

Deep Network Approximation: Beyond ReLU to Diverse Activation Functions

Shijun Zhang*

*Department of Mathematics
Duke University*

SHIJUN.ZHANG@DUKE.EDU

Jianfeng Lu

*Department of Mathematics
Duke University*

JIANFENG@MATH.DUKE.EDU

Hongkai Zhao

*Department of Mathematics
Duke University*

ZHAO@MATH.DUKE.EDU

Abstract

This paper explores the expressive power of deep neural networks for a diverse range of activation functions. An activation function set \mathcal{A} is defined to encompass the majority of commonly used activation functions, such as ReLU, LeakyReLU, ReLU², ELU, SELU, Softplus, GELU, SiLU, Swish, Mish, Sigmoid, Tanh, Arctan, Softsign, dSiLU, and SRS. We demonstrate that for any activation function $\varrho \in \mathcal{A}$, a ReLU network of width N and depth L can be approximated to arbitrary precision by a ϱ -activated network of width $6N$ and depth $2L$ on any bounded set. This finding enables the extension of most approximation results achieved with ReLU networks to a wide variety of other activation functions, at the cost of slightly larger constants.

1 Introduction

In the realm of artificial intelligence (AI), deep neural networks have emerged as a powerful tool. By harnessing the potential of interconnected nodes organized into multiple layers, deep neural networks have showcased notable success in many challenging applications and new territories. The foundation of deep neural networks consists of a linear transformation followed by an activation function. The activation function plays an important role in the successful training of deep neural networks. In recent years, the Rectified Linear Unit (ReLU) [27] has experienced a surge in popularity and demonstrated its effectiveness as an activation function.

The adoption of ReLU has marked a significant improvement of results on challenging datasets in supervised learning [20]. Optimizing deep networks activated by ReLU is simpler compared to networks utilizing other activation functions such as Sigmoid or Tanh, since gradients can propagate when the input to ReLU is positive. It was also shown in the recent work [44] that using ReLU makes the network a less regularizer compared to other smoother activation functions in practice. The effectiveness and simplicity of ReLU have positioned it as the preferred default activation function in the deep learning community. A significant number of publications have extensively investigated the expressive capabilities of deep neural networks, with the majority of them primarily focusing on the ReLU activation function.

* Corresponding author.

In recent developments, various alternative activation functions have been proposed as replacements for ReLU. Notable examples include the Leaky ReLU (LeakyReLU) [24], the Exponential Linear Units (ELU) [9], and the Gaussian Error Linear Unit (GELU) [16]. These alternative activation functions have exhibited improved performance in specific neural network architectures. Among these alternatives, GELU has gained significant popularity in deep learning models, especially in the realm of natural language processing (NLP) tasks. They have been successfully employed in prominent models such as GPT-3 [5], BERT [11], XLNet [39], and various other transformer models. While these recently proposed activation functions have demonstrated promising empirical results, their theoretical underpinnings are still being developed. This paper aims to investigate the expressive capabilities of deep neural networks utilizing these activation functions. In doing so, we establish connections between these functions and ReLU, allowing us to extend most existing approximation results for ReLU networks to encompass other activation functions such as ELU and GELU.

More precisely, we will define an activation function set, denoted as \mathcal{A} , which contains the majority of commonly used activation functions. To the best of our knowledge, activation functions can be broadly categorized into three cases. The first case consists of piecewise smooth functions, e.g., ReLU, LeakyReLU, ReLU² (ReLU squared) [36], ELU, and SELU (Scaled Exponential Linear Unit) [19]. These activation functions are continuous piecewise smooth functions belonging to the set $\mathcal{A}_1 := \cup_{k=0}^{\infty} \mathcal{A}_{1,k}$, where $\mathcal{A}_{1,k}$, for each smoothness index $k \in \mathbb{N}$, is defined as

$$\mathcal{A}_{1,k} := \left\{ \varrho : \mathbb{R} \rightarrow \mathbb{R} \mid \exists a_0 < b_0, \varrho \in C^k((a_0, b_0)), \exists x_0 \in (a_0, b_0), \right. \\ \left. \mathbb{R} \ni \lim_{t \rightarrow 0^-} \frac{\varrho^{(k)}(x_0+t) - \varrho^{(k)}(x_0)}{t} \neq \lim_{t \rightarrow 0^+} \frac{\varrho^{(k)}(x_0+t) - \varrho^{(k)}(x_0)}{t} \in \mathbb{R} \right\}.$$

It is worth noting that $\varrho \in C^k((a_0, b_0)) \setminus C^{k+1}((a_0, b_0))$ is necessary to ensure $\varrho \in \mathcal{A}_{1,k}$. Specifically, at $x_0 \in (a_0, b_0)$, the left and right derivatives of $\varrho^{(k)}$ must exist and be distinct. However, there are no specific requirements placed on ϱ outside (a_0, b_0) . Here and in the sequel, we use $f^{(k)}$ to represent the k -th derivative of a function f . For instance, $f^{(0)}$ refers to the function itself, and $f^{(1)}$ represents the first derivative. The set of functions whose k -th derivative exists and is continuous on a domain Ω is denoted as $C^k(\Omega)$. Specifically, when $k = 0$, we have $C(\Omega) = C^0(\Omega)$, the set of continuous functions on Ω .

The second case consists of smooth variants of ReLU, e.g., Softplus [13], GELU, SiLU (Sigmoid Linear Unit) [12, 16], Swish [29], and Mish [25]. These activation functions are included in the set \mathcal{A}_2 , defined via

$$\mathcal{A}_2 := \left\{ \varrho : \mathbb{R} \rightarrow \mathbb{R} \mid \forall A, \sup_{x \in [-A, A]} |\varrho(x)| < \infty, \exists x_0 \in \mathbb{R}, \varrho''(x_0) \neq 0, \exists T_0 > 0, \right. \\ \left. \mathbb{R} \ni \lim_{x \rightarrow -\infty} (\varrho(x + T_0) - \varrho(x)) \neq \lim_{x \rightarrow \infty} (\varrho(x + T_0) - \varrho(x)) \in \mathbb{R} \right\}.$$

The set \mathcal{A}_2 encompasses a wide range of activation functions, some of which can even be discontinuous. To provide a clearer understanding, we present a refined subset of \mathcal{A}_2 below.

$$\mathcal{A}_2 \supseteq \left\{ \varrho \in C(\mathbb{R}) : \exists x_0 \in \mathbb{R}, \varrho''(x_0) \neq 0, \mathbb{R} \ni \lim_{x \rightarrow -\infty} \varrho'(x) \neq \lim_{x \rightarrow \infty} \varrho'(x) \in \mathbb{R} \right\}.$$

The final case consists of S-shaped functions, e.g., Sigmoid, Tanh, Arctan, Softsign [38]. These functions are part of the set \mathcal{A}_3 , which is defined via

$$\mathcal{A}_3 := \left\{ \varrho : \mathbb{R} \rightarrow \mathbb{R} \mid \sup_{x \in \mathbb{R}} |\varrho(x)| < \infty, \exists x_0 \in \mathbb{R}, \varrho''(x_0) \neq 0, \right. \\ \left. \mathbb{R} \ni \lim_{x \rightarrow -\infty} \varrho(x) \neq \lim_{x \rightarrow \infty} \varrho(x) \in \mathbb{R} \right\}.$$

The set \mathcal{A}_3 can be regarded as a collection of generalized S-shaped functions, which encompasses additional activation functions, such as dSiLU (derivative of SiLU) [12] and SRS (Soft-Root-Sign) [21]. Moreover, the derivatives of Softplus, GELU, SiLU, Swish, and Mish are also classified within \mathcal{A}_3 .

Then the activation function set \mathcal{A} is defined as the union of $\cup_{k=0}^4 \mathcal{A}_{1,k}$, \mathcal{A}_2 , and \mathcal{A}_3 :

$$\mathcal{A} := \left(\cup_{k=0}^4 \mathcal{A}_{1,k} \right) \cup \mathcal{A}_2 \cup \mathcal{A}_3.$$

The definitions of \mathcal{A} , $\mathcal{A}_{1,k}$ for $k \in \mathbb{N}$, \mathcal{A}_2 , and \mathcal{A}_3 will remain consistent throughout the whole paper. It is worth noting that if $\varrho \in \mathcal{A}$, then $w_1 \varrho(w_0 x + b_0) + b_1 \in \mathcal{A}$ provided $w_0 \neq 0 \neq w_1$. Notably, the set \mathcal{A} encompasses the majority of commonly used activation functions, such as ReLU, LeakyReLU, ReLU², ELU, SELU, Softplus, GELU, SiLU, Swish, Mish, Sigmoid, Tanh, Arctan, Softsign, dSiLU, SRS, and their modified versions achieved by employing translation, non-zero scaling, and reflection operations. In Section 2.3, we will present definitions and visual representations of the activation functions mentioned above.

Define the supremum norm of a vector-valued function $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ by

$$\|\mathbf{f}\|_{\sup([-A, A]^d)} := \sup \{ |f_i(\mathbf{x})| : \mathbf{x} \in [-A, A]^d, i \in \{1, 2, \dots, n\} \},$$

where f_i is the i -th component of \mathbf{f} . Let \mathbb{N} denote the set of natural numbers, i.e., $\mathbb{N} := \{0, 1, 2, \dots\}$, and set $\mathbb{N}^+ := \mathbb{N} \setminus \{0\}$. This paper exclusively focuses on fully connected feed-forward neural networks. We denote $\mathcal{NN}_\varrho\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ as the set of functions $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ that can be represented by ϱ -activated networks of width $\leq N \in \mathbb{N}^+$ and depth $\leq L \in \mathbb{N}^+$. In our context, the width of a network refers to the maximum number of neurons in a hidden layer and the depth corresponds to the number of hidden layers. For instance, suppose $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is a vector-valued function realized by a ϱ -activated network, where ϱ is the activation function that can be applied elementwise to a vector input. Then ϕ can be expressed as

$$\phi = \mathcal{L}_L \circ \varrho \circ \mathcal{L}_{L-1} \circ \dots \circ \varrho \circ \mathcal{L}_1 \circ \varrho \circ \mathcal{L}_0,$$

where \mathcal{L}_ℓ is an affine linear map given by $\mathcal{L}_\ell(\mathbf{y}) := \mathbf{W}_\ell \cdot \mathbf{y} + \mathbf{b}_\ell$ for $\ell = 0, 1, \dots, L$. Here, $\mathbf{W}_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ and $\mathbf{b}_\ell \in \mathbb{R}^{N_{\ell+1}}$ are the weight matrix and the bias vector with $N_0 = d$, $N_1, N_2, \dots, N_L \in \mathbb{N}^+$, and $N_{L+1} = n$. Clearly, $\phi \in \mathcal{NN}_\varrho\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, where $N = \max\{N_1, N_2, \dots, N_L\}$.

Our goal is to explore the expressiveness of deep neural networks activated by $\varrho \in \mathcal{A}$. In pursuit of this goal, the following theorem establishes connections between these functions and ReLU. This allows us to extend and generalize most existing approximation results for ReLU networks to activation functions in \mathcal{A} .

Theorem 1.1. *Suppose $\varrho \in \mathcal{A}$ and $\phi_{\text{ReLU}} \in \mathcal{NN}_{\text{ReLU}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ with $N, L, d, n \in \mathbb{N}^+$. Then for any $\varepsilon > 0$ and $A > 0$, there exists $\phi_\varrho \in \mathcal{NN}_\varrho\{6N, 2L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ such that*

$$\|\phi_\varrho - \phi_{\text{ReLU}}\|_{\sup([-A, A]^d)} < \varepsilon.$$

The proof of Theorem 1.1 can be found in Section 3. Theorem 1.1 implies that a ReLU network of width N and depth L can be approximated by a ϱ -activated network of width $6N$ and $2L$ arbitrarily well on any bounded set for any pre-specified $\varrho \in \mathcal{A}$. In other words, $\mathcal{NN}_\varrho\{6N, 2L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ is dense in $\mathcal{NN}_{\text{ReLU}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ in terms of the $\|\cdot\|_{\sup([-A, A]^d)}$ norm for any pre-specified $A > 0$ and $\varrho \in \mathcal{A}$. It is worth mentioning while Theorem 1.1 covers activation functions $\varrho \in \mathcal{A}_{1,k}$ only for $k = 0, 1, 2, 3, 4$, it is possible to obtain analogous results for larger values of $k \in \mathbb{N}$. For more detailed analysis and discussions, please refer to Section 2.1.

Equipped with Theorem 1.1, we can expand most existing approximation results for ReLU networks to encompass various alternative activation functions, albeit with slightly larger constants. To illustrate this point, we present several corollaries below. Theorem 1.1 of [32] implies that a ReLU network of width $C_{d,1}N$ and depth $C_{d,2}L$ can approximate a continuous function $f \in C([0, 1]^d)$ with an error $C_{d,3}\omega_f((N^2L^2 \ln(N+1))^{-1/d})$, where $C_{d,1}$, $C_{d,2}$, and $C_{d,3}$ are constants¹ determined by d , and $\omega_f(\cdot)$ is the modulus of continuity of $f \in C([0, 1]^d)$ defined via

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

By combining this result with Theorem 1.1, an immediate corollary follows.

Corollary 1.2. *Suppose $\varrho \in \mathcal{A}$ and $f \in C([0, 1]^d)$ with $d \in \mathbb{N}^+$. Then for any $N, L \in \mathbb{N}^+$, there exists $\phi \in \mathcal{NN}_{\varrho}\{C_{d,1}N, C_{d,2}L; \mathbb{R}^d \rightarrow \mathbb{R}\}$ such that*

$$\|f - \phi\|_{L^\infty([0, 1]^d)} \leq C_{d,3}\omega_f((N^2L^2 \ln(N+1))^{-1/d}),$$

where $C_{d,1}$, $C_{d,2}$, and $C_{d,3}$ are constants determined by d .

It is demonstrated in Theorem 1.1 of [35] that a ReLU network of width $C_{s,d,1}N \ln(N+1)$ and depth $C_{s,d,2}L \ln(L+1)$ can approximate a smooth function $f \in C^s([0, 1]^d)$ with an error $C_{s,d,3}\|f\|_{C^s([0, 1]^d)}N^{-2s/d}L^{-2s/d}$, where $C_{s,d,1}$, $C_{s,d,2}$, and $C_{s,d,3}$ are constants² determined by s and d . Here, the norm $\|f\|_{C^s([0, 1]^d)}$ for any $f \in C^s([0, 1]^d)$ is defined via

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

where $\partial^\alpha f$ denotes the partial derivative $\mathbf{x} \mapsto \frac{\partial^\alpha}{\partial \mathbf{x}^\alpha} f(\mathbf{x}) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} f(\mathbf{x})$ for any $\mathbf{x} = (x_1, \dots, x_d) \in [0, 1]^d$ and $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$. By combining the aforementioned result with Theorem 1.1, we can promptly deduce the subsequent corollary.

Corollary 1.3. *Suppose $\varrho \in \mathcal{A}$ and $f \in C^s([0, 1]^d)$ with $s, d \in \mathbb{N}^+$. Then for any $N, L \in \mathbb{N}^+$, there exists $\phi \in \mathcal{NN}_{\varrho}\{C_{s,d,1}N \ln(N+1), C_{s,d,2}L \ln(L+1); \mathbb{R}^d \rightarrow \mathbb{R}\}$ such that*

$$\|\phi - f\|_{L^\infty([0, 1]^d)} \leq C_{s,d,3}\|f\|_{C^s([0, 1]^d)}N^{-2s/d}L^{-2s/d},$$

where $C_{s,d,1}$, $C_{s,d,2}$, and $C_{s,d,3}$ are constants determined by s and d .

It is demonstrated in Theorem 1 of [6] that a continuous piecewise linear function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with $q \in \mathbb{N}^+$ pieces can be exactly represented by a ReLU network of width $\lceil 3q/2 \rceil q$ and depth $2\lceil \log_2 q \rceil + 1$. By combining this result with Theorem 1.1, we obtain the following corollary.

Corollary 1.4. *Suppose $\varrho \in \mathcal{A}$ and let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a continuous piecewise linear function with q pieces, where $d, q \in \mathbb{N}^+$. Then for any $\varepsilon > 0$ and $A > 0$, there exists $\phi \in \mathcal{NN}_{\varrho}\{6\lceil 3q/2 \rceil q, 4\lceil \log_2 q \rceil + 2; \mathbb{R}^d \rightarrow \mathbb{R}\}$, such that*

$$|\phi(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

It is demonstrated in [43] that even though a single fixed-size ReLU network has limited expressive capabilities, repeatedly composing it can create surprisingly expressive networks. Specifically, Theorem 1.1 of [43] establishes that $\mathcal{L}_2 \circ \mathbf{g}^{\circ(3r+1)} \circ \mathcal{L}_1$ can approximate a continuous function $f \in C([0, 1]^d)$ with an error $6\sqrt{d}\omega_f(r^{-1/d})$, where $\mathbf{g} \in \mathcal{NN}_{\text{ReLU}}\{69d + 48, 5; \mathbb{R}^{5d+5} \rightarrow \mathbb{R}^{5d+5}\}$, \mathcal{L}_1 and \mathcal{L}_2 are two affine linear maps matching the dimensions, and $\mathbf{g}^{\circ r}$ denotes the r -times composition of \mathbf{g} . By merging this outcome with Theorem 1.1, we can promptly deduce the subsequent corollary.

¹ The values of $C_{d,1}$, $C_{d,2}$, and $C_{d,3}$ are explicitly given in [32].

² The values of $C_{s,d,1}$, $C_{s,d,2}$, and $C_{s,d,3}$ are explicitly provided in [35].

Corollary 1.5. Suppose $\varrho \in \mathcal{A}$ and $f \in C([0, 1]^d)$ with $d \in \mathbb{N}^+$. Then for any $r \in \mathbb{N}^+$ and $p \in [1, \infty)$, there exist $\mathbf{g} \in \mathcal{NN}_{\varrho}\{414d + 288, 10; \mathbb{R}^{5d+5} \rightarrow \mathbb{R}^{5d+5}\}$ and two affine linear maps $\mathcal{L}_1 : \mathbb{R}^d \rightarrow \mathbb{R}^{5d+5}$ and $\mathcal{L}_2 : \mathbb{R}^{5d+5} \rightarrow \mathbb{R}$ such that

$$\|\mathcal{L}_2 \circ \mathbf{g}^{\circ(3r+1)} \circ \mathcal{L}_1 - f\|_{L^p([0,1]^d)} \leq 7\sqrt{d}\omega_f(r^{-1/d}).$$

It is worth highlighting that the approximation error in Corollary 1.5 is measured using the L^p -norm for any $p \in [1, \infty)$. Nevertheless, it is feasible to generalize this result to the L^∞ -norm as well, though it comes with larger associated constants. To accomplish this, we only need to combine Theorem 1.3 of [43] with Theorem 1.1.

The remainder of this paper is organized as follows. In Section 2, we explore some additional related topics. We present two supplementary theorems, Theorems 2.1 and 2.2, in Section 2.1 to complement Theorem 1.1. We also discuss related work in Section 2.2 and provide definitions and illustrations of common activation functions in Section 2.3. Moving forward to Section 3, we establish the proofs of Theorems 1.1, 2.1, and 2.2. In Section 3.1, we introduce the notations used throughout this paper. In Section 3.2, we present several propositions, namely Propositions 3.1, 3.2, 3.3, and 3.4, outlining the underlying ideas for proving Theorems 1.1, 2.1, and 2.2. Subsequently, by assuming the validity of propositions, we provide the proof of Theorem 1.1 in Section 3.3, followed by the subsequent proofs of Theorems 2.1 and 2.2 in Section 3.4. Finally, we prove Propositions 3.1, 3.2, 3.3, and 3.4 in Sections 4, 5, 6, and 7, respectively.

2 Further Discussions

In this section, we explore some additional related topics. We first present two supplementary theorems, namely Theorems 2.1 and 2.2, which complement Theorem 1.1 and are covered in detail in Section 2.1. Additionally, we discuss related work in Section 2.2 and provide comprehensive explanations and visual examples of commonly used activation functions in Section 2.3.

2.1 Additional Results

It is important to note that Theorem 1.1 specifically focuses on activation functions $\varrho \in \mathcal{A}_{1,k}$ with $k = 0, 1, 2, 3, 4$. However, we can also obtain similar results for larger values of $k \in \mathbb{N}$, where $\varrho \in \mathcal{A}_{1,k}$ exhibits even smoother properties. In particular, we establish that for any $\varrho \in C^k(\mathbb{R})$ with $k \in \mathbb{N}$, a $\varrho^{(k)}$ -activated network of width N and depth L can be approximated to arbitrary precision by a ϱ -activated network of width $(k+1)N$ and depth L on any bounded set.

Theorem 2.1. Given any $k \in \mathbb{N}$ and $\varrho \in C^k(\mathbb{R})$, suppose $\phi_{\varrho^{(k)}} \in \mathcal{NN}_{\varrho^{(k)}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ with $N, L, d, n \in \mathbb{N}^+$. Then for any $\varepsilon > 0$ and $A > 0$, there exists $\phi_{\varrho} \in \mathcal{NN}_{\varrho}\{(k+1)N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ such that

$$\|\phi_{\varrho} - \phi_{\varrho^{(k)}}\|_{\sup([-A, A]^d)} < \varepsilon.$$

Furthermore, the following theorem specifically addresses $\varrho \in \mathcal{A}_{1,k}$ for any $k \in \mathbb{N}$. Specifically, we demonstrate that for any $\varrho \in \mathcal{A}_{1,k}$ with $k \in \mathbb{N}$, a ReLU network of width N and depth L can be approximated with arbitrary precision by a ϱ -activated network of width $(k+2)N$ and depth L on any bounded set.

Theorem 2.2. Suppose $\phi_{\text{ReLU}} \in \mathcal{NN}_{\text{ReLU}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ with $N, L, d, n \in \mathbb{N}^+$. Then for any $\varepsilon > 0$, $A > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathcal{A}_{1,k}$, there exists $\phi_{\varrho} \in \mathcal{NN}_{\varrho}\{(k+2)N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ such that

$$\|\phi_{\varrho} - \phi_{\text{ReLU}}\|_{\sup([-A, A]^d)} < \varepsilon.$$

The proofs of Theorems 2.1 and 2.2 are placed in Section 3.

2.2 Related Work

Extensive research has been conducted to explore the approximation capabilities of neural networks, and a multitude of publications have focused on the construction of various neural network architectures to approximate a wide range of target functions. Noteworthy examples of such studies include [1, 2, 4, 7, 8, 10, 14, 15, 18, 22, 23, 26, 28, 30, 31, 32, 33, 37, 40, 41, 42, 45]. During the early stages of this field, the primary focus was on investigating the universal approximation capabilities of single-hidden-layer networks. The universal approximation theorem [10, 17, 18] demonstrated that when a neural network is sufficiently large, it can approximate a particular type of target function with arbitrary precision, without explicitly quantifying the approximation error in relation to the size of the network. Subsequent research, exemplified by [2, 3], delved into analyzing the approximation error of single-hidden-layer networks with a width of n . These studies demonstrated an asymptotic approximation error of $\mathcal{O}(n^{-1/2})$ in the L^2 -norm for target functions possessing certain smoothness properties.

In recent years, the most widely used and effective activation function is ReLU. The adoption of ReLU has marked a significant improvement of results on challenging datasets in supervised learning [20]. Optimizing deep networks activated by ReLU is comparatively simpler than networks utilizing other activation functions such as Sigmoid or Tanh, since gradients can propagate when the input to ReLU is positive. The effectiveness and simplicity of ReLU have positioned it as the preferred default activation function in the deep learning community. Extensive research has investigated the expressive capabilities of deep neural networks, with a majority of studies focusing on the ReLU activation function [23, 30, 31, 34, 40, 41, 42, 43]. In recent advancements, several alternative activation functions have emerged as potential replacements for ReLU. Section 1 provides numerous examples of these alternatives. Although these newly proposed activation functions have shown promising empirical results, their theoretical foundations are still being developed. The objective of this paper is to explore the expressive capabilities of deep neural networks using these activation functions. By establishing connections between these functions and ReLU, we aim to expand most existing approximation results for ReLU networks to encompass a wide range of activation functions.

2.3 Definitions and Illustrations of Common Activation Functions

We will provide definitions and visual representations of activation functions mentioned in Section 1, including ReLU, LeakyReLU, ReLU², ELU, SELU, Softplus, GELU, SiLU, Swish, Mish, Sigmoid, Tanh, Arctan, Softsign, dSiLU, and SRS. The definitions of these sixteen activation functions are presented below. The first five activation functions are given by

$$\begin{aligned} \text{ReLU}(x) &= \max\{0, x\}, & \text{LeakyReLU}(x) &= \begin{cases} x & \text{for } x \geq 0 \\ \alpha x & \text{for } x < 0, \end{cases} \\ \text{ReLU}^2(x) &= \max\{0, x^2\}, & \text{ELU}(x) &= \begin{cases} x & \text{for } x \geq 0 \\ \alpha(e^x - 1) & \text{for } x < 0 \end{cases} \quad \text{with } \alpha \in \mathbb{R}, \end{aligned}$$

228 and

$$229 \quad \text{SELU}(x) = \lambda \begin{cases} x & \text{for } x \geq 0 \\ \alpha(e^x - 1) & \text{for } x < 0 \end{cases} \quad \text{with } \lambda \in (0, \infty) \text{ and } \alpha \in \mathbb{R},$$

230 where e is the base of the natural logarithm. For the last six activation functions, **Arctan** is
 231 the inverse tangent function and the other five activation functions are given by

$$232 \quad \text{Sigmoid}(x) = \frac{1}{1 + e^{-x}}, \quad \text{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \text{Softsign}(x) = \frac{x}{1 + |x|},$$

233

$$234 \quad \text{dSiLU}(x) = \frac{1 + e^{-x} + xe^{-x}}{(1 + e^{-x})^2}, \quad \text{and} \quad \text{SRS}(x) = \frac{x}{x/\alpha + e^{-x/\beta}} \quad \text{with } \alpha, \beta \in (0, \infty).$$

235 The remaining five activation functions are given by

$$236 \quad \text{Softplus}(x) = \ln(1 + e^x), \quad \text{SiLU}(x) = \frac{x}{1 + e^{-x}},$$

237

$$238 \quad \text{Swish}(x) = \frac{x}{1 + e^{-\beta x}} \quad \text{with } \beta \in (0, \infty), \quad \text{Mish}(x) = x \cdot \text{Tanh}(\text{Softplus}(x)),$$

239 and

$$240 \quad \text{GELU}(x) = x \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt \quad \text{with } \mu \in \mathbb{R} \text{ and } \sigma \in (0, \infty).$$

241 Refer to Figure 1 for visual representations of all these activation functions.

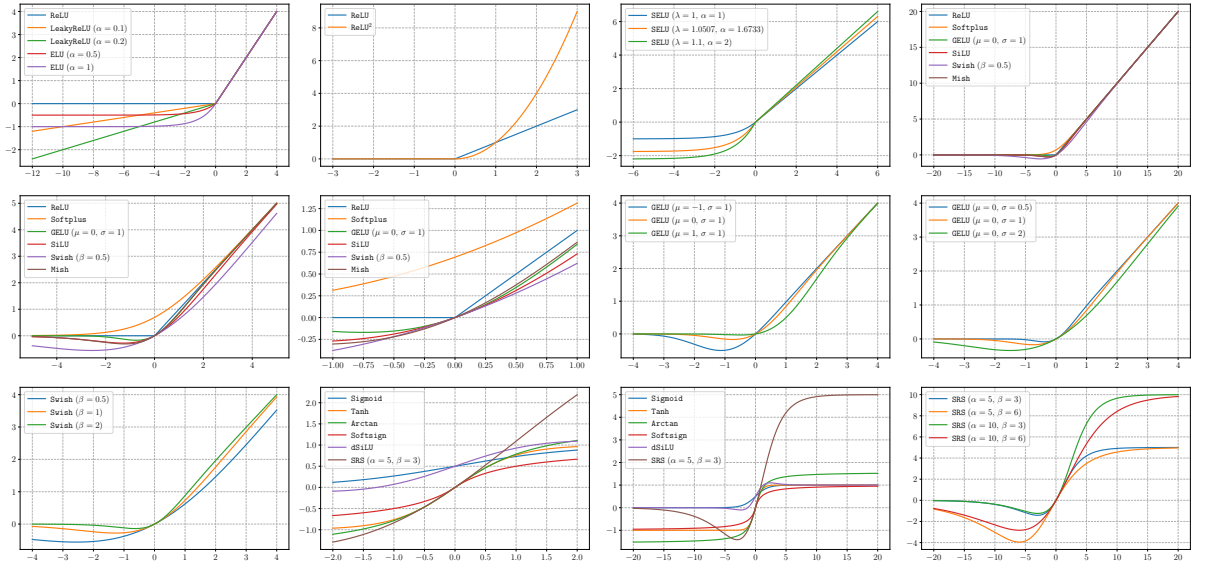


Figure 1: Illustrations of ReLU, LeakyReLU, ReLU², ELU, SELU, Softplus, GELU, SiLU, Swish, Mish, Sigmoid, Tanh, Arctan, Softsign, dSiLU, and SRS.

242 3 Proofs of Theorems in Sections 1 and 2

243 In this section, we will prove the theorems in Sections 1 and 2, i.e., Theorems 1.1, 2.1, and 2.2.
 244 To enhance clarity, Section 3.1 offers a concise overview of the notations employed throughout
 245 this paper. Next in Section 3.2, we present the ideas for proving Theorems 1.1, 2.1, and 2.2.
 246 Moreover, to simplify the proofs, we establish several propositions, which will be proved in later
 247 sections. By assuming the validity of these propositions, we provide the proof of Theorem 1.1
 248 in Section 3.3 and give the proofs of Theorems 2.1 and 2.2 in Section 3.4.

249 3.1 Notations

250 The following is an overview of the basic notations used in this paper.

- 251 • The set difference of two sets A and B is denoted as $A \setminus B := \{x : x \in A, x \notin B\}$.
- 252 • The symbols \mathbb{N} , \mathbb{Z} , \mathbb{Q} , and \mathbb{R} are used to denote the sets of natural numbers (including
253 0), integers, rational numbers, and real numbers, respectively. The set of positive natural
254 numbers is denoted as $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$.
- 255 • The base of the natural logarithm is denoted as e , i.e., $e = \lim_{n \rightarrow \infty} (1 + \frac{1}{n})^n \approx 2.71828$.
- 256 • The indicator (or characteristic) function of a set A , denoted by $\mathbf{1}_A$, is a function that
257 takes the value 1 for elements of A and 0 for elements not in A .
- 258 • The floor and ceiling functions of a real number x can be represented as $\lfloor x \rfloor = \max\{n :$
259 $n \leq x, n \in \mathbb{Z}\}$ and $\lceil x \rceil = \min\{n : n \geq x, n \in \mathbb{Z}\}$.
- 260 • Let $\binom{n}{k}$ denote the coefficient of the x^k term in the polynomial expansion of the binomial
261 power $(1+x)^n$ for any $n, k \in \mathbb{N}$ with $n \geq k$, i.e., $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.
- 262 • Vectors are denoted by bold lowercase letters, such as $\mathbf{a} = (a_1, \dots, a_d) \in \mathbb{R}^d$. On the
263 other hand, matrices are represented by bold uppercase letters. For example, $\mathbf{A} \in \mathbb{R}^{m \times n}$
264 refers to a real matrix of size $m \times n$, and \mathbf{A}^T denotes the transpose of matrix \mathbf{A} .
- 265 • Given any $p \in [1, \infty]$, the p -norm (also known as ℓ^p -norm) of a vector $\mathbf{x} = (x_1, \dots, x_d) \in$
266 \mathbb{R}^d is defined via

$$267 \quad \|\mathbf{x}\|_p = \|\mathbf{x}\|_{\ell^p} := (|x_1|^p + \dots + |x_d|^p)^{1/p} \quad \text{if } p \in [1, \infty)$$

268 and

$$269 \quad \|\mathbf{x}\|_\infty = \|\mathbf{x}\|_{\ell^\infty} := \max \{|x_i| : i = 1, 2, \dots, d\}.$$

- 270 • Let “ \Rightarrow ” denote the uniform convergence. For example, if $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is a vector-valued
271 function and $\mathbf{f}_\delta(\mathbf{x}) \Rightarrow \mathbf{f}(\mathbf{x})$ as $\delta \rightarrow 0^+$ for any $\mathbf{x} \in \Omega \subseteq \mathbb{R}^d$, then for any $\varepsilon > 0$, there
272 exists $\delta_\varepsilon \in (0, 1)$ such that

$$273 \quad \sup_{\mathbf{x} \in \Omega} \|\mathbf{f}_\delta(\mathbf{x}) - \mathbf{f}(\mathbf{x})\|_{\ell^\infty} < \varepsilon \quad \text{for any } \delta \in (0, \delta_\varepsilon).$$

- 274 • A network is labeled as “a network of width N and depth L ” when it satisfies the following
275 two conditions.
 - 276 – The count of neurons in each hidden layer of the network does not exceed N .
 - 277 – The total number of hidden layers in the network is at most L .
- 278 • Suppose $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is a vector-valued function realized by a ϱ -activated network. Then
279 ϕ can be expressed as

$$280 \quad \mathbf{x} = \tilde{\mathbf{h}}_0 \xrightarrow[\mathcal{L}_0]{\mathbf{W}_0, \mathbf{b}_0} \mathbf{h}_1 \xrightarrow{\varrho} \tilde{\mathbf{h}}_1 \quad \dots \quad \xrightarrow[\mathcal{L}_{L-1}]{\mathbf{W}_{L-1}, \mathbf{b}_{L-1}} \mathbf{h}_L \xrightarrow{\varrho} \tilde{\mathbf{h}}_L \xrightarrow[\mathcal{L}_L]{\mathbf{W}_L, \mathbf{b}_L} \mathbf{h}_{L+1} = \phi(\mathbf{x}),$$

281 where $N_0 = d$, $N_1, N_2, \dots, N_L \in \mathbb{N}^+$, $N_{L+1} = n$, $\mathbf{W}_i \in \mathbb{R}^{N_{i+1} \times N_i}$ and $\mathbf{b}_i \in \mathbb{R}^{N_{i+1}}$ are the
282 weight matrix and the bias vector in the i -th affine linear map \mathcal{L}_i , respectively, i.e.,

$$283 \quad \mathbf{h}_{i+1} = \mathbf{W}_i \cdot \tilde{\mathbf{h}}_i + \mathbf{b}_i =: \mathcal{L}_i(\tilde{\mathbf{h}}_i) \quad \text{for } i = 0, 1, \dots, L,$$

and

$$\tilde{\mathbf{h}}_i = \varrho(\mathbf{h}_i) \quad \text{for } i = 1, 2, \dots, L,$$

where ϱ is the activation function that can be applied elementwise to a vector input. Clearly, $\phi \in \mathcal{NN}_\varrho\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, where $N = \max\{N_1, N_2, \dots, N_L\}$. Furthermore, ϕ can be expressed as a composition of functions

$$\phi = \mathcal{L}_L \circ \varrho \circ \mathcal{L}_{L-1} \circ \dots \circ \varrho \circ \mathcal{L}_1 \circ \varrho \circ \mathcal{L}_0.$$

Refer to Figure 2 for an illustration.

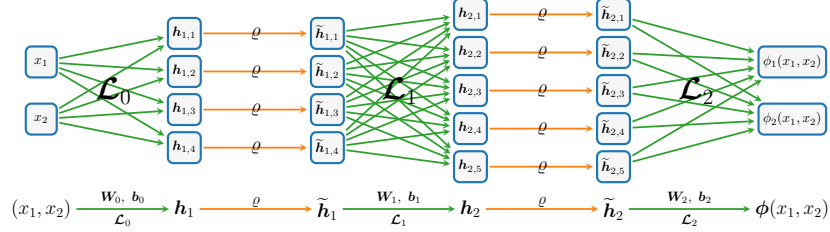


Figure 2: An example of a ϱ -activated network of width 5 and depth 2. The network realizes a vector-valued function $\phi = (\phi_1, \phi_2)$.

3.2 Propositions for Proving Theorems in Sections 1 and 2

We now present the key ideas for proving theorems introduced in Sections 1 and 2, i.e., Theorems 1.1, 2.1, and 2.2. These three theorems collectively convey a narrative wherein a $\tilde{\varrho}$ -activated network can be accurately approximated by a ϱ -activated network, provided certain assumptions are met regarding ϱ and $\tilde{\varrho}$. Consequently, it becomes imperative to establish an auxiliary theorem that allows for the substitution of the network's activation function at the cost of a sufficiently small error.

Proposition 3.1. *Given two functions $\varrho, \tilde{\varrho} : \mathbb{R} \rightarrow \mathbb{R}$ with $\tilde{\varrho} \in C(\mathbb{R})$, suppose for any $M > 0$, there exists $\tilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{\tilde{N}, \tilde{L}; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\eta \in (0, 1)$ such that*

$$\tilde{\varrho}_\eta(x) \rightrightarrows \tilde{\varrho}(x) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Assuming $\phi_{\tilde{\varrho}} \in \mathcal{NN}_{\tilde{\varrho}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, for any $\varepsilon > 0$ and $A > 0$, there exists $\phi_\varrho \in \mathcal{NN}_\varrho\{\tilde{N} \cdot N, \tilde{L} \cdot L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ such that

$$\|\phi_\varrho - \phi_{\tilde{\varrho}}\|_{\sup([-A, A]^d)} < \varepsilon.$$

The proof of Proposition 3.1 can be found in Section 4. The utilization of Proposition 3.1 simplifies our task of proving Theorems 1.1, 2.1, and 2.2. Our focus now shifts to constructing ϱ -activated networks that can effectively approximate both $\varrho^{(k)}$ (assuming $\varrho \in C^k(\mathbb{R})$) and ReLU. To facilitate this construction process, we introduce the following three propositions.

Proposition 3.2. *Given any $n \in \mathbb{N}$ and $a_0 < a < b < b_0$, if $f \in C^n((a_0, b_0))$, then*

$$\frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} f(x + \ell t)}{(-t)^n} \rightrightarrows f^{(n)}(x) \quad \text{as } t \rightarrow 0 \quad \text{for any } x \in [a, b].$$

Proposition 3.3. *Given any $M > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathcal{A}_{1,k}$, there exists $\phi_\varepsilon \in \mathcal{NN}_\varrho\{k+2, 1; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\varepsilon \in (0, 1)$ such that*

$$\phi_\varepsilon(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Proposition 3.4. *Given any $M > 0$ and $\varrho \in \mathcal{A}_2 \cup \mathcal{A}_3$, there exists $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\varepsilon \in (0, 1)$ such that*

$$\phi_\varepsilon(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Propositions 3.2, 3.3, and 3.4 will be proved in Sections 5, 6, and 7, respectively. Let us briefly discuss the key ideas for proving these three propositions.

The essence of proving Proposition 3.2 lies in the application of Cauchy's Mean Value Theorem. Through repeated utilization of such a theorem, we can establish the existence of $|t_n| \in (0, |t|)$ such that

$$\frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} f(x + \ell t)}{(-t)^n} = \frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^n f^{(n)}(x + \ell t_n)}{(-1)^n n!}.$$

Furthermore, we will demonstrate $\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^n = (-1)^n n!$ in Lemma 5.1 later. With the uniform continuity of $f^{(n)}$ on a closed interval, Proposition 3.2 follows straightforwardly. See more details in Section 5.

The proof of Proposition 3.3 can be divided into two main steps. The first step involves demonstrating that

$$\frac{\varrho^{(k)}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} \rightrightarrows \tau(x) := \begin{cases} L_2 x & \text{for } x \geq 0 \\ L_1 x & \text{for } x < 0 \end{cases} \quad \text{for any } x \in [-A, A] \text{ and } A > 0,$$

where τ can be used to generate ReLU and

$$L_1 = \lim_{t \rightarrow 0^-} \frac{\varrho^{(k)}(x_0 + t) - \varrho^{(k)}(x_0)}{t} \neq L_2 = \lim_{t \rightarrow 0^+} \frac{\varrho^{(k)}(x_0 + t) - \varrho^{(k)}(x_0)}{t}.$$

The second step involves employing Proposition 3.2 to approximate $\varrho^{(k)}$ using a ϱ -activated network. By combining these two steps, we can construct a ϱ -activated network that effectively approximates ReLU. For further details, refer to Section 6.

The core of proving Proposition 3.4 is the fact $x \cdot \mathbb{1}_{\{x>0\}} = \text{ReLU}(x)$ for any $x \in \mathbb{R}$. This fact simplifies our proof considerably. Our focus then shifts toward constructing ϱ -activated networks that can effectively approximate x , $\mathbb{1}_{\{x>0\}}$, and xy for any $x, y \in [-A, A]$ and $A > 0$. Additional details can be found in Section 7.

3.3 Proof of Theorem 1.1 with Propositions

The proof of Theorem 1.1 can be easily demonstrated by employing Propositions 3.1, 3.3, and 3.4.

Proof of Theorem 1.1. Since $\mathcal{A} = (\cup_{k=0}^4 \mathcal{A}_{1,k}) \cup \mathcal{A}_2 \cup \mathcal{A}_3$, we can divide the proof into two cases: $\varrho \in \cup_{k=0}^4 \mathcal{A}_{1,k}$ and $\varrho \in \mathcal{A}_2 \cup \mathcal{A}_3$.

We first consider the case $\varrho \in \cup_{k=0}^4 \mathcal{A}_{1,k}$, i.e., $\varrho \in \mathcal{A}_{1,k}$ for some $k \in \{0, 1, 2, 3, 4\}$. By Proposition 3.3, for any $M > 0$, there exist $\tilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{k+2, 1; \mathbb{R} \rightarrow \mathbb{R}\} \subseteq \mathcal{NN}_\varrho\{6, 1; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\eta \in (0, 1)$ such that

$$\tilde{\varrho}_\eta(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Then by Proposition 3.1 with $\tilde{\varrho}$ being ReLU therein, for any $\varepsilon > 0$, $A > 0$, and $\phi_{\text{ReLU}} \in \mathcal{NN}_{\text{ReLU}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, there exists

$$\phi_\varrho \in \mathcal{NN}_\varrho\{6N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\} \subseteq \mathcal{NN}_\varrho\{6N, 2L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$$

such that

$$\|\phi_\varrho - \phi_{\text{ReLU}}\|_{\sup([-A,A]^d)} < \varepsilon.$$

Next, we consider the case $\varrho \in \mathcal{A}_2 \cup \mathcal{A}_3$. By Proposition 3.4, for any $M > 0$, there exist $\tilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\eta \in (0, 1)$ such that

$$\tilde{\varrho}_\eta(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Then by Proposition 3.1 with $\tilde{\varrho}$ being ReLU therein, for any $\varepsilon > 0$, $A > 0$, and $\phi_{\text{ReLU}} \in \mathcal{NN}_{\text{ReLU}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, there exists

$$\phi_\varrho \in \mathcal{NN}_\varrho\{6N, 2L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$$

such that

$$\|\phi_\varrho - \phi_{\text{ReLU}}\|_{\sup([-A,A]^d)} < \varepsilon.$$

So we finish the proof of Theorem 1.1. \square

3.4 Proofs of Theorems 2.1 and 2.2 with Propositions

The proofs of Theorems 2.1 and 2.2 can be straightforwardly demonstrated by utilizing Propositions 3.1, 3.2, and 3.3.

Proof of Theorem 2.1. It follows from $\varrho \in C^k(\mathbb{R})$ that $\varrho \in C^k((-M-1, M+1))$ for any $M > 0$. By Proposition 3.3, we have

$$\frac{\sum_{\ell=0}^k (-1)^\ell \binom{k}{\ell} \varrho(x + \ell t)}{(-t)^k} \rightrightarrows \varrho^{(k)}(x) \quad \text{as } t \rightarrow 0 \quad \text{for any } x \in [M, M].$$

For each $\eta \in (0, 1)$, we define

$$\tilde{\varrho}_\eta(x) := \frac{\sum_{\ell=0}^k (-1)^\ell \binom{k}{\ell} \varrho(x + \ell \eta)}{(-\eta)^k} \quad \text{for any } x \in \mathbb{R}.$$

Clearly, $\tilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{k+1, 1; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\eta \in (0, 1)$ and

$$\tilde{\varrho}_\eta(x) \rightrightarrows \varrho^{(k)}(x) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Then by Proposition 3.1 with $\tilde{\varrho}$ being $\varrho^{(k)}$ therein, for any $\varepsilon > 0$, $A > 0$, and $\phi_{\varrho^{(k)}} \in \mathcal{NN}_{\varrho^{(k)}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, there exists $\phi_\varrho \in \mathcal{NN}_\varrho\{(k+1)N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ such that

$$\|\phi_\varrho - \phi_{\varrho^{(k)}}\|_{\sup([-A,A]^d)} < \varepsilon.$$

So we finish the proof of Theorem 2.1. \square

Proof of Theorem 2.2. By Proposition 3.3, for any $M > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathcal{A}_{1,k}$, there exist $\tilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{k+2, 1; \mathbb{R} \rightarrow \mathbb{R}\}$ for each $\eta \in (0, 1)$ such that

$$\tilde{\varrho}_\eta(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

Then by Proposition 3.1 with $\tilde{\varrho}$ being ReLU therein, for any $\varepsilon > 0$, $A > 0$, and $\phi_{\text{ReLU}} \in \mathcal{NN}_{\text{ReLU}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, there exists $\phi_\varrho \in \mathcal{NN}_\varrho\{(k+2)N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$ such that

$$\|\phi_\varrho - \phi_{\text{ReLU}}\|_{\sup([-A,A]^d)} < \varepsilon.$$

So we finish the proof of Theorem 2.2. \square

4 Proof of Proposition 3.1

We will prove Proposition 3.1 in this section. The crucial aspect of the proof is the observation that $\tilde{\varrho} \in C(\mathbb{R})$ implies $\tilde{\varrho}$ is uniformly continuous on $[-M, M]$ for any $M > 0$. Further information and specific details are provided below.

Proof of Proposition 3.1. For ease of notation, we allow the activation function to be applied elementwise to a vector input. Since $\phi_{\tilde{\varrho}} \in \mathcal{NN}_{\tilde{\varrho}}\{N, L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$, $\phi_{\tilde{\varrho}}$ is realized by a \hat{L} -hidden-layer $\tilde{\varrho}$ -activated network, where $L \geq \hat{L} \in \mathbb{N}^+$. We may assume $\hat{L} = L$ since the proof remains similar if we replace L with \hat{L} when $\hat{L} < L$. Then $\phi_{\tilde{\varrho}}$ can be represented in a form of function compositions

$$\phi_{\tilde{\varrho}}(\mathbf{x}) = \mathcal{L}_L \circ \tilde{\varrho} \circ \mathcal{L}_{L-1} \circ \cdots \circ \tilde{\varrho} \circ \mathcal{L}_1 \circ \tilde{\varrho} \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d,$$

where $N_0 = d$, $N_1, N_2, \dots, N_L \in \mathbb{N}^+$ with $\max\{N_1, N_2, \dots, N_L\} \leq N$, $N_{L+1} = n$, $\mathbf{W}_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ and $\mathbf{b}_\ell \in \mathbb{R}^{N_{\ell+1}}$ are the weight matrix and the bias vector in the ℓ -th affine linear transform $\mathcal{L}_\ell : \mathbf{y} \mapsto \mathbf{W}_\ell \cdot \mathbf{y} + \mathbf{b}_\ell$ for each $\ell \in \{0, 1, \dots, L\}$.

Recall that there exists

$$\tilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{\tilde{N}, \tilde{L}; \mathbb{R} \rightarrow \mathbb{R}\} \quad \text{for each } \eta \in (0, 1)$$

such that

$$\tilde{\varrho}_\eta(t) \rightrightarrows \tilde{\varrho}(t) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } t \in [-M, M],$$

where $M > 0$ is a large number determined later. For each $\eta \in (0, 1)$, we define

$$\phi_{\tilde{\varrho}_\eta}(\mathbf{x}) := \mathcal{L}_L \circ \tilde{\varrho}_\eta \circ \mathcal{L}_{L-1} \circ \cdots \circ \tilde{\varrho}_\eta \circ \mathcal{L}_1 \circ \tilde{\varrho}_\eta \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d.$$

It is easy to verify that

$$\phi_{\tilde{\varrho}_\eta} \in \mathcal{NN}_\varrho\{\tilde{N} \cdot N, \tilde{L} \cdot L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}.$$

Moreover, we will prove

$$\phi_{\tilde{\varrho}_\eta}(\mathbf{x}) \rightrightarrows \phi_{\tilde{\varrho}}(\mathbf{x}) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

For each $\eta \in (0, 1)$ and $\ell = 1, 2, \dots, L+1$, we define

$$\mathbf{h}_\ell(\mathbf{x}) := \mathcal{L}_{\ell-1} \circ \tilde{\varrho} \circ \mathcal{L}_{\ell-2} \circ \cdots \circ \tilde{\varrho} \circ \mathcal{L}_1 \circ \tilde{\varrho} \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d$$

and

$$\mathbf{h}_{\ell, \eta}(\mathbf{x}) := \mathcal{L}_{\ell-1} \circ \tilde{\varrho}_\eta \circ \mathcal{L}_{\ell-2} \circ \cdots \circ \tilde{\varrho}_\eta \circ \mathcal{L}_1 \circ \tilde{\varrho}_\eta \circ \mathcal{L}_0(\mathbf{x}) \quad \text{for any } \mathbf{x} \in \mathbb{R}^d.$$

Note that \mathbf{h}_ℓ and $\mathbf{h}_{\ell, \eta}$ are two maps from \mathbb{R}^d to \mathbb{R}^{N_ℓ} for each $\eta \in (0, 1)$ and $\ell = 1, 2, \dots, L+1$.

For $\ell = 1, 2, \dots, L+1$, we will prove by induction that

$$\mathbf{h}_{\ell, \eta}(\mathbf{x}) \rightrightarrows \mathbf{h}_\ell(\mathbf{x}) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d. \quad (1)$$

First, we consider the case $\ell = 1$. Clearly,

$$\mathbf{h}_{1, \eta}(\mathbf{x}) = \mathcal{L}_0(\mathbf{x}) = \mathbf{h}_1(\mathbf{x}) \rightrightarrows \mathbf{h}_1(\mathbf{x}) \quad \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

This means Equation (1) holds for $\ell = 1$.

Next, supposing Equation (1) holds for $\ell = i \in \{1, 2, \dots, L\}$, our goal is to prove that it also holds for $\ell = i + 1$. Determine $M > 0$ via

$$M = \sup \left\{ \|\mathbf{h}_j(\mathbf{x})\|_\infty + 1 : \mathbf{x} \in [-A, A]^d, \quad j = 1, 2, \dots, L+1 \right\},$$

where the continuity of $\tilde{\varrho}$ guarantees the above supremum is finite, i.e., $M \in [1, \infty)$. By the induction hypothesis, we have

$$\mathbf{h}_{i,\eta}(\mathbf{x}) \rightrightarrows \mathbf{h}_i(\mathbf{x}) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

Clearly, for any $\mathbf{x} \in [-A, A]^d$, we have $\|\mathbf{h}_i(\mathbf{x})\|_{\ell^\infty} \leq M$ and

$$\|\mathbf{h}_{i,\eta}(\mathbf{x})\|_{\ell^\infty} \leq \|\mathbf{h}_i(\mathbf{x})\|_{\ell^\infty} + 1 \leq M \quad \text{for small } \eta > 0.$$

Recall that $\tilde{\varrho}_\eta(t) \rightrightarrows \tilde{\varrho}(t)$ as $\eta \rightarrow 0^+$ for any $t \in [-M, M]$. Then, we have

$$\tilde{\varrho}_\eta \circ \mathbf{h}_{i,\eta}(\mathbf{x}) - \tilde{\varrho} \circ \mathbf{h}_{i,\eta}(\mathbf{x}) \rightrightarrows \mathbf{0} \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

The continuity of $\tilde{\varrho}$ implies the uniform continuity of $\tilde{\varrho}$ on $[-M, M]$, from which we deduce

$$\tilde{\varrho} \circ \mathbf{h}_{i,\eta}(\mathbf{x}) - \tilde{\varrho} \circ \mathbf{h}_i(\mathbf{x}) \rightrightarrows \mathbf{0} \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

Therefore, for any $\mathbf{x} \in [-A, A]^d$, as $\eta \rightarrow 0^+$, we have

$$\tilde{\varrho}_\eta \circ \mathbf{h}_{i,\eta}(\mathbf{x}) - \tilde{\varrho} \circ \mathbf{h}_i(\mathbf{x}) = \underbrace{\tilde{\varrho}_\eta \circ \mathbf{h}_{i,\eta}(\mathbf{x}) - \tilde{\varrho} \circ \mathbf{h}_{i,\eta}(\mathbf{x})}_{\rightrightarrows \mathbf{0}} + \underbrace{\tilde{\varrho} \circ \mathbf{h}_{i,\eta}(\mathbf{x}) - \tilde{\varrho} \circ \mathbf{h}_i(\mathbf{x})}_{\rightrightarrows \mathbf{0}} \rightrightarrows \mathbf{0},$$

implying

$$\mathbf{h}_{i+1,\eta}(\mathbf{x}) = \mathcal{L}_i \circ \tilde{\varrho}_\eta \circ \mathbf{h}_{i,\eta}(\mathbf{x}) \rightrightarrows \mathcal{L}_i \circ \tilde{\varrho} \circ \mathbf{h}_i(\mathbf{x}) = \mathbf{h}_{i+1}(\mathbf{x}).$$

This means Equation (1) holds for $\ell = i + 1$. So we complete the inductive step.

By the principle of induction, we have

$$\phi_{\tilde{\varrho}_\eta}(\mathbf{x}) = \mathbf{h}_{L+1,\eta}(\mathbf{x}) \rightrightarrows \mathbf{h}_{L+1}(\mathbf{x}) = \phi_{\tilde{\varrho}}(\mathbf{x}) \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } \mathbf{x} \in [-A, A]^d.$$

Then for any $\varepsilon > 0$, there exists a small $\eta_0 > 0$ such that

$$\|\phi_{\tilde{\varrho}_{\eta_0}} - \phi_{\tilde{\varrho}}\|_{\sup([-A, A]^d)} < \varepsilon.$$

By defining $\phi_\varrho := \phi_{\tilde{\varrho}_{\eta_0}}$, we have

$$\phi_\varrho = \phi_{\tilde{\varrho}_{\eta_0}} \in \mathcal{NN}_\varrho\{\tilde{N} \cdot N, \tilde{L} \cdot L; \mathbb{R}^d \rightarrow \mathbb{R}^n\}$$

and

$$\|\phi_\varrho - \phi_{\tilde{\varrho}}\|_{\sup([-A, A]^d)} = \|\phi_{\tilde{\varrho}_{\eta_0}} - \phi_{\tilde{\varrho}}\|_{\sup([-A, A]^d)} < \varepsilon.$$

So we finish the proof of Proposition 3.1. □

5 Proof of Proposition 3.2

In this section, our goal is to prove Proposition 3.2. To facilitate the proof, we first introduce a lemma in Section 5.1 that simplifies the process. Subsequently, we provide the detailed proof in Section 5.2.

438 5.1 A Lemma for Proving Proposition 3.2

439 **Lemma 5.1.** *Given any $n \in \mathbb{N}$, it holds that*

$$440 \quad \sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^i = \begin{cases} 0 & \text{if } i \in \{0, 1, \dots, n-1\}, \\ (-1)^n n! & \text{if } i = n. \end{cases}$$

441 *Proof.* To simplify the proof, we claim that there exists a polynomial p_i for each $i \in \{0, 1, \dots, n\}$
442 such that

$$443 \quad \sum_{\ell=0}^n t^\ell \binom{n}{\ell} \ell^i = (1+t)^{n-i} \left(\frac{n!}{(n-i)!} t^i + (1+t)p_i(t) \right) \quad \text{for any } t \in (-1, 0).$$

444 By assuming the validity of the claim, we have

$$445 \quad \begin{aligned} \sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^i &= \lim_{t \rightarrow -1^+} \sum_{\ell=0}^n t^\ell \binom{n}{\ell} \ell^i = \lim_{t \rightarrow -1^+} (1+t)^{n-i} \left(\frac{n!}{(n-i)!} t^i + (1+t)p_i(t) \right) \\ &= \begin{cases} 0 & \text{if } i \in \{0, 1, \dots, n-1\}, \\ (-1)^n n! & \text{if } i = n. \end{cases} \end{aligned}$$

446 It remains to prove the claim and we will establish its validity by induction.

447 First, we consider the case $i = 0$. Clearly,

$$448 \quad \sum_{\ell=0}^n t^\ell \binom{n}{\ell} \ell^0 = \sum_{\ell=0}^n t^\ell \binom{n}{\ell} = (1+t)^n = (1+t)^{n-0} \left(\frac{n!}{(n-0)!} t^0 + (1+t) \cdot p_0(t) \right)$$

449 for any $t \in (-1, 0)$, where $p_0(t) = 0$. That means the claim holds for $i = 0$.

450 Next, assuming the claim holds for $i = j \in \{0, 1, \dots, n-1\}$, we will show it also holds for
451 $i = j+1$. By the induction hypothesis, we have

$$452 \quad \sum_{\ell=0}^n t^\ell \binom{n}{\ell} \ell^j = (1+t)^{n-j} \underbrace{\left(\frac{n!}{(n-j)!} t^j + (1+t)p_j(t) \right)}_{\tilde{p}_j(t)} = (1+t)^{n-j} \tilde{p}_j(t)$$

453 for any $t \in (-1, 0)$, where $\tilde{p}_j(t) = \frac{n!}{(n-j)!} t^j + (1+t)p_j(t)$ is a polynomial. By differentiating
454 both sides of the equation above, we obtain

$$455 \quad \begin{aligned} \sum_{\ell=0}^n \ell t^{\ell-1} \binom{n}{\ell} \ell^j &= (n-j)(1+t)^{n-j-1} \tilde{p}_j(t) + (1+t)^{n-j} \frac{d}{dt} \tilde{p}_j(t) \\ &= (1+t)^{n-j-1} \left((n-j) \tilde{p}_j(t) + (1+t) \frac{d}{dt} \tilde{p}_j(t) \right) \end{aligned}$$

for any $t \in (-1, 0)$, implying

$$\begin{aligned}
\sum_{\ell=0}^n t^\ell \binom{n}{\ell} \ell^{j+1} &= t \sum_{\ell=0}^n \ell t^{\ell-1} \binom{n}{\ell} \ell^j = t(1+t)^{n-j-1} \left((n-j) \tilde{p}_j(t) + (1+t) \frac{d}{dt} \tilde{p}_j(t) \right) \\
&= (1+t)^{n-j-1} \left(t(n-j) \tilde{p}_j(t) + t(1+t) \frac{d}{dt} \tilde{p}_j(t) \right) \\
&= (1+t)^{n-(j+1)} \left(t(n-j) \underbrace{\left(\frac{n!}{(n-j)!} t^j + (1+t) p_j(t) \right)}_{\tilde{p}_j(t)} + t(1+t) \frac{d}{dt} \tilde{p}_j(t) \right) \\
&= (1+t)^{n-(j+1)} \left(\frac{n!(n-j)}{(n-j)!} t^{j+1} + t(n-j)(1+t) p_j(t) + t(1+t) \frac{d}{dt} \tilde{p}_j(t) \right) \\
&= (1+t)^{n-(j+1)} \left(\frac{n!}{(n-(j+1))!} t^{j+1} + (1+t) \underbrace{\left(t(n-j) p_j(t) + t \frac{d}{dt} \tilde{p}_j(t) \right)}_{p_{j+1}(t)} \right) \\
&= (1+t)^{n-(j+1)} \left(\frac{n!}{(n-(j+1))!} t^{j+1} + (1+t) p_{j+1}(t) \right),
\end{aligned}$$

for any $t \in (-1, 0)$, where $p_{j+1}(t) = t(n-j)p_j(t) + t \frac{d}{dt} \tilde{p}_j(t)$ is a polynomial. With the completion of the induction step, we have successfully demonstrated the validity of the claim. Thus, we complete the proof of Lemma 5.1. \square

5.2 Proof of Proposition 3.2 with Lemma 5.1

Equipped with Lemma 5.1, we are prepared to demonstrate the proof of Proposition 3.2.

Proof of Proposition 3.2. We may assume $n \in \mathbb{N}^+$ since the case $n = 0$ is trivial. For each $x \in [a, b]$, we define

$$g_x(t) := \sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} f(x + \ell t) \quad \text{for any } t \in (-c_0, c_0),$$

where $c_0 > 0$ is a small number ensuring that $x + \ell t \in (a_0, b_0)$ for $\ell = 0, 1, \dots, n$. For example, we can set

$$c_0 = \min \left\{ \frac{a - a_0}{n + 1}, \frac{b_0 - b}{n + 1} \right\}.$$

It follows from $f \in C^n((a_0, b_0))$ that $f^{(n)}$ is continuous on (a_0, b_0) , implying $f^{(n)}$ is uniformly continuous on $[a - nc_0, b + nc_0] \subseteq (a_0, b_0)$. For any $\varepsilon > 0$, there exists $\delta_0 \in (0, c_0)$ such that

$$|f^{(n)}(x_1) - f^{(n)}(x_2)| < \frac{\varepsilon}{C_n} \quad \text{if } |x_1 - x_2| < n\delta_0 \quad \text{for any } x_1, x_2 \in [a - nc_0, b + nc_0], \quad (2)$$

where $C_n = \sum_{j=0}^n j^n \binom{n}{j}$.

For each $x \in [a, b]$, we have

$$g_x^{(i)}(t) = \sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^i f^{(i)}(x + \ell t) \quad \text{for any } t \in (-c_0, c_0) \text{ and } i = 0, 1, \dots, n,$$

implying

$$g_x^{(i)}(0) = \sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^i f^{(i)}(x) = 0 \quad \text{for } i = 0, 1, \dots, n-1,$$

where the last equality comes from Lemma 5.1.

Then for any $t \in (-\delta_0, 0) \cup (0, \delta_0)$ and each $x \in [a, b]$, by Cauchy's Mean Value Theorem, there exist $0 < |t_{x,n}| < \dots < |t_{x,1}| < |t| < \delta_0$ such that

$$\begin{aligned} \frac{g_x(t)}{t^n} &= \frac{g_x^{(0)}(t) - g_x^{(0)}(0)}{t^n - 0} = \frac{g_x^{(1)}(t_{x,1})}{nt_{x,1}^{n-1}} = \frac{g_x^{(1)}(t_{x,1}) - g_x^{(1)}(0)}{nt_{x,1}^{n-1} - 0} \\ &= \frac{g_x^{(2)}(t_{x,2})}{n(n-1)t_{x,2}^{n-2}} = \frac{g_x^{(2)}(t_{x,2}) - g_x^{(2)}(0)}{n(n-1)t_{x,2}^{n-2} - 0} = \frac{g_x^{(3)}(t_{x,3})}{n(n-1)(n-2)t_{x,3}^{n-3}} = \dots = \frac{g_x^{(n)}(t_{x,n})}{n!}. \end{aligned}$$

Moreover, for any $t \in (-\delta_0, 0) \cup (0, \delta_0)$ and each $x \in [a, b] \subseteq [a - nc_0, b + nc_0]$, we have

$$|(x + \ell t_{x,n}) - x| = |\ell t_{x,n}| \leq |n t_{x,n}| < n\delta_0 < nc_0 \quad \text{and} \quad x + \ell t_{x,n} \in [a - nc_0, b + nc_0],$$

for $\ell = 0, 1, \dots, n$, from which we deduce

$$|f^{(n)}(x + \ell t_{x,n}) - f^{(n)}(x)| < \frac{\varepsilon}{C_n} = \frac{\varepsilon}{\sum_{j=0}^n j^n \binom{n}{j}},$$

where the strict inequality comes from Equation (2).

Set $\lambda_\ell = \frac{(-1)^\ell \binom{n}{\ell} \ell^n}{(-1)^n n!}$ for $\ell = 0, 1, \dots, n$. By Lemma 5.1, we have

$$\sum_{\ell=0}^n \lambda_\ell = \sum_{\ell=0}^n \frac{(-1)^\ell \binom{n}{\ell} \ell^n}{(-1)^n n!} = \frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^n}{(-1)^n n!} = \frac{(-1)^n n!}{(-1)^n n!} = 1.$$

Therefore, for any $t \in (-\delta_0, 0) \cup (0, \delta_0)$ and each $x \in [a, b]$, we have

$$\begin{aligned} &\left| \frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} f(x + \ell t)}{(-t)^n} - f^{(n)}(x) \right| = \left| \frac{g_x(t)}{(-1)^n t^n} - f^{(n)}(x) \right| = \left| \frac{g_x^{(n)}(t_{x,n})}{(-1)^n n!} - f^{(n)}(x) \right| \\ &= \left| \frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \ell^n f^{(n)}(x + \ell t_{x,n})}{(-1)^n n!} - f^{(n)}(x) \right| = \left| \sum_{\ell=0}^n \lambda_\ell f^{(n)}(x + \ell t_{x,n}) - f^{(n)}(x) \right| \\ &= \left| \sum_{\ell=0}^n \lambda_\ell f^{(n)}(x + \ell t_{x,n}) - \sum_{\ell=0}^n \lambda_\ell f^{(n)}(x) \right| = \sum_{\ell=0}^n |\lambda_\ell| \cdot |f^{(n)}(x + \ell t_{x,n}) - f^{(n)}(x)| \\ &< \sum_{\ell=0}^n |\lambda_\ell| \cdot \frac{\varepsilon}{C_n} = \sum_{\ell=0}^n \frac{\ell^n \binom{n}{\ell}}{n!} \cdot \frac{\varepsilon}{\sum_{j=0}^n j^n \binom{n}{j}} \leq \sum_{\ell=0}^n \ell^n \binom{n}{\ell} \cdot \frac{\varepsilon}{\sum_{j=0}^n j^n \binom{n}{j}} = \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we can conclude that

$$\frac{\sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} f(x + \ell t)}{(-t)^n} \Rightarrow f^{(n)}(x) \quad \text{as } t \rightarrow 0 \quad \text{for any } x \in [a, b].$$

So we finish the proof of Proposition 3.2. \square

6 Proof of Proposition 3.3

The objective of this section is to provide the proof of Proposition 3.3. To streamline the proof process, we first introduce a lemma in Section 6.1. Subsequently, we present the comprehensive proof in Section 6.2.

497 6.1 A Lemma for Proving Proposition 3.3

498 **Lemma 6.1.** *Suppose $f : \mathbb{R} \rightarrow \mathbb{R}$ is a function with $f'(x_0) \neq 0$ for some $x_0 \in \mathbb{R}$. Then for any*
 499 *$M > 0$, it holds that*

$$500 \quad \frac{f(x_0 + \varepsilon x) - f(x_0)}{\varepsilon f'(x_0)} \rightrightarrows x \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

501 *Proof.* Clearly,

$$502 \quad \lim_{t \rightarrow 0} \frac{f(x_0 + t) - f(x_0)}{t} = f'(x_0) \neq 0 \implies \lim_{t \rightarrow 0} \frac{f(x_0 + t) - f(x_0)}{t f'(x_0)} = 1.$$

503 Then for any $\varepsilon \in (0, 1)$ and $M > 0$, there exists a small $\xi_\varepsilon > 0$ such that

$$504 \quad \left| \frac{f(x_0 + t) - f(x_0)}{t f'(x_0)} - 1 \right| < \varepsilon / M \quad \text{for any } t \in (-\xi_\varepsilon, 0) \cup (0, \xi_\varepsilon).$$

505 For each $\varepsilon \in (0, 1)$, we define

$$506 \quad g_\varepsilon(x) := \frac{f(x_0 + \varepsilon x) - f(x_0)}{\varepsilon f'(x_0)} \quad \text{for any } x \in \mathbb{R}.$$

507 Clearly, $g_\varepsilon(0) = 0$, i.e., $|g_\varepsilon(x) - x| = 0 < \varepsilon$ if $x = 0$. Moreover, for any $x \in [-M, 0) \cup (0, M]$
 508 and $\varepsilon \in (0, \xi_\varepsilon/M)$, we have $\varepsilon x \in (-\xi_\varepsilon, 0) \cup (0, \xi_\varepsilon)$, implying

$$509 \quad \begin{aligned} |g_\varepsilon(x) - x| &\leq |x| \cdot |g_\varepsilon(x)/x - 1| \leq M \cdot |g_\varepsilon(x)/x - 1| \\ &= M \cdot \left| \frac{f(x_0 + \varepsilon x) - f(x_0)}{\varepsilon x f'(x_0)} - 1 \right| < M \cdot \frac{\varepsilon}{M} = \varepsilon. \end{aligned}$$

510 Thus, we have

$$511 \quad \frac{f(x_0 + \varepsilon x) - f(x_0)}{\varepsilon f'(x_0)} = g_\varepsilon(x) \rightrightarrows x \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

512 So we finish the proof of Lemma 6.1. □

513 6.2 Proof of Proposition 3.3 with Lemma 6.1

514 With Lemma 6.1 in hand, we are ready to present the proof of Proposition 3.3.

515 *Proof of Proposition 3.3.* Given any $\varepsilon \in (0, 1)$, our goal is to construct $\phi_\varepsilon \in \mathcal{NN}_\varrho\{(k +$
 516 $2), 1; \mathbb{R} \rightarrow \mathbb{R}\}$ with $\varrho \in \mathcal{A}_{1,k}$ to approximate ReLU well on $[-M, M]$.

517 Clearly, there exist $a_0 < b_0$ and $x_0 \in (a_0, b_0)$ such that $\varrho \in C^k((a_0, b_0))$ and

$$518 \quad L_1 = \lim_{t \rightarrow 0^-} \frac{\varrho^{(k)}(x_0 + t) - \varrho^{(k)}(x_0)}{t} \neq L_2 = \lim_{t \rightarrow 0^+} \frac{\varrho^{(k)}(x_0 + t) - \varrho^{(k)}(x_0)}{t}.$$

519 Set

$$520 \quad c_0 = \min \left\{ \frac{b_0 - x_0}{2}, \frac{x_0 - a_0}{2} \right\} \quad \text{and} \quad K = \max \left\{ 1, \left| \frac{1}{L_2 - L_1} \right|, \left| \frac{L_1}{L_2 - L_1} \right| \right\}.$$

521 There exists a small $\delta_\varepsilon \in (0, c_0)$ such that

$$522 \quad \left| \frac{\varrho^{(k)}(x_0 + t) - \varrho^{(k)}(x_0)}{t} - (L_1 \cdot \mathbb{1}_{\{t < 0\}} + L_2 \cdot \mathbb{1}_{\{t > 0\}}) \right| < \varepsilon / (4KM)$$

for any $t \in (-\delta_\varepsilon, 0) \cup (0, \delta_\varepsilon)$. Define

$$\psi_\varepsilon(x) := \frac{\varrho^{(k)}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} \quad \text{for any } x \in \mathbb{R}.$$

Clearly, $\psi_\varepsilon(0) = 0$. Moreover, for any $x \in [-2M, 0) \cup (0, 2M]$ and each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$, we have $\varepsilon x \in (-\delta_\varepsilon, 0) \cup (0, \delta_\varepsilon)$, implying

$$\begin{aligned} & \left| \psi_\varepsilon(x) - (L_1 \cdot \mathbb{1}_{\{x < 0\}} + L_2 \cdot \mathbb{1}_{\{x > 0\}})x \right| \leq |x| \cdot \left| \psi_\varepsilon(x)/x - (L_1 \cdot \mathbb{1}_{\{x < 0\}} + L_2 \cdot \mathbb{1}_{\{x > 0\}}) \right| \\ & = |x| \cdot \left| \frac{\varrho^{(k)}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon x} - (L_1 \cdot \mathbb{1}_{\{\varepsilon x < 0\}} + L_2 \cdot \mathbb{1}_{\{\varepsilon x > 0\}}) \right| < 2M \cdot \frac{\varepsilon}{4KM} = \varepsilon/(2K). \end{aligned}$$

Thus, for each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$, we have

$$\left| \psi_\varepsilon(x) - (L_1 \cdot \mathbb{1}_{\{x < 0\}} + L_2 \cdot \mathbb{1}_{\{x > 0\}})x \right| < \varepsilon/(2K) \quad \text{for any } x \in [-2M, 2M],$$

implying

$$\left| \psi_\varepsilon(x) - \psi(x) \right| < \varepsilon/(2K) \quad \text{for any } x \in [-2M, 2M], \quad (3)$$

where

$$\psi(x) := (L_1 \cdot \mathbb{1}_{\{x < 0\}} + L_2 \cdot \mathbb{1}_{\{x > 0\}})x \quad \text{for any } x \in \mathbb{R}.$$

Moreover, for any $x \in \mathbb{R}$, we have

$$\begin{aligned} \psi(x) - L_1 x &= (L_1 \cdot \mathbb{1}_{\{x < 0\}} + L_2 \cdot \mathbb{1}_{\{x > 0\}})x - L_1 x (\mathbb{1}_{\{x < 0\}} + \mathbb{1}_{\{x > 0\}}) \\ &= (L_2 - L_1) \cdot \mathbb{1}_{\{x > 0\}} \cdot x = (L_2 - L_1) \cdot \text{ReLU}(x), \end{aligned}$$

implying

$$\frac{1}{L_2 - L_1} \psi(x) - \frac{L_1}{L_2 - L_1} x = \text{ReLU}(x).$$

To construct a ϱ -activated network to approximate ReLU well, we only need to construct ϱ -activated networks to effectively approximate $\psi(x)$ and x for any $x \in [-M, M]$. We divide the remaining proof into two cases: $k = 0$ and $k \geq 1$.

Case 1: $k = 0$.

First, let us consider the case of $k = 0$. In this case, $\varrho^{(k)} = \varrho$. For each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$ and any $x \in [-M, M]$, we have $x - M \in [-2M, 0] \subseteq [-2M, 2M]$, and by combining this with Equation (3), we deduce

$$\begin{aligned} & \varepsilon/(2K) > \left| \psi_\varepsilon(x - M) - \psi(x - M) \right| \\ & = \left| \psi_\varepsilon(x - M) - (L_1 \cdot \mathbb{1}_{\{x - M < 0\}} + L_2 \cdot \mathbb{1}_{\{x - M > 0\}})(x - M) \right| \\ & = \left| \psi_\varepsilon(x - M) - L_1(x - M) \right| = \left| \psi_\varepsilon(x - M) + L_1 M - L_1 x \right|. \end{aligned} \quad (4)$$

Define

$$\begin{aligned} \phi_\varepsilon(x) &:= \frac{1}{L_2 - L_1} \psi_\varepsilon(x) - \frac{1}{L_2 - L_1} (\psi_\varepsilon(x - M) + L_1 M) \\ &= \frac{1}{L_2 - L_1} \frac{\varrho(x_0 + \varepsilon x) - \varrho(x_0)}{\varepsilon} - \frac{1}{L_2 - L_1} \left(\frac{\varrho(x_0 + \varepsilon(x - M)) - \varrho(x_0)}{\varepsilon} + L_1 M \right) \end{aligned}$$

for any $x \in \mathbb{R}$. It is easy to verify that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{2, 1; \mathbb{R} \rightarrow \mathbb{R}\} = \mathcal{NN}_\varrho\{k+2, 1; \mathbb{R} \rightarrow \mathbb{R}\}$.
 Moreover, for each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$ and any $x \in [-M, M]$, we have

$$\begin{aligned} |\phi_\varepsilon(x) - \text{ReLU}(x)| &= \left| \underbrace{\frac{1}{L_2-L_1}\psi_\varepsilon(x) - \frac{1}{L_2-L_1}\left(\psi_\varepsilon(x-M) + L_1M\right)}_{\phi_\varepsilon} - \underbrace{\left(\frac{1}{L_2-L_1}\psi(x) - \frac{L_1}{L_2-L_1}x\right)}_{\text{ReLU}} \right| \\ &\leq \left| \frac{1}{L_2-L_1} \right| \cdot |\psi_\varepsilon(x) - \psi(x)| + \left| \frac{1}{L_2-L_1} \right| \cdot \left| \left(\psi_\varepsilon(x-M) + L_1M\right) - L_1x \right| \\ &< K \cdot \frac{\varepsilon}{2K} + K \cdot \frac{\varepsilon}{2K} = \varepsilon, \end{aligned}$$

where the strict inequality comes from Equations (3) and (4). Therefore, we can conclude that

$$\phi_\varepsilon(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

That means we finish the proof for the case of $k = 0$.

Case 2: $k \geq 1$.

Next, let us consider the case of $k \geq 1$. Define

$$\tilde{\phi}_\varepsilon(x) := \frac{1}{L_2-L_1}\psi_\varepsilon(x) - \frac{L_1}{L_2-L_1}x \quad \text{for any } x \in \mathbb{R}.$$

Then by Equation (3), for each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$ and any $x \in [-M, M] \subseteq [-2M, 2M]$, we have

$$\begin{aligned} |\tilde{\phi}_\varepsilon(x) - \text{ReLU}(x)| &= \left| \left(\frac{1}{L_2-L_1}\psi_\varepsilon(x) - \frac{L_1}{L_2-L_1}x \right) - \left(\frac{1}{L_2-L_1}\psi(x) - \frac{L_1}{L_2-L_1}x \right) \right| \\ &= \left| \frac{1}{L_2-L_1}\psi_\varepsilon(x) - \frac{1}{L_2-L_1}\psi(x) \right| \leq \left| \frac{1}{L_2-L_1} \right| \cdot |\psi_\varepsilon(x) - \psi(x)| < K \cdot \frac{\varepsilon}{2K} = \varepsilon/2, \end{aligned} \tag{5}$$

where the strict inequality comes from Equation (3). Our goal is to use a ϱ -activated network to effectively approximate

$$\tilde{\phi}_\varepsilon(x) = \frac{1}{L_2-L_1}\psi_\varepsilon(x) - \frac{L_1}{L_2-L_1}x = \frac{1}{L_2-L_1} \frac{\varrho^{(k)}(x_0+\varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} - \frac{L_1}{L_2-L_1}x$$

for any $x \in [-M, M]$ and $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$. To this end, we need to construct ϱ -activated networks to effectively approximate $\varrho^{(k)}(x_0 + \varepsilon x)$ and x for any $x \in [-M, M]$ and $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$.

Recall that $\varrho \in C^k((a_0, b_0)) \setminus C^{k+1}((a_0, b_0))$ with $k \geq 1$. Then there exists $x_1 \in (a_0, b_0)$ such that $\varrho'(x_1) \neq 0$. For each $\eta \in (0, 1)$, we define

$$g_\eta(x) := \frac{\varrho(x_1 + \eta x) - \varrho(x_1)}{\eta \varrho'(x_1)} \quad \text{for any } x \in \mathbb{R}.$$

By Lemma 6.1,

$$g_\eta(x) = \frac{\varrho(x_1 + \eta x) - \varrho(x_1)}{\eta \varrho'(x_1)} \rightrightarrows x \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

For each $\eta \in (0, 1)$, we define

$$h_\eta(z) := \frac{\sum_{i=0}^k (-1)^i \binom{k}{i} \varrho(z + i\eta)}{(-\eta)^k} \quad \text{for any } z \in \mathbb{R}.$$

Recall that $c_0 = \min \left\{ \frac{b_0 - x_0}{2}, \frac{x_0 - a_0}{2} \right\}$ and $\varrho \in C^k((a_0, b_0))$. By Proposition 3.2,

$$h_\eta(z) = \frac{\sum_{i=0}^k (-1)^i \binom{k}{i} \varrho(z + i\eta)}{(-\eta)^k} \rightrightarrows \varrho^{(k)}(z) \quad \text{as } \eta \rightarrow 0 \quad \text{for any } z \in [x_0 - c_0, x_0 + c_0].$$

Then there exists $\eta_\varepsilon > 0$ such that

$$|g_{\eta_\varepsilon}(x) - x| < \varepsilon/(4K) \quad \text{for any } x \in [-M, M]$$

and

$$|h_{\eta_\varepsilon}(z) - \varrho^{(k)}(z)| < \varepsilon^2/(4K) \quad \text{for any } z \in [x_0 - c_0, x_0 + c_0].$$

Next, we can define the desired ϕ_ε via

$$\begin{aligned} \phi_\varepsilon(x) &:= \frac{1}{L_2 - L_1} \frac{h_{\eta_\varepsilon}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} - \frac{L_1}{L_2 - L_1} g_{\eta_\varepsilon}(x) \\ &= \frac{\sum_{i=0}^k (-1)^i \binom{k}{i} \varrho(x_0 + \varepsilon x + i\eta_\varepsilon) - (-\eta_\varepsilon)^k \varrho^{(k)}(x_0)}{(-\eta_\varepsilon)^k (L_2 - L_1) \varepsilon} - \frac{L_1 \varrho(x_1 + \eta_\varepsilon x) - L_1 \varrho(x_1)}{(L_2 - L_1) \eta_\varepsilon \varrho'(x_1)} \end{aligned}$$

for any $x \in \mathbb{R}$. It is easy to verify that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{k+2, 1; \mathbb{R} \rightarrow \mathbb{R}\}$. Moreover, for each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M}) \subseteq (0, \frac{c_0}{2M})$ and any $x \in [-M, M]$, we have $x_0 + \varepsilon x \in [x_0 - c_0, x_0 + c_0]$, implying

$$\begin{aligned} &|\phi_\varepsilon(x) - \tilde{\phi}_\varepsilon(x)| \\ &= \left| \left(\frac{1}{L_2 - L_1} \frac{h_{\eta_\varepsilon}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} - \frac{L_1}{L_2 - L_1} g_{\eta_\varepsilon} \right) - \left(\frac{1}{L_2 - L_1} \frac{\varrho^{(k)}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} - \frac{L_1}{L_2 - L_1} x \right) \right| \\ &\leq \left| \frac{1}{L_2 - L_1} \right| \cdot \left| \frac{h_{\eta_\varepsilon}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} - \frac{\varrho^{(k)}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0)}{\varepsilon} \right| + \left| \frac{L_1}{L_2 - L_1} \right| \cdot |g_{\eta_\varepsilon}(x) - x| \\ &\leq \frac{1}{\varepsilon} \left| \frac{1}{L_2 - L_1} \right| \cdot \left| h_{\eta_\varepsilon}(x_0 + \varepsilon x) - \varrho^{(k)}(x_0 + \varepsilon x) \right| + K \cdot \frac{\varepsilon}{4K} \leq \frac{1}{\varepsilon} K \cdot \frac{\varepsilon^2}{4K} + K \cdot \frac{\varepsilon}{4K} = \varepsilon/2. \end{aligned}$$

Combining this with Equation (5), we can conclude that

$$|\phi_\varepsilon(x) - \text{ReLU}(x)| \leq |\phi_\varepsilon(x) - \tilde{\phi}_\varepsilon(x)| + |\tilde{\phi}_\varepsilon(x) - \text{ReLU}(x)| < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

for each $\varepsilon \in (0, \frac{\delta_\varepsilon}{2M})$ and any $x \in [-M, M]$. That means

$$\phi_\varepsilon(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

So we finish the proof of Proposition 3.3. □

7 Proof of Proposition 3.4

We will prove Proposition 3.4 in this section. To this end, we first establish two lemmas in Section 7.1, which play important roles in proving Proposition 3.4. Next, we give the detailed proof of Proposition 3.4 based on these two lemmas in Section 7.2.

7.1 Lemmas for Proving Proposition 3.4

Lemma 7.1. *Given any $A > 0$, suppose $\varrho : \mathbb{R} \rightarrow \mathbb{R}$ is a function with $\varrho''(x_0) \neq 0$ for some $x_0 \in \mathbb{R}$. Then there exists*

$$\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 1; \mathbb{R}^2 \rightarrow \mathbb{R}\} \quad \text{for each } \varepsilon \in (0, 1)$$

such that

$$\phi_\varepsilon(x, y) \rightrightarrows xy \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x, y \in [-A, A].$$

597 *Proof.* By L'Hôpital's Rule,

$$598 \quad \lim_{t \rightarrow 0} \frac{\varrho(x_0 + t) + \varrho(x_0 - t) - 2\varrho(x_0)}{t^2} = \lim_{t \rightarrow 0} \frac{\varrho'(x_0 + t) - \varrho'(x_0 - t)}{2t}$$

$$= \lim_{t \rightarrow 0} \frac{\varrho'(x_0 + t) - \varrho'(x_0) + \varrho'(x_0) - \varrho'(x_0 - t)}{2t} = \varrho''(x_0)/2 + \varrho''(x_0)/2 = \varrho''(x_0) \neq 0.$$

599 There exists a small $\delta_\varepsilon \in (0, 1)$ such that

$$600 \quad \left| \frac{\varrho(x_0 + t) + \varrho(x_0 - t) - 2\varrho(x_0)}{t^2 \varrho''(x_0)} - 1 \right| < \varepsilon / (4A^2) \quad \text{for any } t \in (-\delta_\varepsilon, 0) \cup (0, \delta_\varepsilon). \quad (6)$$

601 For each $\varepsilon \in (0, 1)$, we define

$$602 \quad \psi_\varepsilon(z) := \frac{\varrho(x_0 + \varepsilon z) + \varrho(x_0 - \varepsilon z) - 2\varrho(x_0)}{\varepsilon^2 \varrho''(x_0)} \quad \text{for any } z \in \mathbb{R}.$$

603 Clearly, $\psi_\varepsilon(0) = 0$, i.e., $|\psi_\varepsilon(z) - z^2| = 0 < \varepsilon$ if $z = 0$. Moreover, for any $z \in [-2A, 0) \cup (0, 2A]$
 604 and $\varepsilon \in (0, \delta_\varepsilon / (2A))$, we have $\varepsilon z \in (-\delta_\varepsilon, 0) \cup (0, \delta_\varepsilon)$, implying

$$605 \quad \begin{aligned} |\psi_\varepsilon(z) - z^2| &\leq |z^2| \cdot |\psi_\varepsilon(z)/z^2 - 1| \leq 4A^2 \cdot |\psi_\varepsilon(z)/z^2 - 1| \\ &= 4A^2 \left| \frac{\varrho(x_0 + \varepsilon z) + \varrho(x_0 - \varepsilon z) - 2\varrho(x_0)}{(\varepsilon z)^2 \varrho''(x_0)} - 1 \right| < 4A^2 \cdot \frac{\varepsilon}{4A^2} = \varepsilon, \end{aligned}$$

606 where the strict inequality comes from Equation (6). That means

$$607 \quad \psi_\varepsilon(z) \rightrightarrows z^2 \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } z \in [-2A, 2A].$$

608 Therefore, for any $x, y \in [-A, A]$, we have

$$609 \quad \psi_\varepsilon(x) \rightrightarrows x^2, \quad \psi_\varepsilon(y) \rightrightarrows y^2, \quad \text{and} \quad \psi_\varepsilon(x + y) \rightrightarrows (x + y)^2 \quad \text{as } \varepsilon \rightarrow 0^+.$$

610 Then, by defining

$$611 \quad \phi_\varepsilon(x, y) := \frac{1}{2}(\psi_\varepsilon(x + y) - \psi_\varepsilon(x) - \psi_\varepsilon(y)) \quad \text{for any } x, y \in \mathbb{R},$$

612 we have

$$613 \quad \phi_\varepsilon(x, y) \rightrightarrows \frac{1}{2}((x + y)^2 - x^2 - y^2) = xy \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x, y \in [-A, A].$$

614 Furthermore, as shown in Figure 3, $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 1; \mathbb{R}^2 \rightarrow \mathbb{R}\}$. Thus, we finish the proof of Lemma 7.1. \square

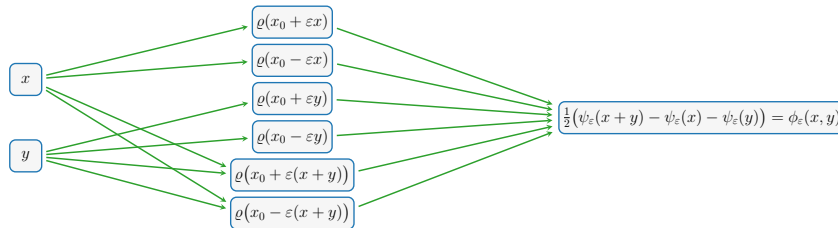


Figure 3: An illustration of the network architecture realizing ϕ_ε .

615

616 **Lemma 7.2.** *Given any $M > 0$ and two functions $g_1, g_{2,\delta} : \mathbb{R} \rightarrow \mathbb{R}$ for each $\delta \in (0, 1)$, suppose*

$$617 \quad \sup_{x \in \mathbb{R}} |g_1(x)| < \infty, \quad \lim_{x \rightarrow -\infty} g_1(x) = 0, \quad \lim_{x \rightarrow \infty} g_1(x) = 1,$$

618 *and*

$$619 \quad g_{2,\delta}(x) \rightrightarrows x \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } x \in [-M, M],$$

620 *Then for any $\varepsilon > 0$, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0, 1)$ such that*

$$621 \quad |g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| < \varepsilon \quad \text{for any } x \in [-M, M].$$

622 *Proof.* Since $\sup_{x \in \mathbb{R}} |g_1(x)| < \infty$, $\lim_{x \rightarrow -\infty} g_1(x) = 0$, and $\lim_{x \rightarrow \infty} g_1(x) = 1$, we have

$$623 \quad K_0 = \sup_{x \in \mathbb{R}} |g_1(x)| \in [1, \infty)$$

624 and there exists $K_1 > 0$ such that

$$625 \quad |g_1(x)| < \varepsilon_1 \text{ for any } x \leq -K_1/4 \quad \text{and} \quad |g_1(x) - 1| < \varepsilon_1 \text{ for any } x \geq K_1/4,$$

626 where $\varepsilon_1 = \varepsilon/(2M)$. It follows that

$$627 \quad |g_1(K_0 K_1 x / \varepsilon) - \mathbb{1}_{\{x > 0\}}| < \varepsilon_1 = \varepsilon/(2M) \quad \text{for any } |x| \geq \varepsilon/(4K_0), \quad (7)$$

628 Recall that $g_{2,\delta}(x) \rightrightarrows x$ as $\delta \rightarrow 0^+$ for any $x \in [-M, M]$. There exists $\delta_\varepsilon \in (0, 1)$ such that

$$629 \quad |g_{2,\delta_\varepsilon} - x| < \varepsilon_2 = \varepsilon/(3K_0) \quad \text{for any } x \in [-M, M]. \quad (8)$$

630 Observe that $\text{ReLU}(x) = x \cdot \mathbb{1}_{\{x > 0\}}$ for any $x \in \mathbb{R}$. Setting $K_\varepsilon = K_0 K_1 / \varepsilon$ and by Equation (8),
631 for any $x \in [-M, M]$, we have

$$\begin{aligned} & |g_1(K_\varepsilon x) g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| = |g_1(K_\varepsilon x) g_{2,\delta_\varepsilon}(x) - x \cdot \mathbb{1}_{\{x > 0\}}| \\ & \leq |g_1(K_\varepsilon x) g_{2,\delta_\varepsilon}(x) - x g_1(K_\varepsilon x)| + |x g_1(K_\varepsilon x) - x \cdot \mathbb{1}_{\{x > 0\}}| \\ 632 & \leq |g_1(K_\varepsilon x)| \cdot |g_{2,\delta_\varepsilon}(x) - x| + |x| \cdot |g_1(K_\varepsilon x) - \mathbb{1}_{\{x > 0\}}| \\ & \leq K_0 \cdot \varepsilon_2 + |x| \cdot |g_1(K_0 K_1 x / \varepsilon) - \mathbb{1}_{\{x > 0\}}|. \end{aligned}$$

633 In the case of $|x| < \varepsilon/(4K_0)$, we have

$$\begin{aligned} & |g_1(K_\varepsilon x) g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| \leq K_0 \cdot \varepsilon_2 + |x| \cdot |g_1(K_0 K_1 x / \varepsilon) - \mathbb{1}_{\{x > 0\}}| \\ 634 & \leq K_0 \cdot \frac{\varepsilon}{3K_0} + \frac{\varepsilon}{4K_0} \cdot (K_0 + 1) \leq \varepsilon/3 + \varepsilon/2 < \varepsilon. \end{aligned}$$

635 We may assume $\varepsilon/(4K_0) \leq M$ since the proof is complete if $\varepsilon/(4K_0) > M$. In the case of
636 $|x| \in [\varepsilon/(4K_0), M]$, by Equation (7), we have

$$\begin{aligned} & |g_1(K_\varepsilon x) g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| \leq K_0 \cdot \varepsilon_2 + |x| \cdot |g_1(K_0 K_1 x / \varepsilon) - \mathbb{1}_{\{x > 0\}}| \\ 637 & \leq K_0 \cdot \varepsilon_2 + M \cdot \varepsilon_1 \leq K_0 \cdot \frac{\varepsilon}{3K_0} + M \cdot \frac{\varepsilon}{2M} \leq \varepsilon/3 + \varepsilon/2 < \varepsilon \end{aligned}$$

638 Therefore, for any $x \in [-M, M]$, we have

$$639 \quad |g_1(K_\varepsilon x) g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| < \varepsilon,$$

640 which means we finish the proof. □

641 7.2 Proof of Proposition 3.4 with Lemmas 7.2 and 7.1

642 Having established Lemmas 7.2 and 7.1 in Section 7.1, we are now prepared to prove Proposi-
643 tion 3.4.

644 *Proof of Proposition 3.4.* For any $\varepsilon \in (0, 1)$, our goal is to construct $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$
645 with $\varrho \in \mathcal{A}_2 \cup \mathcal{A}_3$ to approximate ReLU well on $[-M, M]$. We divide the proof into two cases:
646 $\varrho \in \mathcal{A}_2$ and $\varrho \in \mathcal{A}_3$.

647 **Case 1:** $\varrho \in \mathcal{A}_2$.

648 First, let us consider the case of $\varrho \in \mathcal{A}_2$. Clearly, we have

$$649 \sup_{x \in [-r, r]} |\varrho(x)| < \infty \quad \text{for any } r > 0 \quad (9)$$

650 and there exist $T_0 > 0$ and $x_0 \in \mathbb{R}$ such that $\varrho''(x_0) \neq 0$ and

$$651 L_1 = \lim_{x \rightarrow -\infty} \widehat{\varrho}(x) \neq L_2 = \lim_{x \rightarrow \infty} \widehat{\varrho}(x),$$

652 where

$$653 \widehat{\varrho}(x) := \varrho(x + T_0) - \varrho(x) \quad \text{for any } x \in \mathbb{R}.$$

654 It follows that $\sup_{x \in \mathbb{R}} |\widehat{\varrho}(x)| < \infty$.

655 By defining

$$656 g_1(x) := \frac{\widehat{\varrho}(x) - L_1}{L_2 - L_1} = \frac{\varrho(x + T_0) - \varrho(x) - L_1}{L_2 - L_1} \quad \text{for any } x \in \mathbb{R},$$

657 we have

$$658 \sup_{x \in \mathbb{R}} |g_1(x)| < \infty, \quad \lim_{x \rightarrow -\infty} g_1(x) = 0, \quad \text{and} \quad \lim_{x \rightarrow \infty} g_1(x) = 1.$$

659 Since $\varrho''(x_0) \neq 0$, there exists $x_1 \in \mathbb{R}$ such that $\varrho'(x_1) \neq 0$. For each $\delta \in (0, 1)$, we define

$$660 g_{2,\delta}(x) := \frac{\varrho(x_1 + \delta x) - \varrho(x_1)}{\delta \varrho'(x_1)} \quad \text{for any } x \in \mathbb{R}.$$

661 By Lemma 6.1,

$$662 g_{2,\delta}(x) \rightrightarrows x \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

663 By Lemma 7.2, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0, 1)$ such that

$$664 |g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| < \varepsilon \quad \text{for any } x \in [-M, M]. \quad (10)$$

665 It follows from Equation (9) that

$$\begin{aligned} 666 A &= \sup_{x \in [-M, M]} \max \{ |g_1(K_\varepsilon x)|, |g_{2,\delta_\varepsilon}(x)| \} \\ &= \sup_{x \in [-M, M]} \max \left\{ \left| \frac{\varrho(K_\varepsilon x + T_0) - \varrho(K_\varepsilon x) - L_1}{L_2 - L_1} \right|, \left| \frac{\varrho(x_1 + \delta_\varepsilon x) - \varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)} \right| \right\} < \infty. \end{aligned}$$

667 Since $\varrho''(x_0) \neq 0$, by Lemma 7.1, there exists

$$668 \Gamma_\eta \in \mathcal{NN}_\varrho\{6, 1; \mathbb{R}^2 \rightarrow \mathbb{R}\} \quad \text{for each } \eta \in (0, 1)$$

669 such that

$$670 \Gamma_\eta(u, v) \rightrightarrows uv \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } u, v \in [-A, A].$$

671 Then there exists $\eta_\varepsilon \in (0, 1)$ such that

$$672 \quad |\Gamma_{\eta_\varepsilon}(u, v) - uv| < \varepsilon \quad \text{for any } u, v \in [-A, A],$$

673 implying

$$674 \quad \left| \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) - g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) \right| < \varepsilon \quad \text{for any } x \in [-M, M]. \quad (11)$$

675 Define

$$676 \quad \phi_\varepsilon(x) := \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) \quad \text{for any } x \in \mathbb{R}.$$

677 Then, by Equations (10) and (11), we have

$$\begin{aligned} & |\phi_\varepsilon(x) - \text{ReLU}(x)| = \left| \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) - \text{ReLU}(x) \right| \\ 678 \quad & \leq \left| \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) - g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) \right| + \left| g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x) \right| \\ & < \varepsilon + \varepsilon = 2\varepsilon \end{aligned}$$

679 for any $x \in [-M, M]$, from which we deduce

$$680 \quad \phi_\varepsilon(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

681 We still need to demonstrate that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$. By defining

$$682 \quad \psi_\varepsilon(x) := \left(\frac{\varrho(K_\varepsilon x + T_0) - \varrho(K_\varepsilon x) - L_1}{L_2 - L_1}, \frac{\varrho(x_1 + \delta_\varepsilon x) - \varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)} \right) \quad \text{for any } x \in \mathbb{R},$$

683 we have $\psi_\varepsilon \in \mathcal{NN}_\varrho\{3, 1; \mathbb{R} \rightarrow \mathbb{R}^2\}$ and

$$\begin{aligned} 684 \quad \phi_\varepsilon(x) &= \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) \\ &= \Gamma_{\eta_\varepsilon}\left(\frac{\varrho(K_\varepsilon x + T_0) - \varrho(K_\varepsilon x) - L_1}{L_2 - L_1}, \frac{\varrho(x_1 + \delta_\varepsilon x) - \varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)}\right) = \Gamma_{\eta_\varepsilon} \circ \psi_\varepsilon(x) \end{aligned}$$

685 for any $x \in \mathbb{R}$. Recall that $\Gamma_{\eta_\varepsilon} \in \mathcal{NN}_\varrho\{6, 1; \mathbb{R}^2 \rightarrow \mathbb{R}\}$. Therefore, we have $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$, as required.

687 **Case 2:** $\varrho \in \mathcal{A}_3$.

688 Let us now turn to the case of $\varrho \in \mathcal{A}_3$. Clearly, we have $\sup_{x \in \mathbb{R}} |\varrho(x)| < \infty$, $\varrho''(x_0) \neq 0$ for
689 some $x_0 \in \mathbb{R}$, and

$$690 \quad L_1 = \lim_{x \rightarrow -\infty} \varrho(x) \neq L_2 = \lim_{x \rightarrow \infty} \varrho(x).$$

691 By defining

$$692 \quad g_1(x) := \frac{\varrho(x) - L_1}{L_2 - L_1} \quad \text{for any } x \in \mathbb{R},$$

693 we have

$$694 \quad \sup_{x \in \mathbb{R}} |g_1(x)| < \infty, \quad \lim_{x \rightarrow -\infty} g_1(x) = 0, \quad \text{and} \quad \lim_{x \rightarrow \infty} g_1(x) = 1.$$

695 Since $\varrho''(x_0) \neq 0$, there exists x_1 such that $\varrho'(x_1) \neq 0$. For each $\delta \in (0, 1)$, we define

$$696 \quad g_{2,\delta}(x) := \frac{\varrho(x_1 + \delta x) - \varrho(x_1)}{\delta \varrho'(x_1)} \quad \text{for any } x \in \mathbb{R}.$$

697 By Lemma 6.1,

$$698 \quad g_{2,\delta}(x) \rightrightarrows x \quad \text{as } \delta \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

By Lemma 7.2, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0, 1)$ such that

$$|g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x)| < \varepsilon \quad \text{for any } x \in [-M, M]. \quad (12)$$

The fact $\sup_{x \in \mathbb{R}} |\varrho(x)| < \infty$ implies

$$\begin{aligned} A &= \sup_{x \in [-M, M]} \max \{|g_1(K_\varepsilon x)|, |g_{2,\delta_\varepsilon}(x)|\} \\ &= \sup_{x \in [-M, M]} \max \left\{ \left| \frac{\varrho(K_\varepsilon x) - L_1}{L_2 - L_1} \right|, \left| \frac{\varrho(x_1 + \delta_\varepsilon x) - \varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)} \right| \right\} < \infty. \end{aligned}$$

Since $\varrho''(x_0) \neq 0$, by Lemma 7.1, there exists

$$\Gamma_\eta \in \mathcal{NN}_\varrho\{6, 1; \mathbb{R}^2 \rightarrow \mathbb{R}\} \quad \text{for each } \eta \in (0, 1)$$

such that

$$\Gamma_\eta(u, v) \rightrightarrows uv \quad \text{as } \eta \rightarrow 0^+ \quad \text{for any } u, v \in [-A, A].$$

Then there exists $\eta_\varepsilon \in (0, 1)$ such that

$$|\Gamma_{\eta_\varepsilon}(u, v) - uv| < \varepsilon \quad \text{for any } u, v \in [-A, A],$$

implying

$$\left| \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) - g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) \right| < \varepsilon \quad \text{for any } x \in [-M, M]. \quad (13)$$

Define

$$\phi_\varepsilon(x) := \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) \quad \text{for any } x \in \mathbb{R}.$$

Next, by Equations (12) and (13), we have

$$\begin{aligned} |\phi_\varepsilon(x) - \text{ReLU}(x)| &= \left| \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) - \text{ReLU}(x) \right| \\ &\leq \left| \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) - g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) \right| + \left| g_1(K_\varepsilon x) \cdot g_{2,\delta_\varepsilon}(x) - \text{ReLU}(x) \right| \\ &< \varepsilon + \varepsilon = 2\varepsilon \end{aligned}$$

for any $x \in [-M, M]$, from which we deduce

$$\phi_\varepsilon(x) \rightrightarrows \text{ReLU}(x) \quad \text{as } \varepsilon \rightarrow 0^+ \quad \text{for any } x \in [-M, M].$$

It remains to show $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$. By defining

$$\psi_\varepsilon(x) := \left(\frac{\varrho(K_\varepsilon x) - L_1}{L_2 - L_1}, \frac{\varrho(x_1 + \delta_\varepsilon x) - \varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)} \right) \quad \text{for any } x \in \mathbb{R},$$

we have $\psi_\varepsilon \in \mathcal{NN}_\varrho\{2, 1; \mathbb{R} \rightarrow \mathbb{R}^2\}$ and

$$\phi_\varepsilon(x) = \Gamma_{\eta_\varepsilon}(g_1(K_\varepsilon x), g_{2,\delta_\varepsilon}(x)) = \Gamma_{\eta_\varepsilon}\left(\frac{\varrho(K_\varepsilon x) - L_1}{L_2 - L_1}, \frac{\varrho(x_1 + \delta_\varepsilon x) - \varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)}\right) = \Gamma_{\eta_\varepsilon} \circ \psi_\varepsilon(x)$$

for any $x \in \mathbb{R}$. Recall that $\Gamma_{\eta_\varepsilon} \in \mathcal{NN}_\varrho\{6, 1; \mathbb{R}^2 \rightarrow \mathbb{R}\}$. Hence, we can conclude that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{6, 2; \mathbb{R} \rightarrow \mathbb{R}\}$. This result completes the proof of Proposition 3.4. \square

Acknowledgments

Jianfeng Lu was partially supported by NSF grants CCF-1910571 and DMS-2012286. Hongkai Zhao was partially supported by NSF grant DMS-2012860 and DMS-2309551.

References

- [1] Chenglong Bao, Qianxiao Li, Zuowei Shen, Cheng Tai, Lei Wu, and Xueshuang Xiang. Approximation analysis of convolutional neural networks. *East Asian Journal on Applied Mathematics*, 13(3):524–549, 2023.
- [2] Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.
- [3] Andrew R. Barron and Jason M. Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv e-prints*, page arXiv:1809.03090, September 2018.
- [4] Helmut. Bölcskei, Philipp. Grohs, Gitta. Kutyniok, and Philipp. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [6] Kuan-Lin Chen, Harinath Garudadri, and Bhaskar D Rao. Improved bounds on neural complexity for representing piecewise linear functions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7167–7180. Curran Associates, Inc., 2022.
- [7] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [8] Charles K. Chui, Shao-Bo Lin, and Ding-Xuan Zhou. Construction of neural networks for realization of localized deep learning. *Frontiers in Applied Mathematics and Statistics*, 4:14, 2018.
- [9] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- [10] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

- [12] Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. Special issue on deep reinforcement learning.
- [13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.
- [14] Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive Approximation*, 55:259–367, 2022.
- [15] Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05):803–859, 2020.
- [16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv e-prints*, page arXiv:1606.08415, June 2016.
- [17] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.
- [18] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
- [19] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- [21] Dandan Li and Yuan Zhou. Soft-Root-Sign: A new bounded neural activation function. In *Pattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16–18, 2020, Proceedings, Part III*, page 310–319, Berlin, Heidelberg, 2020. Springer-Verlag.
- [22] Qianxiao Li, Ting Lin, and Zuowei Shen. Deep learning via dynamical systems: An approximation perspective. *Journal of the European Mathematical Society*, 25(5):1671–1709, 2023.
- [23] Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.
- [24] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML, Workshop on Deep Learning for Audio, Speech, and Language Processing*. Atlanta, Georgia, USA, 2013.
- [25] Diganta Misra. Mish: A self regularized non-monotonic activation function. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

- [26] Hadrien Montanelli and Haizhao Yang. Error bounds for deep ReLU networks using the Kolmogorov-Arnold superposition theorem. *Neural Networks*, 129:1–6, 2020.
- [27] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML’10*, page 807–814, Madison, WI, USA, 2010. Omnipress.
- [28] Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020.
- [29] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv e-prints*, page arXiv:1710.05941, October 2017.
- [30] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.
- [31] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.
- [32] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation: Achieving arbitrary accuracy with fixed number of neurons. *Journal of Machine Learning Research*, 23(276):1–60, 2022.
- [33] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation in terms of intrinsic parameters. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19909–19934. PMLR, 17–23 Jul 2022.
- [34] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network architecture beyond width and depth. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 5669–5681. Curran Associates, Inc., 2022.
- [35] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of ReLU networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157:101–135, 2022.
- [36] Jonathan W. Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022.
- [37] Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.
- [38] Joseph Turian, James Bergstra, and Yoshua Bengio. Quadratic features and deep architectures for chunking. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short ’09*, page 245–248, USA, 2009. Association for Computational Linguistics.

- [39] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [40] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.
- [41] Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 639–649. PMLR, 06–09 Jul 2018.
- [42] Shijun Zhang. Deep neural network approximation via function compositions. *PhD Thesis, National University of Singapore*, 2020. URL <https://scholarbank.nus.edu.sg/handle/10635/186064>.
- [43] Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. On enhancing expressive power via compositions of single fixed-size relu network. *arXiv e-prints*, page arXiv:2301.12353, January 2023.
- [44] Shijun Zhang, Hongkai Zhao, Yimin Zhong, and Haomin Zhou. Why shallow networks struggle with approximating and learning high frequency: A numerical study. *arXiv e-prints*, page arXiv:2306.17301, June 2023.
- [45] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.