Deep Network Approximation: Beyond ReLU to Diverse Activation Functions

Shijun Zhang* SHIJUN.ZHANG@DUKE.EDU
Department of Mathematics
Duke University

Jianfeng Lu JIANFENG@MATH.DUKE.EDU
Department of Mathematics
Duke University

Hongkai Zhao ZHAO@MATH.DUKE.EDU
Department of Mathematics
Duke University

Abstract

This paper explores the expressive power of deep neural networks for a diverse range of activation functions. An activation function set $\mathscr{A}$ is defined to encompass the majority of commonly used activation functions, such as `ReLU`, `LeakyReLU`, `ReLU`^{2}, `ELU`, `SELU`, `Softplus`, `GELU`, `SiLU`, `Swish`, `Mish`, `Sigmoid`, `Tanh`, `Arctan`, `Softsign`, `dSiLU`, and `SRS`. We demonstrate that for any activation function $\varrho \in \mathscr{A}$, a `ReLU` network of width N and depth L can be approximated to arbitrary precision by a ϱ-activated network of width $4N$ 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.

Keywords: deep neural networks, rectified linear unit, diverse activation functions, expressive power, nonlinear approximation

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`) (Nair and Hinton, 2010) 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 (Krizhevsky et al., 2012). Optimizing deep networks activated by `ReLU` is simpler compared to networks utilizing other activation functions such as

* Corresponding author.

Sigmoid or Tanh, since gradients can propagate when the input to ReLU is positive. It was also shown in the recent work (Zhang et al., 2023b) 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.

In recent developments, various alternative activation functions have been proposed as replacements for ReLU. Notable examples include the Leaky ReLU (LeakyReLU) (Maas et al., 2013), the Exponential Linear Units (ELU) (Clevert et al., 2016), and the Gaussian Error Linear Unit (GELU) (Hendrycks and Gimpel, 2016). 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 (Brown et al., 2020), BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), 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 $\mathscr{A}$, 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, $\mathtt{ReLU}^2$ (ReLU squared) (Siegel and Xu, 2022), ELU, and SELU (Scaled Exponential Linear Unit) (Klambauer et al., 2017). These activation functions are continuous piecewise smooth functions belonging to the set $\mathscr{A}_1 \coloneqq \cup_{k=0}^{\infty}\mathscr{A}_{1,k}$, where $\mathscr{A}_{1,k}$, for each smoothness index $k\in\mathbb{N}$, is defined as

$$\mathscr{A}_{1,k} \coloneqq \Big\{\varrho:\mathbb{R}\to\mathbb{R} \;\Big|\; \exists\, a_0<b_0,\ \varrho\in C^k\big((a_0,b_0)\big),\quad \exists\, x_0\in(a_0,b_0), \\ \mathbb{R}\ni \lim_{t\to 0^-}\tfrac{\varrho^{(k)}(x_0+t)-\varrho^{(k)}(x_0)}{t} \neq \lim_{t\to 0^+}\tfrac{\varrho^{(k)}(x_0+t)-\varrho^{(k)}(x_0)}{t}\in\mathbb{R}\Big\}.$$

It is worth noting that $\varrho\in C^k\big((a_0,b_0)\big)\backslash C^{k+1}\big((a_0,b_0)\big)$ is necessary to ensure $\varrho\in\mathscr{A}_{1,k}$. 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 $f:U\subseteq\mathbb{R}\to\mathbb{R}$. For instance, $f^{(0)}$ refers to the function itself, and $f^{(1)}$ represents the first derivative. Let $\mathbb{N}$ denote the set of natural numbers, i.e., $\mathbb{N}\coloneqq\{0,1,2,\cdots\}$, and set $\mathbb{N}^+\coloneqq\mathbb{N}\backslash\{0\}$. Given a function $f:\Omega\subseteq\mathbb{R}^d\to\mathbb{R}$, we denote $\partial^{\boldsymbol{\alpha}}f$ as the partial derivative $\boldsymbol{x}\mapsto\frac{\partial^{\boldsymbol{\alpha}}}{\partial\boldsymbol{x}^{\boldsymbol{\alpha}}}f(\boldsymbol{x}) = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}}\frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}}\cdots\frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}f(\boldsymbol{x})$ for any $\boldsymbol{x}=(x_1,\cdots,x_d)\in\Omega$ and $\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_d)\in\mathbb{N}^d$. Let $C^k(\Omega)$ denote the set of all functions $f:\Omega\subseteq\mathbb{R}^d\to\mathbb{R}$, in which the partial derivatives $\frac{\partial^{\boldsymbol{\alpha}}}{\partial\boldsymbol{x}^{\boldsymbol{\alpha}}}f$ exist and are continuous for any $\boldsymbol{\alpha}\in\mathbb{N}^d$ with $\sum_{i=1}^d\alpha_i\le k$. 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` (Glorot et al., 2011), `GELU`, `SiLU` (Sigmoid Linear Unit) (Hendrycks and Gimpel, 2016; Elfwing et al., 2018), `Swish` (Ramachandran et al., 2017), and `Mish` (Misra, 2020). These activation functions are included in the set $\mathscr{A}_2$, defined via

$$\mathscr{A}_2 \coloneqq \Big\{\varrho:\mathbb{R}\to\mathbb{R} \;\Big|\; \forall\, A,\ \sup_{x\in[-A,A]}|\varrho(x)|<\infty,\quad \exists\, x_0\in\mathbb{R},\ \varrho''(x_0)\neq 0,\quad \exists\, T_0>0,$$
$$\mathbb{R}\ni \lim_{x\to-\infty}\big(\varrho(x+T_0)-\varrho(x)\big)\neq \lim_{x\to\infty}\big(\varrho(x+T_0)-\varrho(x)\big)\in\mathbb{R}\Big\}.$$

The set $\mathscr{A}_2$ 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 $\mathscr{A}_2$ below.

$$\Big\{\varrho\in C(\mathbb{R}): \exists\, x_0\in\mathbb{R},\ \varrho''(x_0)\neq 0,\quad \mathbb{R}\ni\lim_{x\to-\infty}\varrho'(x)\neq\lim_{x\to\infty}\varrho'(x)\in\mathbb{R}\Big\}\subseteq\mathscr{A}_2.$$

The final case consists of S-shaped functions, e.g., `Sigmoid`, `Tanh`, `Arctan`, `Softsign`(Turian et al., 2009). These functions are part of the set $\mathscr{A}_3$, which is defined via

$$\mathscr{A}_3 \coloneqq \Big\{\varrho:\mathbb{R}\to\mathbb{R} \;\Big|\; \sup_{x\in\mathbb{R}}|\varrho(x)|<\infty,\quad \exists\, x_0\in\mathbb{R},\ \varrho''(x_0)\neq 0,$$
$$\mathbb{R}\ni\lim_{x\to-\infty}\varrho(x)\neq\lim_{x\to\infty}\varrho(x)\in\mathbb{R}\Big\}.$$

The set $\mathscr{A}_3$ can be regarded as a collection of generalized S-shaped functions, which encompasses additional activation functions, such as `dSiLU` (derivative of `SiLU`) (Elfwing et al., 2018) and `SRS` (Soft-Root-Sign) (Li and Zhou, 2020). Moreover, the derivatives of `Softplus`, `GELU`, `SiLU`, `Swish`, and `Mish` are also classified within $\mathscr{A}_3$.

Then the activation function set $\mathscr{A}$ is defined as the union of $\cup_{k=0}^{2}\mathscr{A}_{1,k}$, $\mathscr{A}_2$, and $\mathscr{A}_3$:

$$\mathscr{A} \coloneqq \big(\cup_{k=0}^{2}\mathscr{A}_{1,k}\big)\cup\mathscr{A}_2\cup\mathscr{A}_3.$$

The definitions of $\mathscr{A}$, $\mathscr{A}_{1,k}$ for $k\in\mathbb{N}$, $\mathscr{A}_2$, and $\mathscr{A}_3$ will remain consistent throughout the whole paper. It is worth noting that if $\varrho\in\mathscr{A}$, then $w_1\varrho(w_0x+b_0)+b_1\in\mathscr{A}$ provided $w_0\neq 0\neq w_1$. Notably, the set $\mathscr{A}$ encompasses the majority of commonly used activation functions, such as `ReLU`, `LeakyReLU`, `ReLU`^{2}, `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 $\boldsymbol{f}:\mathbb{R}^d\to\mathbb{R}^n$ by

$$\|\boldsymbol{f}\|_{\sup([-A,A]^d)} \coloneqq \sup\big\{|f_i(\boldsymbol{x})| : \boldsymbol{x}\in[-A,A]^d,\ i\in\{1,2,\cdots,n\}\big\},$$

where f_i is the i-th component of $\boldsymbol{f}$. This paper exclusively focuses on fully connected feed-forward neural networks. We denote $\mathcal{NN}_\varrho\{N,\,L;\,\mathbb{R}^d\to\mathbb{R}^n\}$ as the set of functions $\boldsymbol{\phi}:\mathbb{R}^d\to\mathbb{R}^n$ 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 $\boldsymbol{\phi}:\mathbb{R}^d\to\mathbb{R}^n$ 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 $\boldsymbol{\phi}$ can be expressed as

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

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

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

Theorem 1. *Suppose $\varrho \in \mathscr{A}$ and $\boldsymbol{\phi}_{\texttt{ReLU}} \in \mathcal{NN}_{\texttt{ReLU}}\{N,\ L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ with $N, L, d, n \in \mathbb{N}^+$. Then for any $\varepsilon > 0$ and $A > 0$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\{4N,\ 2L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ such that*

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

The proof of Theorem 1 can be found in Section 3. Theorem 1 implies that a `ReLU` network of width N and depth L can be approximated by a ϱ-activated network of width $4N$ and $2L$ arbitrarily well on any bounded set for any pre-specified $\varrho \in \mathscr{A}$. In other words, $\mathcal{NN}_\varrho\{4N,\ 2L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ is dense in $\mathcal{NN}_{\texttt{ReLU}}\{N,\ L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ in terms of the $\|\cdot\|_{\sup([-A,A]^d)}$ norm for any pre-specified $A > 0$ and $\varrho \in \mathscr{A}$. It is worth mentioning while Theorem 1 covers activation functions $\varrho \in \mathscr{A}_{1,k}$ only for $k = 0, 1, 2$, 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, 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 (Shen et al., 2022a) 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 $C_{d,3}\,\omega_f\big(\big(N^2L^2\ln(N+1)\big)^{-1/d}\big)$, where $C_{d,1}$, $C_{d,2}$, and $C_{d,3}$ are constants[1] determined by d, and $\omega_f(\cdot)$ is the modulus of continuity of $f \in C([0,1]^d)$ defined via

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

By combining this result with Theorem 1, an immediate corollary follows.

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

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

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

[1] The values of $C_{d,1}$, $C_{d,2}$, and $C_{d,3}$ are explicitly given in (Shen et al., 2022a).

It is demonstrated in Theorem 1.1 of (Shen et al., 2022) 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 constants[2] 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

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

By combining the aforementioned result with Theorem 1, we can promptly deduce the subsequent corollary.

Corollary 3. *Suppose $\varrho\in\mathscr{A}$ and $f\in C^s([0,1]^d)$ with $s,d\in\mathbb{N}^+$. Then for any $N,L\in\mathbb{N}^+$, there exists $\phi\in\mathcal{NN}_\varrho\big\{C_{s,d,1}N\ln(N+1),\ C_{s,d,2}L\ln(L+1);\ \mathbb{R}^d\to\mathbb{R}\big\}$ such that*

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

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 (Chen et al., 2022) that a continuous piecewise linear function $f:\mathbb{R}^d\to\mathbb{R}$ 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, we obtain the following corollary.

Corollary 4. *Suppose $\varrho\in\mathscr{A}$ and let $f:\mathbb{R}^d\to\mathbb{R}$ 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 $\phi\in\mathcal{NN}_\varrho\big\{4\lceil 3q/2\rceil q,\ 4\lceil\log_2 q\rceil+2;\ \mathbb{R}^d\to\mathbb{R}\big\}$, such that*

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

It is demonstrated in (Zhang et al., 2023a) 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 (Zhang et al., 2023a) establishes that $\mathcal{L}_2\circ\boldsymbol{g}^{\circ(3r+1)}\circ\boldsymbol{\mathcal{L}}_1$ can approximate a continuous function $f\in C([0,1]^d)$ with an error $6\sqrt{d}\,\omega_f(r^{-1/d})$, where $\boldsymbol{g}\in\mathcal{NN}_{\texttt{ReLU}}\{69d+48,\ 5;\ \mathbb{R}^{5d+5}\to\mathbb{R}^{5d+5}\}$, $\boldsymbol{\mathcal{L}}_1$ and $\mathcal{L}_2$ are two affine linear maps matching the dimensions, and $\boldsymbol{g}^{\circ r}$ denotes the r-times composition of $\boldsymbol{g}$. By merging this outcome with Theorem 1, we can promptly deduce the subsequent corollary.

Corollary 5. *Suppose $\varrho\in\mathscr{A}$ 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 $\boldsymbol{g}\in\mathcal{NN}_\varrho\{276d+192,\ 10;\ \mathbb{R}^{5d+5}\to\mathbb{R}^{5d+5}\}$ and two affine linear maps $\boldsymbol{\mathcal{L}}_1:\mathbb{R}^d\to\mathbb{R}^{5d+5}$ and $\mathcal{L}_2:\mathbb{R}^{5d+5}\to\mathbb{R}$ such that*

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

It is worth highlighting that the approximation error in Corollary 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 (Zhang et al., 2023a) with Theorem 1.

[2] The values of $C_{s,d,1}$, $C_{s,d,2}$, and $C_{s,d,3}$ are explicitly provided in (Shen et al., 2022).

The remainder of this paper is organized as follows. In Section 2, we explore some additional related topics. We present two supplementary theorems, Theorems 6 and 7, in Section 2.1 to complement Theorem 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, 6, and 7. In Section 3.1, we introduce the notations used throughout this paper. In Section 3.2, we present several propositions, namely Propositions 8, 9, 10, and 11, outlining the underlying ideas for proving Theorems 1, 6, and 7. Subsequently, by assuming the validity of propositions, we provide the proof of Theorem 1 in Section 3.3, followed by the subsequent proofs of Theorems 6 and 7 in Section 3.4. Finally, we prove Propositions 8, 9, 10, and 11 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 6 and 7, which complement Theorem 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 specifically focuses on activation functions $\varrho \in \mathscr{A}_{1,k}$ with $k = 0, 1, 2$. However, we can also obtain similar results for larger values of $k \in \mathbb{N}$, where $\varrho \in \mathscr{A}_{1,k}$ 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 6. *Given any $k \in \mathbb{N}$ and $\varrho \in C^k(\mathbb{R})$, suppose $\boldsymbol{\phi}_{\varrho^{(k)}} \in \mathcal{NN}_{\varrho^{(k)}}\{N,\ L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ with $N, L, d, n \in \mathbb{N}^+$. Then for any $\varepsilon > 0$ and $A > 0$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\{(k+1)N,\ L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ such that*

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

Furthermore, the following theorem specifically addresses $\varrho \in \mathscr{A}_{1,k}$ for any $k \in \mathbb{N}$. Specifically, we demonstrate that for any $\varrho \in \mathscr{A}_{1,k}$ 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 7. *Suppose $\boldsymbol{\phi}_{\texttt{ReLU}} \in \mathcal{NN}_{\texttt{ReLU}}\{N,\ L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ with $N, L, d, n \in \mathbb{N}^+$. Then for any $\varepsilon > 0$, $A > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathscr{A}_{1,k}$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\{(k+2)N,\ L;\ \mathbb{R}^d \to \mathbb{R}^n\}$ such that*

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

The proofs of Theorems 6 and 7 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 (Cybenko, 1989; Hornik et al., 1989; Barron, 1993; Yarotsky, 2018, 2017; Bölcskei et al., 2019; Zhou, 2020; Chui et al., 2018; Gribonval et al., 2022; Gühring et al., 2020; Suzuki, 2019; Nakada and Imaizumi, 2020; Chen et al., 2019; Bao et al., 2023; Li et al., 2023; Montanelli and Yang, 2020; Shen et al., 2019, 2020; Lu et al., 2021; Zhang, 2020; Shen et al., 2022b,a). 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 (Cybenko, 1989; Hornik, 1991; Hornik et al., 1989) 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 (Barron, 1993; Barron and Klusowski, 2018), 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 (Krizhevsky et al., 2012). 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 (Yarotsky, 2018, 2017; Shen et al., 2019, 2020; Lu et al., 2021; Zhang et al., 2023a; Shen et al., 2022; Zhang, 2020). 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`, $\texttt{ReLU}^2$, `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

$$\texttt{ReLU}(x) = \max\{0, x\}, \qquad \texttt{LeakyReLU}(x) = \begin{cases} x & \text{for } x \geq 0 \\ \alpha x & \text{for } x < 0, \end{cases}$$

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

and

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

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

$$\texttt{Sigmoid}(x) = \frac{1}{1+e^{-x}}, \qquad \texttt{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \qquad \texttt{Softsign}(x) = \frac{x}{1+|x|},$$

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

The remaining five activation functions are given by

$$\texttt{Softplus}(x) = \ln(1+e^x), \qquad \texttt{SiLU}(x) = \frac{x}{1+e^{-x}},$$

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

and

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

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

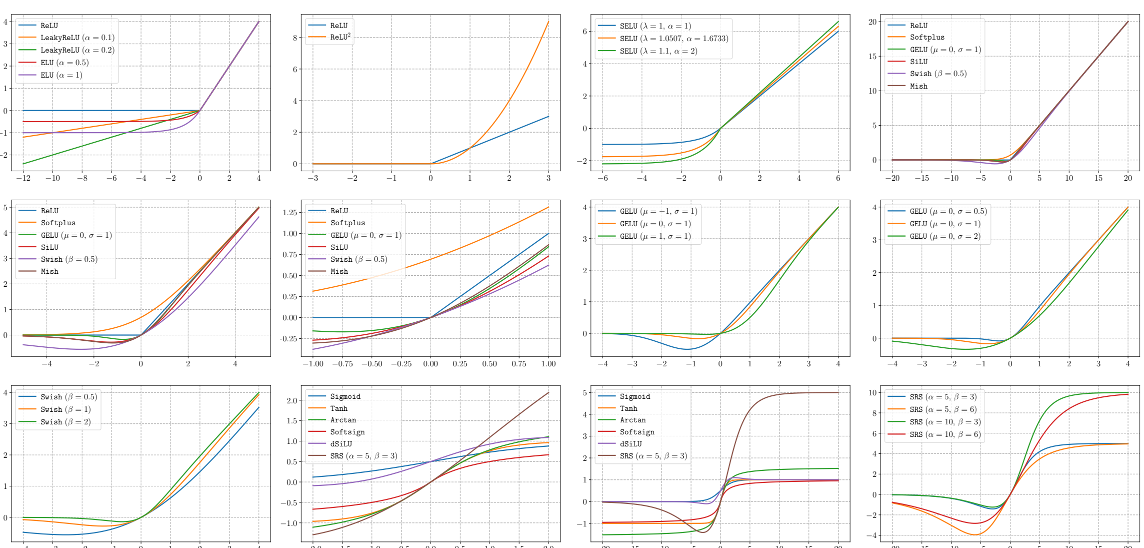

Figure 1: Illustrations of `ReLU`, `LeakyReLU`, `ReLU`^{2}, `ELU`, `SELU`, `Softplus`, `GELU`, `SiLU`, `Swish`, `Mish`, `Sigmoid`, `Tanh`, `Arctan`, `Softsign`, `dSiLU`, and `SRS`.

3 Proofs of Theorems in Sections 1 and 2

In this section, we will prove the theorems in Sections 1 and 2, i.e., Theorems 1, 6, and 7. To enhance clarity, Section 3.1 offers a concise overview of the notations employed throughout this paper. Next in Section 3.2, we present the ideas for proving Theorems 1, 6, and 7. Moreover, to simplify the proofs, we establish several propositions, which will be proved in later sections. By assuming the validity of these propositions, we provide the proof of Theorem 1 in Section 3.3 and give the proofs of Theorems 6 and 7 in Section 3.4.

3.1 Notations

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

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

  $$\|\boldsymbol{x}\|_p = \|\boldsymbol{x}\|_{\ell^p} \coloneqq \big(|x_1|^p + \cdots + |x_d|^p\big)^{1/p} \quad \text{if } p \in [1,\infty)$$

  and

  $$\|\boldsymbol{x}\|_\infty = \|\boldsymbol{x}\|_{\ell^\infty} \coloneqq \max\big\{|x_i| : i = 1, 2, \cdots, d\big\}.$$

- Let "$\rightrightarrows$" denote the uniform convergence. For example, if $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n$ is a vector-valued function and $\boldsymbol{f}_\delta(\boldsymbol{x}) \rightrightarrows \boldsymbol{f}(\boldsymbol{x})$ as $\delta \to 0^+$ for any $\boldsymbol{x} \in \Omega \subseteq \mathbb{R}^d$, then for any $\varepsilon > 0$, there exists $\delta_\varepsilon \in (0,1)$ such that

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

- A network is labeled as "a network of width N and depth L" when it satisfies the following two conditions.
  - The count of neurons in each hidden layer of the network does not exceed N.
  - The total number of hidden layers in the network is at most L.
- Suppose $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^n$ is a vector-valued function realized by a ϱ-activated network. Then $\boldsymbol{\phi}$ can be expressed as

$$\boldsymbol{x} = \widetilde{\boldsymbol{h}}_0 \xrightarrow[\boldsymbol{\mathcal{L}}_0]{\boldsymbol{W}_0,\ \boldsymbol{b}_0} \boldsymbol{h}_1 \xrightarrow{\varrho} \widetilde{\boldsymbol{h}}_1 \quad \cdots \quad \xrightarrow[\boldsymbol{\mathcal{L}}_{L-1}]{\boldsymbol{W}_{L-1},\ \boldsymbol{b}_{L-1}} \boldsymbol{h}_L \xrightarrow{\varrho} \widetilde{\boldsymbol{h}}_L \xrightarrow[\boldsymbol{\mathcal{L}}_L]{\boldsymbol{W}_L,\ \boldsymbol{b}_L} \boldsymbol{h}_{L+1} = \boldsymbol{\phi}(\boldsymbol{x}),$$

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

$$\boldsymbol{h}_{i+1} = \boldsymbol{W}_i \cdot \widetilde{\boldsymbol{h}}_i + \boldsymbol{b}_i =: \boldsymbol{\mathcal{L}}_i(\widetilde{\boldsymbol{h}}_i) \quad \text{for } i = 0, 1, \cdots, L,$$

  and

$$\widetilde{\boldsymbol{h}}_i = \varrho(\boldsymbol{h}_i) \quad \text{for } i = 1, 2, \cdots, L,$$

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

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

  Refer to Figure 2 for an illustration.

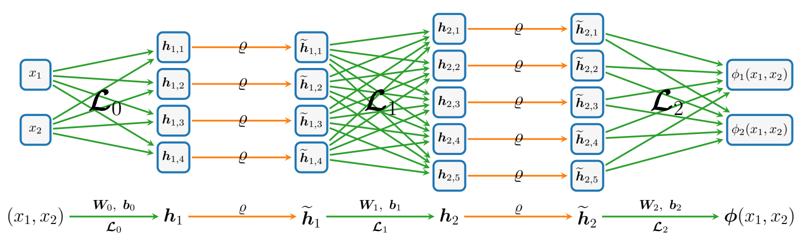

Figure 2: An example of a ϱ-activated network of width 5 and depth 2. The network realizes a vector-valued function $\boldsymbol{\phi} = (\phi_1, \phi_2)$.

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, 6, and 7. These three theorems collectively convey a narrative wherein a $\widetilde{\varrho}$-activated network can be accurately approximated by a ϱ-activated network, provided certain assumptions are met regarding ϱ and $\widetilde{\varrho}$. 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 8. *Given two functions $\varrho, \widetilde{\varrho} : \mathbb{R} \to \mathbb{R}$ with $\widetilde{\varrho} \in C(\mathbb{R})$, suppose for any $M > 0$, there exists $\widetilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{\widetilde{N},\ \widetilde{L};\ \mathbb{R} \to \mathbb{R}\}$ for each $\eta \in (0,1)$ such that*

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

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

$$\left\|\boldsymbol{\phi}_\varrho - \boldsymbol{\phi}_{\widetilde{\varrho}}\right\|_{\sup([-A,A]^d)} < \varepsilon.$$

The proof of Proposition 8 can be found in Section 4. The utilization of Proposition 8 simplifies our task of proving Theorems 1, 6, and 7. 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 9. *Given any $n \in \mathbb{N}$ and $a_0 < a < b < b_0$, if $f \in C^n\big((a_0, b_0)\big)$, then*

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

Proposition 10. *Given any $M > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathscr{A}_{1,k}$, there exists $\phi_\varepsilon \in \mathcal{NN}_\varrho\{k+2,\ 1;\ \mathbb{R} \to \mathbb{R}\}$ for each $\varepsilon \in (0,1)$ such that*

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

Proposition 11. *Given any $M > 0$ and $\varrho \in \mathscr{A}_2 \cup \mathscr{A}_3$, there exists $\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\ 2;\ \mathbb{R} \to \mathbb{R}\}$ for each $\varepsilon \in (0,1)$ such that*

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

Propositions 9, 10, and 11 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 9 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 12 later. With the uniform continuity of $f^{(n)}$ on a closed interval, Proposition 9 follows straightforwardly. See more details in Section 5.

The proof of Proposition 10 can be divided into two main steps. The first step involves demonstrating that

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

where τ can be used to generate ReLU and

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

The second step involves employing Proposition 9 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 11 is the fact $x \cdot \mathbb{1}_{\{x>0\}} = \mathtt{ReLU}(x)$ 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 with Propositions

The proof of Theorem 1 can be easily demonstrated by employing Propositions 8, 10, and 11.

Proof of Theorem 1. Since $\mathscr{A} = \big(\cup_{k=0}^{2} \mathscr{A}_{1,k}\big) \cup \mathscr{A}_2 \cup \mathscr{A}_3$, we can divide the proof into two cases: $\varrho \in \cup_{k=0}^{2}\mathscr{A}_{1,k}$ and $\varrho \in \mathscr{A}_2 \cup \mathscr{A}_3$.

We first consider the case $\varrho \in \cup_{k=0}^{2}\mathscr{A}_{1,k}$, i.e., $\varrho \in \mathscr{A}_{1,k}$ for some $k \in \{0,1,2\}$. By Proposition 10, for any $M > 0$, there exist $\widetilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{k+2,\ 1;\ \mathbb{R}\to\mathbb{R}\} \subseteq \mathcal{NN}_\varrho\{4,\ 1;\ \mathbb{R}\to\mathbb{R}\}$ for each $\eta \in (0,1)$ such that

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

Then by Proposition 8 with $\widetilde{\varrho}$ being ReLU therein, for any $\varepsilon > 0$, $A > 0$, and $\boldsymbol{\phi}_{\mathtt{ReLU}} \in \mathcal{NN}_{\mathtt{ReLU}}\{N,\ L;\ \mathbb{R}^d\to\mathbb{R}^n\}$, there exists

$$\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\big\{4N,\ L;\ \mathbb{R}^d\to\mathbb{R}^n\big\} \subseteq \mathcal{NN}_\varrho\big\{4N,\ 2L;\ \mathbb{R}^d\to\mathbb{R}^n\big\}$$

such that

$$\big\|\boldsymbol{\phi}_\varrho - \boldsymbol{\phi}_{\mathtt{ReLU}}\big\|_{\sup([-A,A]^d)} < \varepsilon.$$

Next, we consider the case $\varrho \in \mathscr{A}_2 \cup \mathscr{A}_3$. By Proposition 11, for any $M > 0$, there exist $\widetilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{4,\ 2;\ \mathbb{R}\to\mathbb{R}\}$ for each $\eta \in (0,1)$ such that

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

Then by Proposition 8 with $\widetilde{\varrho}$ being ReLU therein, for any $\varepsilon > 0$, $A > 0$, and $\boldsymbol{\phi}_{\mathtt{ReLU}} \in \mathcal{NN}_{\mathtt{ReLU}}\{N,\ L;\ \mathbb{R}^d\to\mathbb{R}^n\}$, there exists

$$\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\big\{4N,\ 2L;\ \mathbb{R}^d\to\mathbb{R}^n\big\}$$

such that

$$\big\|\boldsymbol{\phi}_\varrho - \boldsymbol{\phi}_{\mathtt{ReLU}}\big\|_{\sup([-A,A]^d)} < \varepsilon.$$

So we finish the proof of Theorem 1. ∎

3.4 Proofs of Theorems 6 and 7 with Propositions

The proofs of Theorems 6 and 7 can be straightforwardly demonstrated by utilizing Propositions 8, 9, and 10.

Proof of Theorem 6. It follows from $\varrho \in C^k(\mathbb{R})$ that $\varrho \in C^k\big((-M-1, M+1)\big)$ for any $M > 0$. By Proposition 10, we have

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

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

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

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

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

Then by Proposition 8 with $\widetilde{\varrho}$ being $\varrho^{(k)}$ therein, for any $\varepsilon > 0$, $A > 0$, and $\boldsymbol{\phi}_{\varrho^{(k)}} \in \mathcal{NN}_{\varrho^{(k)}}\{N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\}$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\big\{(k+1)N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$ such that

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

So we finish the proof of Theorem 6. ∎

Proof of Theorem 7. By Proposition 10, for any $M > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathscr{A}_{1,k}$, there exist $\widetilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{k+2,\, 1;\, \mathbb{R} \to \mathbb{R}\}$ for each $\eta \in (0,1)$ such that

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

Then by Proposition 8 with $\widetilde{\varrho}$ being $\mathtt{ReLU}$ therein, for any $\varepsilon > 0$, $A > 0$, and $\boldsymbol{\phi}_{\mathtt{ReLU}} \in \mathcal{NN}_{\mathtt{ReLU}}\{N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\}$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\big\{(k+2)N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$ such that

$$\big\|\boldsymbol{\phi}_\varrho - \boldsymbol{\phi}_{\mathtt{ReLU}}\big\|_{\sup([-A,A]^d)} < \varepsilon.$$

So we finish the proof of Theorem 7. ∎

4 Proof of Proposition 8

We will prove Proposition 8 in this section. The crucial aspect of the proof is the observation that $\widetilde{\varrho} \in C(\mathbb{R})$ implies $\widetilde{\varrho}$ is uniformly continuous on $[-M, M]$ for any $M > 0$. Further information and specific details are provided below.

Proof of Proposition 8. For ease of notation, we allow the activation function to be applied elementwise to a vector input. Since $\boldsymbol{\phi}_{\widetilde{\varrho}} \in \mathcal{NN}_{\widetilde{\varrho}}\big\{N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$, $\boldsymbol{\phi}_{\widetilde{\varrho}}$ is realized by a $\widehat{L}$-hidden-layer $\widetilde{\varrho}$-activated network, where $L \ge \widehat{L} \in \mathbb{N}^+$. We may assume $\widehat{L} = L$ since the proof remains similar if we replace L with $\widehat{L}$ when $\widehat{L} < L$. Then $\boldsymbol{\phi}_{\widetilde{\varrho}}$ can be represented in a form of function compositions

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

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

Recall that there exists

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

such that

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

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

$$\boldsymbol{\phi}_{\widetilde{\varrho}_\eta}(\boldsymbol{x}) \coloneqq \boldsymbol{\mathcal{L}}_L \circ \widetilde{\varrho}_\eta \circ \boldsymbol{\mathcal{L}}_{L-1} \circ \ \cdots \ \circ \widetilde{\varrho}_\eta \circ \boldsymbol{\mathcal{L}}_1 \circ \widetilde{\varrho}_\eta \circ \boldsymbol{\mathcal{L}}_0(\boldsymbol{x}) \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^d.$$

It is easy to verify that

$$\boldsymbol{\phi}_{\widetilde{\varrho}_\eta} \in \mathcal{NN}_\varrho\big\{\widetilde{N} \cdot N,\ \ \widetilde{L} \cdot L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}.$$

Moveover, we will prove

$$\boldsymbol{\phi}_{\widetilde{\varrho}_\eta}(\boldsymbol{x}) \rightrightarrows \boldsymbol{\phi}_{\widetilde{\varrho}}(\boldsymbol{x}) \quad \text{as } \eta \to 0^+ \quad \text{for any } \boldsymbol{x} \in [-A, A]^d.$$

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

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

and

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

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

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

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

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

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

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

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

$$M = \sup\Big\{\|\boldsymbol{h}_j(\boldsymbol{x})\|_{\ell^\infty} + 1 : \boldsymbol{x} \in [-A, A]^d, \quad j = 1, 2, \cdots, L+1\Big\},$$

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

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

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

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

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

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

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

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

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

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

implying

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

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

By the principle of induction, we have

$$\boldsymbol{\phi}_{\widetilde{\varrho}_\eta}(\boldsymbol{x}) = \boldsymbol{h}_{L+1,\eta}(\boldsymbol{x}) \rightrightarrows \boldsymbol{h}_{L+1}(\boldsymbol{x}) = \boldsymbol{\phi}_{\widetilde{\varrho}}(\boldsymbol{x}) \quad \text{as } \eta \to 0^+ \quad \text{for any } \boldsymbol{x} \in [-A, A]^d.$$

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

$$\big\|\boldsymbol{\phi}_{\widetilde{\varrho}_{\eta_0}} - \boldsymbol{\phi}_{\widetilde{\varrho}}\big\|_{\sup([-A,A]^d)} < \varepsilon.$$

By defining $\boldsymbol{\phi}_\varrho \coloneqq \boldsymbol{\phi}_{\widetilde{\varrho}_{\eta_0}}$, we have

$$\boldsymbol{\phi}_\varrho = \boldsymbol{\phi}_{\widetilde{\varrho}_{\eta_0}} \in \mathcal{NN}_\varrho\big\{\widetilde{N} \cdot N,\ \ \widetilde{L} \cdot L;\ \mathbb{R}^d \to \mathbb{R}^n\big\}$$

and

$$\big\|\boldsymbol{\phi}_\varrho - \boldsymbol{\phi}_{\widetilde{\varrho}}\big\|_{\sup([-A,A]^d)} = \big\|\boldsymbol{\phi}_{\widetilde{\varrho}_{\eta_0}} - \boldsymbol{\phi}_{\widetilde{\varrho}}\big\|_{\sup([-A,A]^d)} < \varepsilon.$$

So we finish the proof of Proposition 8. ∎

5 Proof of Proposition 9

In this section, our goal is to prove Proposition 9. 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.

5.1 A Lemma for Proving Proposition 9

Lemma 12. *Given any $n \in \mathbb{N}$, it holds that*

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

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

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

By assuming the validity of the claim, we have

$$\begin{aligned}\sum_{\ell=0}^{n}(-1)^{\ell}\binom{n}{\ell}\ell^{i} = \lim_{t\to-1^{+}}\sum_{\ell=0}^{n} t^{\ell}\binom{n}{\ell}\ell^{i} &= \lim_{t\to-1^{+}}(1+t)^{n-i}\Big(\frac{n!}{(n-i)!}t^{i} + (1+t)p_i(t)\Big) \\ &= \begin{cases} 0 & \text{if } i \in \{0,1,\cdots,n-1\}, \\ (-1)^{n}\, n! & \text{if } i = n. \end{cases}\end{aligned}$$

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

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

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

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

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

$$\sum_{\ell=0}^{n} t^{\ell}\binom{n}{\ell}\ell^{j} = (1+t)^{n-j}\Big(\underbrace{\frac{n!}{(n-j)!}t^{j} + (1+t)p_j(t)}_{\widetilde{p}_j(t)}\Big) = (1+t)^{n-j}\widetilde{p}_j(t)$$

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

$$\begin{aligned}\sum_{\ell=0}^{n} \ell t^{\ell-1}\binom{n}{\ell}\ell^{j} &= (n-j)(1+t)^{n-j-1}\widetilde{p}_j(t) + (1+t)^{n-j}\tfrac{d}{dt}\widetilde{p}_j(t) \\ &= (1+t)^{n-j-1}\Big((n-j)\widetilde{p}_j(t) + (1+t)\tfrac{d}{dt}\widetilde{p}_j(t)\Big)\end{aligned}$$

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

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

for any $t \in (-1,0)$, where $p_{j+1}(t) = t(n-j)p_j(t) + t\frac{d}{dt}\widetilde{p}_j(t)$ 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 12. ■

5.2 Proof of Proposition 9 with Lemma 12

Equipped with Lemma 12, we are prepared to demonstrate the proof of Proposition 9.

Proof of Proposition 9. We may assume $n \in \mathbb{N}^+$ since the case $n = 0$ is trivial. For each $x \in [a,b]$, we define

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

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

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

It follows from $f \in C^n\big((a_0,b_0)\big)$ 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

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

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

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

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

implying

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

where the last equality comes from Lemma 12.

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 $0 < |t_{x,n}| < \cdots < |t_{x,1}| < |t| < \delta_0$ such that

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

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

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

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

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

where the strict inequality comes from Equation (2).

Set $\lambda_\ell = \frac{(-1)^\ell \binom{n}{\ell} \ell^n}{(-1)^n\, n!}$ for $\ell = 0, 1, \cdots, n$. By Lemma 12, 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

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

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

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

So we finish the proof of Proposition 9. ∎

6 Proof of Proposition 10

The objective of this section is to provide the proof of Proposition 10. To streamline the proof process, we first introduce a lemma in Section 6.1. Subsequently, we present the comprehensive proof in Section 6.2.

6.1 A Lemma for Proving Proposition 10

Lemma 13. *Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function with $f'(x_0) \neq 0$ for some $x_0 \in \mathbb{R}$. Then for any $M > 0$, it holds that*

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

Proof. Clearly,

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

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

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

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

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

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

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

Thus, we have

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

So we finish the proof of Lemma 13. ∎

6.2 Proof of Proposition 10 with Lemma 13

With Lemma 13 in hand, we are ready to present the proof of Proposition 10.

Proof of Proposition 10. Given any $\varepsilon \in (0,1)$, our goal is to construct $\phi_\varepsilon \in \mathcal{NN}_\varrho\{(k+2), 1;\ \mathbb{R} \to \mathbb{R}\}$ with $\varrho \in \mathscr{A}_{1,k}$ to approximate `ReLU` well on $[-M, M]$.

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

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

Set

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

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

$$\Big|\tfrac{\varrho^{(k)}(x_0+t)-\varrho^{(k)}(x_0)}{t} - \big(L_1 \cdot \mathbb{1}_{\{t<0\}} + L_2 \cdot \mathbb{1}_{\{t>0\}}\big)\Big| < \varepsilon/(4KM)$$

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

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

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

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

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

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

implying

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

where

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

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

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

implying

$$\tfrac{1}{L_2-L_1}\psi(x) - \tfrac{L_1}{L_2-L_1}x = \mathtt{ReLU}(x).$$

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 $k \ge 1$.

Case 1: $k = 0$.

First, let us consider the case of $k = 0$. In this case, $\varrho^{(k)} = \varrho$. For each $\varepsilon \in \big(0, \frac{\delta_\varepsilon}{2M}\big)$ 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

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

Define

$$\begin{aligned}\phi_\varepsilon(x) &\coloneqq \tfrac{1}{L_2-L_1}\psi_\varepsilon(x) - \tfrac{1}{L_2-L_1}\Big(\psi_\varepsilon(x-M) + L_1 M\Big)\\ &= \tfrac{1}{L_2-L_1}\tfrac{\varrho(x_0+\varepsilon x)-\varrho(x_0)}{\varepsilon} - \tfrac{1}{L_2-L_1}\Big(\tfrac{\varrho(x_0+\varepsilon(x-M))-\varrho(x_0)}{\varepsilon} + L_1 M\Big)\end{aligned}$$

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

$$\begin{aligned}|\phi_\varepsilon(x) - \mathtt{ReLU}(x)| &= \Bigg|\underbrace{\tfrac{1}{L_2-L_1}\psi_\varepsilon(x) - \tfrac{1}{L_2-L_1}\Big(\psi_\varepsilon(x-M) + L_1 M\Big)}_{\phi_\varepsilon} - \Big(\underbrace{\tfrac{1}{L_2-L_1}\psi(x) - \tfrac{L_1}{L_2-L_1}x}_{\mathtt{ReLU}}\Big)\Bigg|\\ &\le \big|\tfrac{1}{L_2-L_1}\big|\cdot\big|\psi_\varepsilon(x) - \psi(x)\big| + \big|\tfrac{1}{L_2-L_1}\big|\cdot\Big|\Big(\psi_\varepsilon(x-M) + L_1 M\Big) - L_1 x\Big|\\ &< K\cdot\tfrac{\varepsilon}{2K} + K\cdot\tfrac{\varepsilon}{2K} = \varepsilon,\end{aligned}$$

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

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

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

Case 2: $k \ge 1$.

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

$$\widetilde{\phi}_\varepsilon(x) \coloneqq \tfrac{1}{L_2-L_1}\psi_\varepsilon(x) - \tfrac{L_1}{L_2-L_1}x \quad \text{for any } x \in \mathbb{R}.$$

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

$$\begin{aligned}\big|\widetilde{\phi}_\varepsilon(x) - \mathtt{ReLU}(x)\big| &= \Big|\Big(\tfrac{1}{L_2-L_1}\psi_\varepsilon(x) - \tfrac{L_1}{L_2-L_1}x\Big) - \Big(\tfrac{1}{L_2-L_1}\psi(x) - \tfrac{L_1}{L_2-L_1}x\Big)\Big|\\ = \Big|\tfrac{1}{L_2-L_1}\psi_\varepsilon(x) - \tfrac{1}{L_2-L_1}\psi(x)\Big| &\le \big|\tfrac{1}{L_2-L_1}\big|\cdot\big|\psi_\varepsilon(x) - \psi(x)\big| < K\cdot\tfrac{\varepsilon}{2K} = \varepsilon/2,\end{aligned} \tag{5}$$

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

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

for any $x \in [-M, M]$ and $\varepsilon \in \big(0, \tfrac{\delta_\varepsilon}{2M}\big)$. 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 $\varepsilon \in \big(0, \tfrac{\delta_\varepsilon}{2M}\big)$.

Recall that $\varrho \in C^k\big((a_0, b_0)\big)\backslash C^{k+1}\big((a_0, b_0)\big)$ with $k \ge 1$. 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

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

By Lemma 13,

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

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 $\varrho \in C^k\big((a_0,b_0)\big)$. By Proposition 9,

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

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

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

and

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

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

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

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

$$\big|\phi_\varepsilon(x) - \texttt{ReLU}(x)\big| \le \big|\phi_\varepsilon(x) - \widetilde{\phi}_\varepsilon(x)\big| + \big|\widetilde{\phi}_\varepsilon(x) - \texttt{ReLU}(x)\big| < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

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

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

So we finish the proof of Proposition 10. ■

7 Proof of Proposition 11

We will prove Proposition 11 in this section. To this end, we first establish two lemmas in Section 7.1, which play important roles in proving Proposition 11. Next, we give the detailed proof of Proposition 11 based on these two lemmas in Section 7.2.

7.1 Lemmas for Proving Proposition 11

Lemma 14. *Given any $A>0$, suppose $\varrho:\mathbb{R}\to\mathbb{R}$ is a function with $\varrho''(x_0)\neq 0$ for some $x_0\in\mathbb{R}$. Then there exists*

$$\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\ 1;\ \mathbb{R}^2\to\mathbb{R}\} \quad \text{for each } \varepsilon\in(0,1)$$

such that

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

Proof. By L'Hôpital's Rule,

$$\begin{aligned}\lim_{t\to 0}\frac{\varrho(x_0+t)+\varrho(x_0-t)-2\varrho(x_0)}{t^2} &= \lim_{t\to 0}\frac{\varrho'(x_0+t)-\varrho'(x_0-t)}{2t}\\ =\lim_{t\to 0}\frac{\varrho'(x_0+t)-\varrho'(x_0)+\varrho'(x_0)-\varrho'(x_0-t)}{2t} &= \varrho''(x_0)/2+\varrho''(x_0)/2=\varrho''(x_0)\neq 0.\end{aligned}$$

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

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

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

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

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

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

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

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

Therefore, for any $x,y\in[-A,A]$, we have $x+y,x-y\in[-2A,2A]$, implying

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

Then, by defining

$$\phi_\varepsilon(x,y) := \tfrac{1}{4}\big(\psi_\varepsilon(x+y)-\psi_\varepsilon(x-y)\big) \quad \text{for any } x,y\in\mathbb{R},$$

we have

$$\phi_\varepsilon(x,y)\rightrightarrows \tfrac{1}{4}\big((x+y)^2-(x-y)^2\big)=xy \quad \text{as } \varepsilon\to 0^+ \quad \text{for any } x,y\in[-A,A].$$

Furthermore, as shown in Figure 3, $\phi_\varepsilon\in\mathcal{NN}_\varrho\{4,\ 1;\ \mathbb{R}^2\to\mathbb{R}\}$. Thus, we finish the proof of Lemma 14. ∎

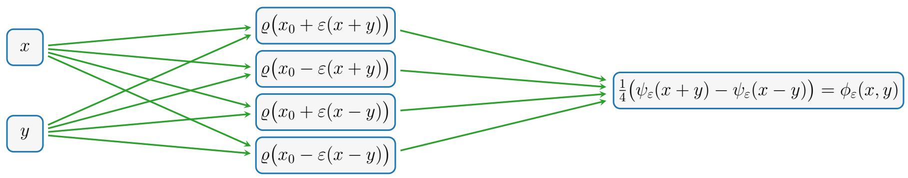

Figure 3: An illustration of the network architecture realizing ϕ_ε by Equation (7.1).

Lemma 15. *Given any $M>0$ and two functions $g_1, g_{2,\delta}:\mathbb{R}\to\mathbb{R}$ for each $\delta\in(0,1)$, suppose*

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

and

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

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

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

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

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

and there exists $K_1>0$ such that

$$\big|g_1(x)\big|<\varepsilon_1 \text{ for any } x\le -K_1/4\quad\text{and}\quad \big|g_1(x)-1\big|<\varepsilon_1 \text{ for any } x\ge K_1/4,$$

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

$$\big|g_1(K_0K_1x/\varepsilon)-\mathbb{1}_{\{x>0\}}\big|<\varepsilon_1=\varepsilon/(2M)\quad\text{for any } |x|\ge\varepsilon/(4K_0),\tag{7}$$

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

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

Observe that $\texttt{ReLU}(x)=x\cdot\mathbb{1}_{\{x>0\}}$ for any $x\in\mathbb{R}$. Setting $K_\varepsilon=K_0K_1/\varepsilon$ and by Equation (8), for any $x\in[-M,M]$, we have

$$\begin{aligned}
\big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x)-\texttt{ReLU}(x)\big| &= \big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x)-x\cdot\mathbb{1}_{\{x>0\}}\big|\\
&\le \big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x)-xg_1(K_\varepsilon x)\big|+\big|xg_1(K_\varepsilon x)-x\cdot\mathbb{1}_{\{x>0\}}\big|\\
&\le \big|g_1(K_\varepsilon x)\big|\cdot\big|g_{2,\delta_\varepsilon}(x)-x\big|+|x|\cdot\big|g_1(K_\varepsilon x)-\mathbb{1}_{\{x>0\}}\big|\\
&\le K_0\cdot\varepsilon_2+|x|\cdot\big|g_1(K_0K_1x/\varepsilon)-\mathbb{1}_{\{x>0\}}\big|.
\end{aligned}$$

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

$$\begin{aligned}
\big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x)-\texttt{ReLU}(x)\big| &\le K_0\cdot\varepsilon_2+|x|\cdot\big|g_1(K_0K_1x/\varepsilon)-\mathbb{1}_{\{x>0\}}\big|\\
&\le K_0\cdot\tfrac{\varepsilon}{3K_0}+\tfrac{\varepsilon}{4K_0}\cdot(K_0+1)\le\varepsilon/3+\varepsilon/2<\varepsilon.
\end{aligned}$$

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

$$\begin{aligned}\left|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x) - \mathtt{ReLU}(x)\right| &\le K_0 \cdot \varepsilon_2 + |x| \cdot \left|g_1(K_0K_1x/\varepsilon) - \mathbb{1}_{\{x>0\}}\right| \\ &\le K_0 \cdot \varepsilon_2 + M \cdot \varepsilon_1 \le K_0 \cdot \tfrac{\varepsilon}{3K_0} + M \cdot \tfrac{\varepsilon}{2M} \le \varepsilon/3 + \varepsilon/2 < \varepsilon\end{aligned}$$

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

$$\left|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x) - \mathtt{ReLU}(x)\right| < \varepsilon,$$

which means we finish the proof. ■

7.2 Proof of Proposition 11 with Lemmas 14 and 15

Having established Lemmas 14 and 15 in Section 7.1, we are now prepared to prove Proposition 11.

Proof of Proposition 11. For any $\varepsilon \in (0, 1)$, our goal is to construct $\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\ 2;\ \mathbb{R} \to \mathbb{R}\}$ with $\varrho \in \mathscr{A}_2 \cup \mathscr{A}_3$ to approximate $\mathtt{ReLU}$ well on $[-M, M]$. We divide the proof into two cases: $\varrho \in \mathscr{A}_2$ and $\varrho \in \mathscr{A}_3$.

Case 1: $\varrho \in \mathscr{A}_2$.

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

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

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

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

where

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

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

By defining

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

we have

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

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

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

By Lemma 13,

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

By Lemma 15, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0,1)$ such that

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

It follows from Equation (9) that

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

Since $\varrho''(x_0) \neq 0$, by Lemma 14, there exists

$$\Gamma_\eta \in \mathcal{NN}_\varrho\{4,\ 1;\ \mathbb{R}^2 \to \mathbb{R}\} \quad \text{for each } \eta \in (0,1)$$

such that

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

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

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

Define

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

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

$$\begin{aligned} &\big|\phi_\varepsilon(x) - \mathtt{ReLU}(x)\big| = \Big|\Gamma_{\eta_\varepsilon}\Big(g_1(K_\varepsilon x),\ g_{2,\delta_\varepsilon}(x)\Big) - \mathtt{ReLU}(x)\Big| \\ &\le \Big|\Gamma_{\eta_\varepsilon}\Big(g_1(K_\varepsilon x),\ g_{2,\delta_\varepsilon}(x)\Big) - g_1(K_\varepsilon x)\cdot g_{2,\delta_\varepsilon}(x)\Big| + \Big|g_1(K_\varepsilon x)\cdot g_{2,\delta_\varepsilon}(x) - \mathtt{ReLU}(x)\Big| \\ &< \varepsilon + \varepsilon = 2\varepsilon \end{aligned}$$

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

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

We still need to demonstrate that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\ 2;\ \mathbb{R}\to\mathbb{R}\}$. By defining

$$\boldsymbol{\psi}_\varepsilon(x) \coloneqq \Big(\tfrac{\varrho(K_\varepsilon x+T_0)-\varrho(K_\varepsilon x)-L_1}{L_2-L_1},\ \tfrac{\varrho(x_1+\delta_\varepsilon x)-\varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)}\Big) \quad \text{for any } x \in \mathbb{R},$$

we have $\boldsymbol{\psi}_\varepsilon \in \mathcal{NN}_\varrho\{3,\ 1;\ \mathbb{R}\to\mathbb{R}^2\}$ and

$$\begin{aligned} \phi_\varepsilon(x) &= \Gamma_{\eta_\varepsilon}\Big(g_1(K_\varepsilon x),\ g_{2,\delta_\varepsilon}(x)\Big) \\ &= \Gamma_{\eta_\varepsilon}\Big(\tfrac{\varrho(K_\varepsilon x+T_0)-\varrho(K_\varepsilon x)-L_1}{L_2-L_1},\ \tfrac{\varrho(x_1+\delta_\varepsilon x)-\varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)}\Big) = \Gamma_{\eta_\varepsilon} \circ \boldsymbol{\psi}_\varepsilon(x) \end{aligned}$$

for any $x \in \mathbb{R}$. Recall that $\Gamma_{\eta_\varepsilon} \in \mathcal{NN}_\varrho\{4,\, 1;\; \mathbb{R}^2 \to \mathbb{R}\}$. Therefore, we have $\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\, 2;\; \mathbb{R} \to \mathbb{R}\}$, as required.

Case 2: $\varrho \in \mathscr{A}_3$.

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

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

By defining

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

we have

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

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

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

By Lemma 13,

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

By Lemma 15, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0,1)$ such that

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

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

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

Since $\varrho''(x_0) \neq 0$, by Lemma 14, there exists

$$\Gamma_\eta \in \mathcal{NN}_\varrho\{4,\, 1;\; \mathbb{R}^2 \to \mathbb{R}\} \quad \text{for each } \eta \in (0,1)$$

such that

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

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

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

Define

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

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

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

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

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

It remains to show $\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\, 2;\ \mathbb{R}\to\mathbb{R}\}$. By defining

$$\boldsymbol{\psi}_\varepsilon(x) := \Big(\tfrac{\varrho(K_\varepsilon x)-L_1}{L_2-L_1},\ \tfrac{\varrho(x_1+\delta_\varepsilon x)-\varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)}\Big) \quad \text{for any } x \in \mathbb{R},$$

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

$$\phi_\varepsilon(x) = \Gamma_{\eta_\varepsilon}\Big(g_1(K_\varepsilon x),\, g_{2,\delta_\varepsilon}(x)\Big) = \Gamma_{\eta_\varepsilon}\Big(\tfrac{\varrho(K_\varepsilon x)-L_1}{L_2-L_1},\ \tfrac{\varrho(x_1+\delta_\varepsilon x)-\varrho(x_1)}{\delta_\varepsilon \varrho'(x_1)}\Big) = \Gamma_{\eta_\varepsilon}\circ \boldsymbol{\psi}_\varepsilon(x)$$

for any $x \in \mathbb{R}$. Recall that $\Gamma_{\eta_\varepsilon} \in \mathcal{NN}_\varrho\{4,\, 1;\ \mathbb{R}^2 \to \mathbb{R}\}$. Hence, we can conclude that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{4,\, 2;\ \mathbb{R}\to\mathbb{R}\}$. This result completes the proof of Proposition 11. ■

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

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. ISSN 2079–7370. doi: https://doi.org/10.4208/eajam.2022-270.070123. URL http://global-sci.org/intro/article_detail/eajam/21721.html.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993. ISSN 0018-9448. URL https://doi.org/10.1109/18.256500.

Andrew R. Barron and Jason M. Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv e-prints*, art. arXiv:1809.03090, September 2018. URL https://arxiv.org/abs/1809.03090.

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. doi: 10.1137/18M118709X. URL https://doi.org/10.1137/18M118709X.

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. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/2f4b6febe0b70805c3be75e5d6a66918-Paper-Conference.pdf.

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. URL https://proceedings.neurips.cc/paper/2019/file/fd95ec8df5dbeea25aa8e6c808bad583-Paper.pdf.

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. ISSN 2297-4687. doi: 10.3389/fams.2018.00014. URL https://www.frontiersin.org/article/10.3389/fams.2018.00014.

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. URL http://arxiv.org/abs/1511.07289.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2:303–314, 1989. URL https://doi.org/10.1007/BF02551274.

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. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.12.012. URL https://www.sciencedirect.com/science/article/pii/S0893608017302976. Special issue on deep reinforcement learning.

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. URL https://proceedings.mlr.press/v15/glorot11a.html.

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive Approximation*, 55:259–367, 2022. URL https://doi.org/10.1007/s00365-021-09543-4.

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. doi: 10.1142/S0219530519410021. URL https://doi.org/10.1142/S0219530519410021.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv e-prints*, art. arXiv:1606.08415, June 2016. doi: 10.48550/arXiv.1606.08415.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(91)90009-T. URL http://www.sciencedirect.com/science/article/pii/089360809190009T.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL http://www.sciencedirect.com/science/article/pii/0893608089900208.

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. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf.

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. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

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. ISBN 978-3-030-60635-0. doi: 10.1007/978-3-030-60636-7_26. URL https://doi.org/10.1007/978-3-030-60636-7_26.

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. URL https://doi.org/10.4171/JEMS/1221.

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. URL https://doi.org/10.1137/20M134695X.

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.

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. URL https://www.bmvc2020-conference.com/assets/papers/0928.pdf.

Hadrien Montanelli and Haizhao Yang. Error bounds for deep ReLU networks using the Kolmogorov-Arnold superposition theorem. *Neural Networks*, 129:1–6, 2020. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.12.013. URL http://www.sciencedirect.com/science/article/pii/S0893608019304058.

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. ISBN 9781605589077.

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. URL http://jmlr.org/papers/v21/20-002.html.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv e-prints*, art. arXiv:1710.05941, October 2017. doi: 10.48550/arXiv.1710.05941.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.07.011. URL http://www.sciencedirect.com/science/article/pii/S0893608019301996.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020. ISSN 1991-7120. URL https://doi.org/10.4208/cicp.OA-2020-0149.

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, 2022a. URL http://jmlr.org/papers/v23/21-1404.html.

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 2022b. URL https://proceedings.mlr.press/v162/shen22g.html.

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. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/257be12f31dfa7cc158dda99822c6fd1-Abstract-Conference.html.

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. ISSN 0021-7824. doi: https://doi.org/10.1016/j.matpur.2021.07.009. URL https://www.sciencedirect.com/science/article/pii/S0021782421001124.

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. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.005. URL https://www.sciencedirect.com/science/article/pii/S1063520321001056.

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. URL https://openreview.net/forum?id=H1ebTsActm.

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.

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. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2017.07.002. URL http://www.sciencedirect.com/science/article/pii/S0893608017301545.

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. URL http://proceedings.mlr.press/v75/yarotsky18a.html.

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.

Shijun Zhang, Jianfeng Lu, and Hongkai Zhao. On enhancing expressive power via compositions of single fixed-size ReLU network. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41452–41487. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/zhang23ad.html.

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*, art. arXiv:2306.17301, June 2023b. doi: 10.48550/arXiv.2306.17301.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2019.06.004. URL http://www.sciencedirect.com/science/article/pii/S1063520318302045.