x}) := \sum_{\|\alpha\|_1 \leq s-1} \tilde{\phi}\left(\frac{\phi_\alpha(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{x} - \psi(\mathbf{x}))\right), \quad \text{for any } \mathbf{x} \in \mathbb{R}^d. \quad (4.4)$$

Step 3: Approximation error estimation.

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

$$\begin{aligned} & \left| \sum_{\|\alpha\|_1 \leq s-1} \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha + \sum_{\|\alpha\|_1 = s} \frac{\partial^\alpha f(\psi(\mathbf{x}) + \xi_{\mathbf{x}} \mathbf{h})}{\alpha!} \mathbf{h}^\alpha - \sum_{\|\alpha\|_1 \leq s-1} \tilde{\phi}\left(\phi_\alpha(\psi(\mathbf{x})), P_\alpha(\mathbf{x} - \psi(\mathbf{x}))\right) \right| \\ & \leq \sum_{\|\alpha\|_1 = s} \left| \frac{\partial^\alpha f(\psi(\mathbf{x}) + \xi_{\mathbf{x}} \mathbf{h})}{\alpha!} \mathbf{h}^\alpha \right| + \sum_{\|\alpha\|_1 \leq s-1} \left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha - \tilde{\phi}\left(\phi_\alpha(\psi(\mathbf{x})), P_\alpha(\mathbf{h})\right) \right| := \mathcal{J}_1 + \mathcal{J}_2. \end{aligned}$$

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

$$\mathcal{J}_1 = \sum_{\|\alpha\|_1 = s} \left| \frac{\partial^\alpha f(\psi(\mathbf{x}) + \xi_{\mathbf{x}} \mathbf{h})}{\alpha!} \mathbf{h}^\alpha \right| \leq \sum_{\|\alpha\|_1 = s} \left| \frac{1}{\alpha!} \mathbf{h}^\alpha \right| \leq (s+1)^{d-1} K^{-s}.$$

Now let us estimate the second part \mathcal{J}_2 as follows.

$$\begin{aligned} \mathcal{J}_2 &= \sum_{\|\alpha\|_1 \leq s-1} \left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha - \tilde{\phi}\left(\frac{\phi_\alpha(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right) \right| \\ &\leq \sum_{\|\alpha\|_1 \leq s-1} \left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha - \tilde{\phi}\left(\frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right) \right| \\ &\quad + \sum_{\|\alpha\|_1 \leq s-1} \left| \tilde{\phi}\left(\frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right) - \tilde{\phi}\left(\phi_\alpha(\psi(\mathbf{x})), P_\alpha(\mathbf{h})\right) \right| \\ &:= \mathcal{J}_{2,1} + \mathcal{J}_{2,2}. \end{aligned}$$

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

561 have, for any $\mathbf{x} \in Q_\theta$,

$$\begin{aligned}
\mathcal{J}_{2,1} &= \sum_{\|\alpha\|_1 \leq s-1} \left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha - \tilde{\phi}\left(\frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right) \right| \\
562 \quad &\leq \sum_{\|\alpha\|_1 \leq s-1} \left(\left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} \mathbf{h}^\alpha - \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} P_\alpha(\mathbf{h}) \right| + \left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} P_\alpha(\mathbf{h}) - \tilde{\phi}\left(\frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right) \right| \right) \\
&\leq \sum_{\|\alpha\|_1 \leq s-1} \left(\frac{1}{\alpha!} |\mathbf{h}^\alpha - P_\alpha(\mathbf{h})| + \mathcal{E}_1 \right) \leq \sum_{\|\alpha\|_1 \leq s-1} (\mathcal{E}_2 + \mathcal{E}_1) \leq s^d (\mathcal{E}_1 + \mathcal{E}_2).
\end{aligned}$$

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

$$564 \quad |\tilde{\phi}(x_1, x_2) - \tilde{\phi}(\bar{x}_1, x_2)| \leq |\tilde{\phi}(x_1, x_2) - x_1 x_2| + |\tilde{\phi}(\bar{x}_1, x_2) - \bar{x}_1 x_2| + |x_1 x_2 - \bar{x}_1 x_2| \leq 2\mathcal{E}_1 + 3|x_1 - \bar{x}_1|.$$

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

$$\begin{aligned}
\mathcal{J}_{2,2} &= \sum_{\|\alpha\|_1 \leq s-1} \left| \tilde{\phi}\left(\frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!}, P_\alpha(\mathbf{h})\right) - \tilde{\phi}(\phi_\alpha(\psi(\mathbf{x})), P_\alpha(\mathbf{h})) \right| \\
568 \quad &\leq \sum_{\|\alpha\|_1 \leq s-1} \left(2\mathcal{E}_1 + 3 \left| \frac{\partial^\alpha f(\psi(\mathbf{x}))}{\alpha!} - \phi_\alpha(\psi(\mathbf{x})) \right| \right) \leq \sum_{\|\alpha\|_1 \leq s-1} (2\mathcal{E}_1 + 3\mathcal{E}_3) \leq s^d (2\mathcal{E}_1 + 3\mathcal{E}_3).
\end{aligned}$$

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

$$\begin{aligned}
|f(\mathbf{x}) - \phi(\mathbf{x})| &\leq \mathcal{J}_1 + \mathcal{J}_2 \leq \mathcal{J}_1 + \mathcal{J}_{2,1} + \mathcal{J}_{2,2} \\
570 \quad &\leq (s+1)^{d-1} K^{-s} + s^d (\mathcal{E}_1 + \mathcal{E}_2) + s^d (2\mathcal{E}_1 + 3\mathcal{E}_3) \\
&\leq (s+1)^d (K^{-s} + 3\mathcal{E}_1 + \mathcal{E}_2 + 3\mathcal{E}_3).
\end{aligned}$$

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

$$573 \quad |f(\mathbf{x}) - \phi(\mathbf{x})| \leq (s+1)^d (K^{-s} + 3\mathcal{E}_1 + \mathcal{E}_2 + 3\mathcal{E}_3), \quad \text{for any } \mathbf{x} \in [0, 1]^d \setminus \Omega(K, \delta, d).$$

574 Recall that $(N+1)^{-7s(L+1)} \leq (N+1)^{-2s(L+1)} \leq (N+1)^{-2s} 2^{-2sL} \leq N^{-2s} L^{-2s}$ and $K =$
575 $\lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor \geq \frac{N^{2/d} L^{2/d}}{8}$. Then we have

$$\begin{aligned}
&(s+1)^d (K^{-s} + 3\mathcal{E}_1 + \mathcal{E}_2 + 3\mathcal{E}_3) \\
&= (s+1)^d \left(K^{-s} + 648(N+1)^{-2s(L+1)} + 9s(N+1)^{-7s(L+1)} + 6N^{-2s} L^{-2s} \right) \\
576 \quad &\leq (s+1)^d \left(8^s N^{-2s/d} L^{-2s/d} + (654 + 9s) N^{-2s} L^{-2s} \right) \\
&\leq (s+1)^d (8^s + 654 + 9s) N^{-2s/d} L^{-2s/d} \leq 84(s+1)^d 8^s N^{-2s/d} L^{-2s/d}.
\end{aligned}$$

577 What remaining is to estimate the width and depth of ϕ . Recall that $\psi \in \text{NN}(\text{width} \leq$
578 $d(4N+5); \text{depth} \leq 4(L+1))$, $\tilde{\phi} \in \text{NN}(\text{width} \leq 9N+10; \text{depth} \leq 2s(L+1))$, $P_\alpha \in$
579 $\text{NN}(\text{width} \leq 9N+s+6; \text{depth} \leq 7s^2(L+1))$, and $\phi_\alpha \in \text{NN}(\text{width} \leq 8s(2N+3)\log_2(4N); \text{depth} \leq$
580 $(5L+8)\log_2(2L))$ for $\alpha \in \mathbb{N}$ with $\|\alpha\|_1 \leq s-1$. By Equation (4.4), ϕ can be implemented
581 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
582 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 $xy = 2\left(\left(\frac{x+y}{2}\right)^2 - \left(\frac{x}{2}\right)^2 - \left(\frac{y}{2}\right)^2\right)$.
- $f(x_1, x_2, \dots, x_d) = x_1 x_2 \dots x_d$. We approximate $f(x_1, x_2, \dots, x_d) = x_1 x_2 \dots x_d$ 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 $Cz_1 z_2 \dots z_k$, 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 [21] 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 [21], 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*

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

Proof. Define a set of teeth functions $T_i : [0, 1] \rightarrow [0, 1]$ by induction as follows. Let

$$T_1(x) = \begin{cases} 2x, & x \leq \frac{1}{2}, \\ 2(1-x), & x > \frac{1}{2}, \end{cases}$$

and

$$T_i = T_{i-1} \circ T_1, \quad \text{for } i = 2, 3, \dots.$$

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 .

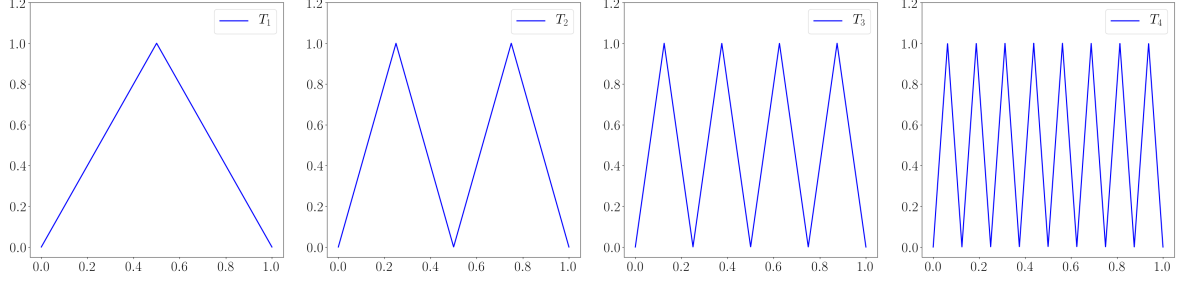


Figure 5: Illustrations of teeth functions T_1 , T_2 , T_3 , and T_4 .

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

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

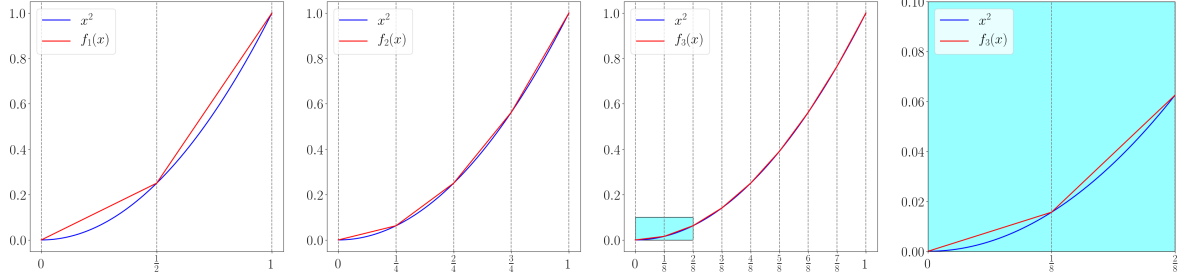


Figure 6: Illustrations of f_1 , f_2 , and f_3 .

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

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

and

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

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 $(k-1)2^{k-1} + 1 \leq N \leq k2^k$. 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 $k2^k + 1 \leq 3N$ 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 identity function as their activation

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

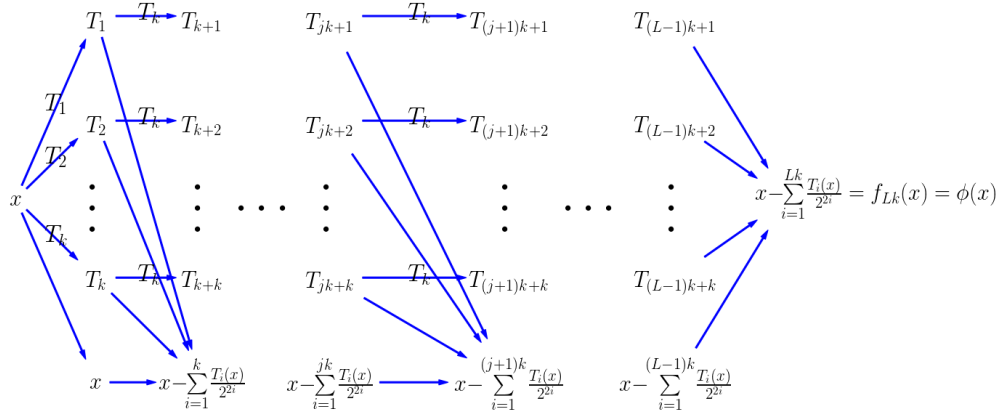


Figure 7: An illustration of the target ReLU FNN for approximating x^2 . We drop the ReLU activation function in this figure since $T_i(x)$ is always positive for all $i \in \mathbb{N}^+$ and $x \in [0, 1]$. Each arrow with T_k means that there is a ReLU FNN approximating T_k and mapping the function 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 $\mathbf{W}_1 \in \mathbb{R}^{N_1 \times N_2}$, $\mathbf{W}_2 \in \mathbb{R}^{N_2 \times N_3}$, and σ is an identity map that can be applied to vectors or matrices elementwisely, then $\mathbf{W}_1 \sigma(\mathbf{W}_2 \mathbf{x}) = \mathbf{W}_3 \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^{N_3}$, where $\mathbf{W}_3 = \mathbf{W}_1 \cdot \mathbf{W}_2 \in \mathbb{R}^{N_1 \times N_3}$.

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

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

where the last inequality comes from $N \leq k2^k \leq 2^{2k}$. So we finish the proof. \square

We have constructed a ReLU FNN to approximate $f(x) = x^2$. By the fact $xy = 2\left(\left(\frac{x+y}{2}\right)^2 - \left(\frac{x}{2}\right)^2 - \left(\frac{y}{2}\right)^2\right)$, 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*

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

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

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

Together with the fact

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

we construct the target function ϕ as

$$\phi(x, y) := 2\left(\psi\left(\frac{x+y}{2}\right) - \psi\left(\frac{x}{2}\right) - \psi\left(\frac{y}{2}\right)\right), \quad \text{for any } x, y \in \mathbb{R}.$$

It follows that

$$\begin{aligned} |xy - \phi(x, y)| &= \left| 2\left(\left(\frac{x+y}{2}\right)^2 - \left(\frac{x}{2}\right)^2 - \left(\frac{y}{2}\right)^2\right) - 2\left(\psi\left(\frac{x+y}{2}\right) - \psi\left(\frac{x}{2}\right) - \psi\left(\frac{y}{2}\right)\right) \right| \\ &\leq 2\left|\left(\frac{x+y}{2}\right)^2 - \psi\left(\frac{x+y}{2}\right)\right| + 2\left|\left(\frac{x}{2}\right)^2 - \psi\left(\frac{x}{2}\right)\right| + 2\left|\left(\frac{y}{2}\right)^2 - \psi\left(\frac{y}{2}\right)\right| \leq 6N^{-L}. \end{aligned}$$

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

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

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

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

$$\left|\psi\left(\frac{x-a}{b-a}, \frac{y-a}{b-a}\right) - \frac{x-a}{b-a} \frac{y-a}{b-a}\right| \leq 6N^{-L}, \quad \text{for any } x, y \in [a, b].$$

It follows that

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

Define

$$\phi(x, y) := (b-a)^2\psi\left(\frac{x-a}{b-a}, \frac{y-a}{b-a}\right) + a(x+y) - a^2, \quad \text{for any } x, y \in \mathbb{R}.$$

Hence,

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

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

The next lemma constructs a ReLU FNN to approximate a multivariable function $f(x_1, x_2, \dots, x_k) = x_1x_2 \cdots x_k$ 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*

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

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

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

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

682 • ϕ_i is a ReLU FNN with width $9(N+1)+i-1$ and depth $7kiL$ for each $i \in \{1, 2, \dots, k-1\}$.
 683

684 • The following inequality holds for any $i \in \{1, 2, \dots, k-1\}$ and $x_1, x_2, \dots, x_{i+1} \in [0, 1]$

$$685 \quad |\phi_i(x_1, \dots, x_{i+1}) - x_1 x_2 \dots x_{i+1}| \leq 9i(N+1)^{-7kL}. \quad (5.3)$$

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

687 1. When $i = 1$, it is obvious that the two required conditions are true: 1) $9(N+1)+i-1 =$
 688 $9(N+1)$ and $iL = L$ if $i = 1$; 2) Equation (5.2) implies Equation (5.3) for $i = 1$.

689 2. Now assume ϕ_i has been defined, then define

$$690 \quad \phi_{i+1}(x_1, \dots, x_{i+2}) := \phi_1(\phi_i(x_1, \dots, x_{i+1}), x_{i+2}), \quad \text{for any } x_1, \dots, x_{i+2} \in \mathbb{R}.$$

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

694 By the hypothesis of induction, we have

$$695 \quad |\phi_i(x_1, \dots, x_{i+1}) - x_1 x_2 \dots x_{i+1}| \leq 9i(N+1)^{-7kL}.$$

696 Recall the fact $9i(N+1)^{-7kL} \leq 9k2^{-7k} \leq 9k\frac{1}{90k} = 0.1$ for any $N, L, k \in \mathbb{N}^+$ and
 697 $i \in \{1, 2, \dots, k-1\}$. It follows that

$$698 \quad \phi_i(x_1, \dots, x_{i+1}) \in [-0.1, 1.1], \quad \text{for any } x_1, \dots, x_{i+1} \in [0, 1].$$

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

$$\begin{aligned} & |\phi_{i+1}(x_1, \dots, x_{i+2}) - x_1 x_2 \dots x_{i+2}| = |\phi_1(\phi_i(x_1, \dots, x_{i+1}), x_{i+2}) - x_1 x_2 \dots x_{i+2}| \\ 700 & \leq |\phi_1(\phi_i(x_1, \dots, x_{i+1}), x_{i+2}) - \phi_i(x_1, \dots, x_{i+1})x_{i+2}| + |\phi_i(x_1, \dots, x_{i+1})x_{i+2} - x_1 x_2 \dots x_{i+2}| \\ & \leq 9(N+1)^{-7kL} + 9i(N+1)^{-7kL} = 9(i+1)(N+1)^{-7kL}. \end{aligned}$$

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

$$702 \quad |\phi(x_1, \dots, x_k) - x_1 x_2 \dots x_k| \leq 9(k-1)(N+1)^{-7kL}, \quad \text{for any } x_1, x_2, \dots, x_k \in [0, 1].$$

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

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

706 *Proof of Proposition 4.1.* Denote $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_d]^T$ and let $[z_1, z_2, \dots, z_k]^T$ be the vector
 707 such that

$$708 \quad z_\ell = x_j, \quad \text{if } \sum_{i=1}^{j-1} \alpha_i < \ell \leq \sum_{i=1}^j \alpha_i, \quad \text{for } j = 1, 2, \dots, d.$$

709 That is,

$$710 \quad [z_1, z_2, \dots, z_k]^T = \left[\overbrace{x_1, \dots, x_1}^{\alpha_1 \text{ times}}, \overbrace{x_2, \dots, x_2}^{\alpha_2 \text{ times}}, \dots, \overbrace{x_d, \dots, x_d}^{\alpha_d \text{ times}} \right]^T \in \mathbb{R}^k.$$

711 Then we have $P(\mathbf{x}) = \mathbf{x}^\alpha = z_1 z_2 \dots z_k$.

712 We construct the target ReLU FNN in two steps. First, there exists a linear map
 713 ϕ_1 that duplicates inputs in \mathbf{x} to form a new vector $[z_1, z_2, \dots, z_k]^T$. Second, by Lemma
 714 5.3, there exists such a ReLU FNN ϕ_2 with width $9(N+1) + k - 2$ and depth $7k(k-1)L$
 715 such that ϕ_2 maps $[z_1, z_2, \dots, z_k]^T$ to $P(\mathbf{x}) = z_1 z_2 \dots z_k$ within the target accuracy. Hence,
 716 we can construct our final target ReLU FNN via $\phi_2 \circ \phi_1(\mathbf{x}) = \phi(\mathbf{x})$. By incorporating
 717 the linear map in ϕ_1 into the first linear map of ϕ , we can treat ϕ as a ReLU FNN with
 718 width $9(N+1) + k - 2$ and depth $7k(k-1)L$ with a desired approximation accuracy. So,
 719 we finish the proof. \square

720 5.2 Proof of Proposition 4.3 for step function approximation

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

725 **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
 726 $x_0 < x_1 < \dots < x_{N_1(N_2+1)}$ and $y_i \geq 0$ for $i = 0, 1, \dots, N_1(N_2+1)$, there exists $\phi \in \text{NN}(\#input =$
 727 $1; \text{widthvec} = [2N_1, 2N_2 + 1])$ satisfying the following conditions.*

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

730 The above lemma is Proposition 2.1 of [19] and the reader is referred to [19] for its
 731 proof. Essentially, this lemma shows the equivalence of one-hidden-layer ReLU FNNs of
 732 size $\mathcal{O}(N^2)$ and two-hidden-layer ones of size $\mathcal{O}(N)$ to fit $\mathcal{O}(N^2)$ samples.

733 The next lemma below shows that special shallow and wide ReLU FNNs can be
 734 represented by deep and narrow ones. This lemma was proposed as Proposition 2.2 in
 735 [19].

736 **Lemma 5.5.** *Given any $N, L \in \mathbb{N}^+$, for arbitrary $\phi_1 \in \text{NN}(\#input = 1; \text{widthvec} =$
 737 $[N, NL])$, there exists $\phi_2 \in \text{NN}(\#input = 1; \text{width} \leq 2N + 4; \text{depth} \leq L + 2)$ such that
 738 $\phi_1(x) = \phi_2(x)$ for any $x \in \mathbb{R}$.*

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

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

741 **Case 1:** $d = 1$.

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

$$743 \quad \left\{ \left(\frac{m}{M}, m \right) : m = 0, 1, \dots, M-1 \right\} \cup \left\{ \left(\frac{m+1}{M} - \delta, m \right) : m = 0, 1, \dots, M-2 \right\} \cup \{(1, M-1), (2, 0)\}.$$

Its cardinality is $2M + 1 = N \cdot ((2NL - 1) + 1) + 1$. By Lemma 5.4 with $N_1 = N$ and $N_2 = 2NL - 1$, there exist $\phi_1 \in \text{NN}(\text{widthvec} = [2N, 2(2NL - 1) + 1]) = \text{NN}(\text{widthvec} = [2N, 4NL - 1])$ 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 $m = 0, 1, \dots, M - 2$;
- ϕ_1 is linear on $[\frac{M-1}{M}, 1]$ and each interval $[\frac{m}{M}, \frac{m+1}{M} - \delta]$ for $m = 0, 1, \dots, M - 2$.

Then

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

Now consider the sample set

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

Its cardinality is $2L + 1 = 1 \cdot ((2L - 1) + 1) + 1$. By Lemma 5.4 with $N_1 = 1$ and $N_2 = 2L - 1$, there exists $\phi_2 \in \text{NN}(\text{widthvec} = [2, 2(2L - 1) + 1]) = \text{NN}(\text{widthvec} = [2, 4L - 1])$ 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 $\ell = 0, 1, \dots, L - 2$;
- ϕ_2 is linear on $[\frac{L-1}{ML}, \frac{1}{M}]$ and each interval $[\frac{\ell}{ML}, \frac{\ell+1}{ML} - \delta]$ for $\ell = 0, 1, \dots, L - 2$.

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

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

Define

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

Notice that each $k \in \{0, 1, \dots, ML - 1\} = \{0, 1, \dots, K - 1\}$ can be uniquely represented by $k = mL + \ell$ for $m \in \{0, 1, \dots, M - 1\}$ and $\ell \in \{0, 1, \dots, L - 1\}$. 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 $m \in \{0, 1, \dots, M - 1\}$, $\ell \in \{0, 1, \dots, L - 1\}$, we have

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

By Lemma 5.5,

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

and

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

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

Case 2: $d \geq 2$.

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

$$\{(\frac{k}{K}, \frac{k}{K}) : k = 0, 1, \dots, K - 1\} \cup \{(\frac{k+1}{K} - \delta, \frac{k}{K}) : k = 0, 1, \dots, K - 2\} \cup \{(1, \frac{K-1}{K}), (2, 1)\},$$

whose cardinality is $2K + 1 = \lfloor N^{1/d} \rfloor ((2\lfloor N^{1/d} \rfloor \lfloor L^{2/d} \rfloor - 1) + 1) + 1$. 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

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

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

Then

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

By Lemma 5.5,

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

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 [19].

Lemma 5.6. *For any $N, L \in \mathbb{N}^+$, any $\theta_{m,\ell} \in \{0, 1\}$ for $m = 0, 1, \dots, M-1$, $\ell = 0, 1, \dots, L-1$, 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 $m = 0, 1, \dots, M-1$, $\ell = 0, 1, \dots, L-1$.*

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 $i = 0, 1, \dots, N^2L^2 - 1$, there exists a ReLU FNN ϕ with width $8N + 10$ and depth $5L + 6$ such that $\phi(i) = \theta_i$, for $i = 0, 1, \dots, N^2L^2 - 1$.*

^⑧Fitting 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 .

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

803 Denote $M = N^2L$, for each $i \in \{0, 1, \dots, N^2L^2 - 1\}$, there exists a unique representation
 804 $i = mL + \ell$ for $m = 0, 1, \dots, M - 1$ and $L = 0, 1, \dots, L - 1$. So we define, for $m = 0, 1, \dots, M - 1$
 805 and $\ell = 0, 1, \dots, L - 1$,

$$806 \quad a_{m,\ell} := \theta_i, \quad \text{where } i = mL + \ell.$$

807 Then we set $b_{m,0} = 0$ for $m = 0, 1, \dots, M - 1$ and $b_{m,\ell} = a_{m,\ell-1}$ for $m = 0, 1, \dots, M - 1$ and
 808 $\ell = 1, \dots, L - 1$.

809 By Lemma 5.6, there exist $\phi_1, \phi_2 \in \text{NN}$ such that

$$810 \quad \phi_1(m, \ell) = \sum_{j=1}^{\ell} a_{m,j} \quad \text{and} \quad \phi_2(m, \ell) = \sum_{j=1}^{\ell} b_{m,j},$$

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

$$812 \quad \{(mL, m) : m = 0, 1, \dots, M\} \cup \{((m+1)L - 1, m) : m = 0, 1, \dots, M - 1\} \subseteq \mathbb{R}^2.$$

813 Its cardinality is $2M + 1 = N \cdot ((2NL - 1) + 1) + 1$. By Lemma 5.4 with $N_1 = N$ and
 814 $N_2 = 2NL - 1$, there exists $\psi \in \text{NN}(\#input = 1; \text{widthvec} = [2N, 2(2NL - 1) + 1]) =$
 815 $\text{NN}(\#input = 1; \text{widthvec} = [2N, 4NL - 1])$ such that

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

818 It follows that

$$819 \quad \psi(i) = m \quad \text{where } i = mL + \ell, \quad \text{for } m = 0, 1, \dots, M - 1 \text{ and } \ell = 0, 1, \dots, L - 1.$$

820 Define

$$821 \quad \phi(x) := \phi_1(\psi(x), x - L\psi(x)) - \phi_2(\psi(x), x - L\psi(x)), \quad \text{for any } x \in \mathbb{R}.$$

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

$$\begin{aligned} 824 \quad \phi(i) &= \phi_1(\psi(i), i - L\psi(i)) - \phi_2(\psi(i), i - L\psi(i)) \\ &= \phi_1(m, \ell) - \phi_2(m, \ell) \\ &= \sum_{j=1}^{\ell} a_{m,j} - \sum_{j=1}^{\ell} b_{m,j} = a_{m,\ell} = \theta_i. \end{aligned}$$

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

$$826 \quad \phi_1, \phi_2 \in \text{NN}(\text{width} \leq 4N + 5; \text{depth} \leq 3L + 4).$$

827 And by Lemma 5.5,

$$828 \quad \psi \in \text{NN}(\text{widthvec} = [2N, 4NL - 1]) \subseteq \text{NN}(\text{width} \leq 4N + 4; \text{depth} \leq 2L + 2).$$

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

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

832 *Proof of Proposition 4.4.* Denote $J = \lceil 2s \log_2(NL + 1) \rceil$. For each $\xi_i \in [0, 1]$, there exist
 833 $\xi_{i,1}, \xi_{i,2}, \dots, \xi_{i,J} \in \{0, 1\}$ such that

$$834 \quad |\xi_i - \text{Bin}_{0,\xi_{i,1}\xi_{i,2}\dots\xi_{i,J}}| \leq 2^{-J}, \quad \text{for } i = 0, 1, \dots, N^2L^2 - 1.$$

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

$$837 \quad \phi_j(i) = \xi_{i,j}, \quad \text{for } i = 0, 1, \dots, N^2L^2 - 1, j = 1, 2, \dots, J.$$

838 Define

$$839 \quad \tilde{\phi}(x) := \sum_{j=1}^J 2^{-j} \phi_j(x), \quad \text{for any } x \in \mathbb{R}.$$

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

$$841 \quad |\tilde{\phi}(i) - \xi_i| = \left| \sum_{j=1}^J 2^{-j} \phi_j(i) - \xi_i \right| = \left| \sum_{j=1}^J 2^{-j} \xi_{i,j} - \xi_i \right| = |\text{Bin}_{0,\xi_{i,1}\xi_{i,2}\dots\xi_{i,J}} - \xi_i| \leq 2^{-J}.$$

842 Notice that

$$843 \quad 2^{-J} = 2^{-\lceil 2s \log_2(NL+1) \rceil} \leq 2^{-2s \log_2(NL+1)} = (NL + 1)^{-2s} \leq N^{-2s} L^{-2s}.$$

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

$$845 \quad \begin{aligned} J &= \lceil 2s \log_2(NL + 1) \rceil \leq 2s(1 + \log_2(NL + 1)) \leq 2s(1 + \log_2(2N) + \log_2 L) \\ &\leq 2s(1 + \log_2(2N))(1 + \log_2 L) \leq 2s \log_2(4N) \log_2(2L), \end{aligned}$$

846 and $\phi_j \in \text{NN}(\text{width} \leq 8N + 10; \text{depth} \leq 5L + 6)$. Then $\tilde{\phi} = \sum_{j=1}^J 2^{-j} \phi_j$ can be implemented
 847 by a ReLU FNN with width $2s(8N + 10) \log_2(4N) + 2 \leq 8s(2N + 3) \log_2(4N)$ and depth
 848 $(5L + 6) \log_2(2L)$.

849 Finally, we define

$$850 \quad \phi(x) = \min \{ \max \{ 0, \tilde{\phi}(x) \}, 1 \}, \quad \text{for any } x \in \mathbb{R}.$$

851 Then $0 \leq \phi(x) \leq 1$ for any $x \in \mathbb{R}$ and ϕ can be implemented by a ReLU FNN with width
 852 $8s(2N + 3) \log_2(4N)$ and depth $(5L + 6) \log_2(2L) + 2 \leq (5L + 8) \log_2(2L)$. Notice that

$$853 \quad \tilde{\phi}(i) = \sum_{j=1}^J 2^{-j} \phi_j(i) = \sum_{j=1}^J 2^{-j} \xi_{i,j} \in [0, 1], \quad \text{for } i = 0, 1, \dots, N^2L^2 - 1.$$

854 It follows that

$$855 \quad |\phi(i) - \xi_i| = \left| \min \{ \max \{ 0, \tilde{\phi}(i) \}, 1 \} - \xi_i \right| = |\tilde{\phi}(i) - \xi_i| \leq N^{-2s} L^{-2s}, \quad \text{for } i = 0, 1, \dots, N^2L^2 - 1.$$

856 The proof is complete. □

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 was partially supported by National University of Singapore under Grant R-146-000-251-133, and by the Ministry of Education in Singapore under the grant MOE2018-T2-2-147.

References

- [1] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.
- [2] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.
- [3] P. Bartlett, V. Maiorov, and R. Meir. Almost linear VC dimension bounds for piecewise polynomial networks. *Neural Computation*, 10:217–3, 1998.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] G. Cybenko. Approximation by superpositions of a sigmoidal function. *MCSS*, 2:303–314, 1989.

- [8] W. E and Q. Wang. Exponential convergence of the deep neural network approximation for analytic functions. *CoRR*, abs/1807.00297, 2018.
- [9] 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.
- [10] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.
- [11] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.
- [12] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, June 1994.
- [13] S. Liang and R. Srikant. Why deep neural networks? *CoRR*, abs/1610.04161, 2016.
- [14] 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.
- [15] H. Montanelli, H. Yang, and Q. Du. Deep relu networks overcome the curse of dimensionality for bandlimited functions. 2019.
- [16] 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.
- [17] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296 – 330, 2018.
- [18] 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.
- [19] Z. Shen, H. Yang, and S. Zhang. Deep Network Approximation Characterized by Number of Neurons. *arXiv e-prints*, page arXiv:1906.05497, Jun 2019.
- [20] Z. Shen, H. Yang, and S. Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74 – 84, 2019.
- [21] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103 – 114, 2017.

- 928 [22] D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU
929 networks. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st*
930 *Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning*
931 *Research*, pages 639–649. PMLR, 06–09 Jul 2018.
- 932 [23] D.-X. Zhou. Universality of deep convolutional neural networks. *Applied and Com-*
933 *putational Harmonic Analysis*, 2019.