Journal of Machine Learning Research 25 (2024) 1-39 Submitted 7/23; Revised 10/23; Published 1/24

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

Editor: Joan Bruna

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`, $\texttt{ReLU}^2$, `ELU`, `CELU`, `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 $3N$ 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, albeit with slightly increased constants. Significantly, we establish that the (width, depth) scaling factors can be further reduced from $(3, 2)$ to $(1, 1)$ if ϱ falls within a specific subset of $\mathscr{A}$. This subset includes activation functions such as `ELU`, `CELU`, `SELU`, `Softplus`, `GELU`, `SiLU`, `Swish`, and `Mish`.

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

1. Introduction

In the realm of artificial intelligence, 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 an affine 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.

* Corresponding author.

The adoption of `ReLU` has led to significant improvements in 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 `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 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.

1.1 Definition of Activation Function Set

To the best of our knowledge, the majority of commonly used activation functions can be generally classified into three distinct categories. The initial category primarily comprises piecewise smooth functions, e.g., `ReLU`, `LeakyReLU`, `ReLU`^{2} (`ReLU` squared) (Siegel and Xu, 2022), `ELU`, `CELU` (Continuously Differentiable `ELU`) (Barron, 2017), and `SELU` (Scaled `ELU`) (Klambauer et al., 2017). All these activation functions are included in $\bigcup_{k=0}^{\infty}\mathscr{A}_{1,k}$, where $\mathscr{A}_{1,k}$, for each smoothness index $k\in\mathbb{N}$, is defined as

$$\mathscr{A}_{1,k}\coloneqq\big\{\varrho:\mathbb{R}\to\mathbb{R}\;\big|\;\mathscr{K}_k(\varrho)\neq\varnothing\big\},$$

where $\mathscr{K}_k(\varrho)$ represents the set of k-th order "kinks" of $\varrho:\mathbb{R}\to\mathbb{R}$. A point $x_0\in\mathbb{R}$ is referred to as a k-th order "kink" of ϱ if there exist $a_0,b_0\in\mathbb{R}$ such that $a_0<x_0<b_0$, $\varrho\in C^k\big((a_0,b_0)\big)$, and

$$\mathbb{R}\ni\lim_{t\to0^-}\frac{\varrho^{(k)}(x_0+t)-\varrho^{(k)}(x_0)}{t}\neq\lim_{t\to0^+}\frac{\varrho^{(k)}(x_0+t)-\varrho^{(k)}(x_0)}{t}\in\mathbb{R}.$$

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 $\varrho^{(k)}\in C\big((a_0,b_0)\big)$ 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 $f : \Omega \subseteq \mathbb{R} \to \mathbb{R}$. 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 $\partial^{\boldsymbol{\alpha}} f$ exist and are continuous for any $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\sum_{i=1}^d \alpha_i \leq k$. In particular, when $k = 0$, we denote $C(\Omega)$ as $C^0(\Omega)$, which represents the set of continuous functions on Ω.

The second category primarily encompasses smooth variations of `ReLU`, e.g., `Softplus` (Glorot et al., 2011), `GELU`, `SiLU` (Sigmoid Linear Unit) (Elfwing et al., 2018; Hendrycks and Gimpel, 2016), `Swish` (Ramachandran et al., 2017), and `Mish` (Misra, 2020). All these activation functions are encompassed in the set $\mathscr{A}_2$, which is defined via

$$\mathscr{A}_2 \coloneqq \Big\{ \varrho : \mathbb{R} \to \mathbb{R} \;\Big|\; \forall\, x \in \mathbb{R},\ \varrho(x) \coloneqq (x + b_0) \cdot h(x) + b_1, \quad b_0, b_1 \in \mathbb{R}, \quad h \in \mathscr{S} \Big\},$$

where $\mathscr{S}$ represents a collection of functions referred to as S-shaped functions, defined as

$$\mathscr{S} \coloneqq \Big\{ h : \mathbb{R} \to \mathbb{R} \;\Big|\; \sup_{x \in \mathbb{R}} |h(x)| < \infty, \quad \mathbb{R} \ni \lim_{x \to -\infty} h(x) \neq \lim_{x \to \infty} h(x) \in \mathbb{R} \Big\}.$$

Evidently, activation functions such as `GELU`, `SiLU`, `Swish`, and `Mish` are members of $\mathscr{A}_2$. It is worth highlighting that `Softplus`, `ELU`, `CELU`, and `SELU` also belong to $\mathscr{A}_2$, even though this may not be immediately apparent. Let us take `Softplus` as an example to illustrate this point, and similar reasoning applies to the other cases. Define $h(x) \coloneqq \frac{\ln(1+e^x) - \ln 2}{x}$ for any $x \neq 0$ and $h(0) = \frac{1}{2}$, where e represents the base of the natural logarithm. Consequently, we have $\texttt{Softplus}(x) = \ln(1 + e^x) = x \cdot h(x) + \ln 2$ for any $x \in \mathbb{R}$. It is then straightforward to verify that `Softplus` is indeed a member of $\mathscr{A}_2$. We would like to point out that the primary idea behind defining $\mathscr{A}_2$ is to replace the step function $\mathbb{1}_{\{x>0\}}$ in $\texttt{ReLU}(x) = x \cdot \mathbb{1}_{\{x>0\}}$ with a (smooth) S-shaped function. This insight allows us to create numerous examples within $\mathscr{A}_2$. For instance, one can define $\varrho(x) \coloneqq x \cdot h(x)$, where $h : \mathbb{R} \to [0, 1]$ represents a cumulative distribution function of a real-valued random variable. Notably, `GELU` is defined in this manner, with h being the (standard) Gaussian cumulative distribution function.

The final category is primarily composed of S-shaped activation functions with particular regularity, e.g., `Sigmoid`, `Tanh`, `Arctan`, and `Softsign` (Turian et al., 2009). All these functions are included in the set $\mathscr{A}_3$, which is defined via

$$\mathscr{A}_3 \coloneqq \big\{ \varrho : \mathbb{R} \to \mathbb{R} \;\big|\; \varrho \in \mathscr{S}, \quad \exists\, x_0 \in \mathbb{R},\ \varrho''(x_0) \neq 0 \big\}.$$

The set $\mathscr{A}_3$ includes a wide range of activation functions, with certain ones featuring discontinuities. In addition to the examples mentioned earlier, there exist numerous functions in the set $\mathscr{A}_3$, such as `dSiLU` (the derivative of `SiLU`) as introduced in (Elfwing et al., 2018), and `SRS` (Soft-Root-Sign) discussed in (Li and Zhou, 2020). Furthermore, the derivatives of `Softplus`, `GELU`, `SiLU`, `Swish`, and `Mish` fall into the category of $\mathscr{A}_3$.

Then the activation function set $\mathscr{A}$ is defined as the union of $\mathscr{A}_{1,0}$, $\mathscr{A}_{1,1}$, $\mathscr{A}_2$, and $\mathscr{A}_3$, which can be expressed as

$$\mathscr{A} \coloneqq \big(\mathscr{A}_{1,0} \cup \mathscr{A}_{1,1}\big) \cup \mathscr{A}_2 \cup \mathscr{A}_3.$$

Throughout the entirety of this paper, the definitions of $\mathscr{A}$, $\mathscr{A}_{1,k}$ for $k \in \mathbb{N}$, $\mathscr{A}_2$, and $\mathscr{A}_3$ will remain consistent. It is worth noting that if $\varrho \in \mathscr{A}$, then its variant $x \mapsto w_1\varrho(w_0x+b_0)+b_1$ is also in $\mathscr{A}$ provided $w_0w_1 \neq 0$. Notably, the set $\mathscr{A}$ encompasses the majority of commonly used activation functions, such as `ReLU`, `LeakyReLU`, `ReLU`^{2}, `ELU`, `CELU`, `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 bounded vector-valued function $\boldsymbol{f} : \Omega \subseteq \mathbb{R}^d \to \mathbb{R}^n$ via

$$\|\boldsymbol{f}\|_{\sup(\Omega)} \coloneqq \sup\big\{|f_i(\boldsymbol{x})| : \boldsymbol{x} \in \Omega,\ i \in \{1, 2, \cdots, n\}\big\},$$

where f_i is the i-th component of $\boldsymbol{f}$ for $i = 1, 2, \cdots, n$. 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 vector-valued 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, respectively, 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 $N = \max\{N_1, N_2, \cdots, N_L\}$.

1.2 Main Results

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 $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\{3N,\, 2L;\, \mathbb{R}^d \to \mathbb{R}^n\}$ 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 $3N$ and $2L$ arbitrarily well on any bounded set for any pre-specified $\varrho \in \mathscr{A}$. In other words, $\mathcal{NN}_\varrho\{3N,\, 2L;\, \mathbb{R}^d \to \mathbb{R}^n\}$ 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 $\varrho \in \mathscr{A}$. Indeed, this implies that networks activated by $\varrho \in \mathscr{A}$ possess, at the very least, a comparable level of expressive capability as `ReLU` networks, providing valuable insights for the development of new activation functions. Constraining ϱ to $\mathscr{A}$ is a relatively simple process, and it serves to ensure the effective

expressiveness of ϱ-activated networks. This, in turn, enables us to direct our attention more towards the learning and numerical properties of ϱ.

It is worth mentioning while Theorem 1 covers activation functions $\varrho \in \mathscr{A}_{1,k}$ only for $k = 0, 1$, 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. Additionally, we would like to emphasize that the (width, depth) scaling factors appeared in Theorem 1 as $(3, 2)$ have the potential to be reduced to $(2, 1)$ or even $(1, 1)$ under certain circumstances, as elaborated in Table 1 later on.

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*

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

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

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.

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

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 $\phi \in \mathcal{NN}_\varrho\{3\lceil 3q/2 \rceil q,\, 4\lceil \log_2 q \rceil + 2;\, \mathbb{R}^d \to \mathbb{R}\}$, such that*

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

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 $\boldsymbol{g} \in \mathcal{NN}_\varrho\{207d+144,\, 10;\, \mathbb{R}^{5d+5} \to \mathbb{R}^{5d+5}\}$ 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*

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

This corollary is not immediately apparent and necessitates a derivation: $\boldsymbol{g}_\varepsilon \approx \boldsymbol{g}$ implies $\boldsymbol{g}_\varepsilon^{\circ(3r+1)} \approx \boldsymbol{g}^{\circ(3r+1)}$ for small $\varepsilon > 0$, where $\boldsymbol{g}$ and $\boldsymbol{g}_\varepsilon$ represent `ReLU` and ϱ-activated networks, respectively. The proof relies on mathematical induction and the following equation

$$\big\|\boldsymbol{g}_\varepsilon^{\circ(m+1)} - \boldsymbol{g}^{\circ(m+1)}\big\|_{\sup(\mathcal{K})} \le \big\|\boldsymbol{g}_\varepsilon \circ \boldsymbol{g}_\varepsilon^{\circ m} - \boldsymbol{g} \circ \boldsymbol{g}_\varepsilon^{\circ m}\big\|_{\sup(\mathcal{K})} + \big\|\boldsymbol{g} \circ \boldsymbol{g}_\varepsilon^{\circ m} - \boldsymbol{g} \circ \boldsymbol{g}^{\circ m}\big\|_{\sup(\mathcal{K})}$$

for any compact set $\mathcal{K}$ and $m \in \mathbb{N}$, where the first term of the above equation is constrained by $\boldsymbol{g}_\varepsilon \approx \boldsymbol{g}$, and the second term is controlled by the induction hypothesis and the uniform continuity of $\boldsymbol{g}$ on a compact set. 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.

The remainder of this paper is organized as follows. In Section 2, we explore some additional related topics. We present four supplementary theorems, Theorems 6, 7, 8, and 9 in Section 2.1 to complement Theorem 1, and we summarize main results of this paper in Table 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, 7, 8, and 9. In Section 3.1, we introduce the notations used throughout this paper. In Section 3.2, we present several propositions, namely Propositions 10, 11, 12, and 13, outlining the underlying ideas for proving Theorems 1, 6, 7, 8, and 9. 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, 7, 8, and 9 in Section 3.4. Finally, we prove Propositions 10, 11, 12, and 13 in Sections 4, 5, 6, and 7, respectively.

2. Further Discussions

In this section, we explore some additional related topics. We first present four supplementary theorems, namely Theorems 6, 7, 8, and 9, 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 $k = 0, 1$. 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*

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

Moving forward, let us delve into the concept of optimality, particularly concerning the potential for further reducing the (width, depth) scaling factors in Theorem 1. Despite our diligent efforts, we have yet to establish lower bounds that correspond to the (width, depth) scaling factors we have identified in this study. Consequently, it remains an open question whether the (width, depth) scaling factors we have found are the best possible in the general context. Nonetheless, by targeting specific categories of activation functions, we have succeeded in deriving improved scaling factors.

Next, we introduce two specific scenarios aimed at diminishing the (width, depth) scaling factors in Theorem 1. The first scenario is directed towards the situation where ϱ belongs to the set $\mathscr{A}_2$. In this particular case, we demonstrate that the (width, depth) scaling factors can be reduced to $(2, 1)$, as exemplified in Theorem 8 below.

Theorem 8. *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 $\varepsilon > 0$, $A > 0$, and $\varrho \in \mathscr{A}_2$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\{2N, L; \mathbb{R}^d \to \mathbb{R}^n\}$ such that*

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

The second scenario revolves around a specific subset of $\mathscr{A}_2$, denoted as $\widetilde{\mathscr{A}_2}$, and it is defined as

$$\widetilde{\mathscr{A}_2} \coloneqq \Big\{ \varrho : \mathbb{R} \to \mathbb{R} \;\Big|\; \forall\, x \in \mathbb{R},\ \varrho(x) \coloneqq (x + b_0) \cdot h(x) + b_1, \quad b_0, b_1 \in \mathbb{R}, \quad h \in \widetilde{\mathscr{S}} \Big\}.$$

where $\widetilde{\mathscr{S}}$ represents a refined subset of $\mathscr{S}$ and is defined as

$$\widetilde{\mathscr{S}} \coloneqq \Big\{ h : \mathbb{R} \to \mathbb{R} \;\Big|\; h \in \mathscr{S}, \quad \Big(\lim_{x \to -\infty} h(x)\Big) \cdot \Big(\lim_{x \to \infty} h(x)\Big) = 0 \Big\}.$$

The only difference between $\mathscr{A}_2$ and $\widetilde{\mathscr{A}_2}$ lies in the limits defined for h therein. In $\mathscr{A}_2$, it only requires the existence and distinctness of $\lim_{x\to-\infty} h(x)$ and $\lim_{x\to\infty} h(x)$. In contrast, $\widetilde{\mathscr{A}_2}$ introduces an additional requirement where either $\lim_{x\to-\infty} h(x)$ or $\lim_{x\to\infty} h(x)$ must equal 0. As illustrated in the theorem below, the (width, depth) scaling factors in this particular instance can be reduced to (1,1), which represents the optimal achievable outcome.

Theorem 9. *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 $\varepsilon > 0$, $A > 0$, and $\varrho \in \widetilde{\mathscr{A}_2}$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\{N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\}$ such that*

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

The proofs of Theorems 6, 7, 8, and 9 are available in Section 3. To facilitate the reading, we provide a summary and comparison of our primary findings in Table 1.

Table 1: Summary of main results.

	conditions on ϱ	(width, depth) scaling factors
Theorem 1	$\varrho \in \mathscr{A} = \big(\mathscr{A}_{1,0} \cup \mathscr{A}_{1,1}\big) \cup \mathscr{A}_2 \cup \mathscr{A}_3$, e.g., `Sigmoid` and `Tanh`	$(3, 2)$
Theorem 7	$\varrho \in \mathscr{A}_{1,k}$ for any $k \in \mathbb{N}$, e.g., $\texttt{ReLU}^2 \in \mathscr{A}_{1,1}$ and $\texttt{ReLU}^3 \in \mathscr{A}_{1,2}$	$(k+2, 1)$
Theorem 8	$\varrho \in \mathscr{A}_2$, e.g., $\varrho(x) \coloneqq x \cdot \texttt{Softsign}(x)$ and $\varrho(x) \coloneqq x \cdot \texttt{Arctan}(x)$	$(2, 1)$
Theorem 9	$\varrho \in \widetilde{\mathscr{A}_2} \subseteq \mathscr{A}_2$, e.g., `ELU`, `CELU`, `SELU`, `Softplus`, `GELU`, `SiLU`, `Swish`, and `Mish`	$(1, 1)$

It is important to highlight that Theorem 9 suggests that the activation function within $\widetilde{\mathscr{A}_2}$ is, at the very least, not inferior to `ReLU` in terms of approximation. Refer to Table 2 for the comparison of approximation errors when using a single active neuron activated by $\varrho \in \widetilde{\mathscr{A}_2}$ to approximate `ReLU`. Definitions and visual depictions of the activation functions referenced in Table 2 are provided in Section 2.3.

Let us briefly discuss how to estimate the approximation errors in Table 2. According to the definition of $\varrho \in \widetilde{\mathscr{A}_2}$, it can be represented as $\varrho(x) = x \cdot h(x)$ for any $x \in \mathbb{R}$, where h is an S-shaped function with either $\lim_{x\to-\infty} h(x)$ or $\lim_{x\to\infty} h(x)$ equal to 0. Without loss of generality (through scaling), it can be assumed that $\lim_{x\to-\infty} h(x) = 0$ and $\lim_{x\to\infty} h(x) = 1$. Then for any $x \in \mathbb{R}$ and $K > 0$, we have

$$\texttt{ReLU}(x) - \frac{\varrho(Kx)}{K} = x \cdot \mathbb{1}_{\{x>0\}} - \frac{Kx \cdot h(Kx)}{K} = \frac{Kx \cdot \big(\mathbb{1}_{\{Kx>0\}} - h(Kx)\big)}{K} \in \Big[\frac{m}{K}, \frac{M}{K}\Big],$$

where $m, M \in \mathbb{R} \cup \{-\infty, \infty\}$ are given by

$$m = \inf\big\{y \cdot \big(\mathbb{1}_{\{y>0\}} - h(y)\big) : y \in \mathbb{R}\big\} \quad \text{and} \quad M = \sup\big\{y \cdot \big(\mathbb{1}_{\{y>0\}} - h(y)\big) : y \in \mathbb{R}\big\}.$$

Table 2: Comparison of approximation errors when using a single active neuron activated by $\varrho \in \widetilde{\mathscr{A}_2}$ to approximate the `ReLU` activation function.

ϱ	approximation error (for any $x \in \mathbb{R}$ and $K > 0$)	constant estimate
`ELU` ($\alpha > 0$) or `CELU` ($\alpha > 0$)	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_1 \cdot \alpha/K$	$C_1 = 1$
`Softplus`	$0 \leq \texttt{ReLU}(x) - \big(\varrho(Kx) - \ln 2\big)/K \leq C_2/K$	$C_2 = \ln 2 \approx 0.693$
`GELU` ($\mu = 0$, $\sigma > 0$)	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_3 \cdot \sigma/K$	$C_3 \approx 0.170$
`SiLU`	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_4/K$	$C_4 \approx 0.278$
`Swish` ($\beta > 0$)	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_5/(\beta \cdot K)$	$C_5 = C_4 \approx 0.278$
`Mish`	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_6/K$	$C_6 \approx 0.309$
$\varrho(x) \coloneqq x \cdot \texttt{dSiLU}(x)$	$-\widetilde{C}_7/K \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_7/K$	$\widetilde{C}_7 \approx 0.265,\ C_7 \approx 0.131$
$\varrho(x) \coloneqq x \cdot \big(\texttt{Softsign}(x)/2 + 1/2\big)$	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_8/K$	$C_8 = 0.5$
$\varrho(x) \coloneqq x \cdot \big(\texttt{Arctan}(x)/\pi + 1/2\big)$	$0 \leq \texttt{ReLU}(x) - \varrho(Kx)/K \leq C_9/K$	$C_9 = 1/\pi \approx 0.318$

To ensure $-\infty < m \leq M < \infty$, we just need the set $\big\{y \cdot \big(\mathbb{1}_{\{y>0\}} - h(y)\big) : y \in \mathbb{R}\big\}$ to be bounded. We can then estimate the values of m and M by examining the characteristics of the derivatives, including higher-order ones. Refer to Figure 1 for visual representations demonstrating that the activation functions in $\widetilde{\mathscr{A}_2}$ require only a single active neuron to approximate the `ReLU` activation function.

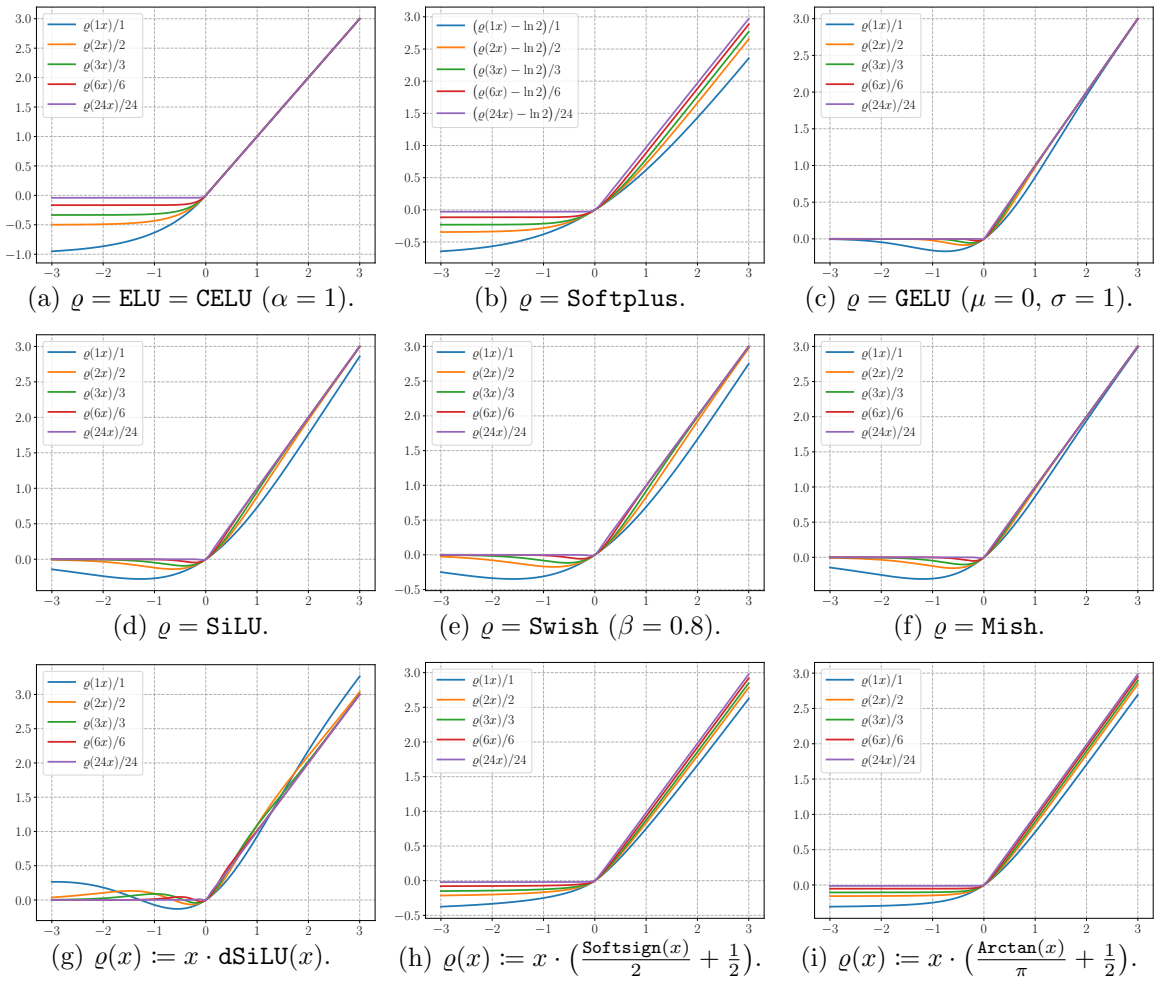
We would like to emphasize that, as we will demonstrate later in Proposition 13, any activation function $\varrho \in \widetilde{\mathscr{A}_2}$ can effectively employ just one active neuron to provide a arbitrarily accurate approximation of `ReLU` within a *bounded* subset. It is worth noting that, by introducing a mild condition, this approximation extends to the entire real number line $\mathbb{R}$, rather than being limited to a bounded subset. To elucidate this, when considering any activation function $\varrho \in \widetilde{\mathscr{A}_2}$, we can identify two affine linear maps $\mathcal{L}_1, \mathcal{L}_2 : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{L}_2 \circ \varrho \circ \mathcal{L}_1(y) = y \cdot \widetilde{h}(y)$ for any $y \in \mathbb{R}$, where $\widetilde{h} : \mathbb{R} \to \mathbb{R}$ is an S-shaped function with

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

Therefore, by ensuring that the set $\big\{y \cdot \big(\mathbb{1}_{\{y>0\}} - \widetilde{h}(y)\big) : y \in \mathbb{R}\big\}$ remains bounded, we provide adequate conditions for ϱ to employ a single active neuron to accurately approximate the `ReLU` activation function on the entire real number line $\mathbb{R}$.

Finally, let us briefly identify the types of functions that do not belong to the previously mentioned activation function sets $\mathscr{A}_{1,k}$ for $k \in \mathbb{N}$, $\mathscr{A}_2$, and $\mathscr{A}_3$. In essence, we aim to characterize the activation function ϱ for which a fixed-size ϱ-activated network cannot achieve an arbitrarily accurate approximation of the `ReLU` activation function by solely adjusting its parameters. Notably, polynomials serve as evident examples in this context. In fact, all rational functions also fall into this category. Below, a concise outline of the proof using the method of contradiction will be provided. Let us assume that ϱ is a rational function, and a fixed-size ϱ-activated network has the capability to approximate `ReLU` arbitrarily well. It is important to note that a fixed-size ϱ-activated network can be represented as a rational function of a certain degree. Consequently, a rational function of a certain degree can approximate `ReLU` arbitrarily well, and by extension, the absolute value function, which is the sum of $\texttt{ReLU}(x)$ and $\texttt{ReLU}(-x)$. However, this scenario leads to a contradiction since the

approximation error has a lower bound that depends on the degree of the rational function when using it to approximate the absolute value function (e.g., see Gosea and Antoulas, 2020).



(a) $\varrho = \texttt{ELU} = \texttt{CELU}$ ($\alpha = 1$). (b) $\varrho = \texttt{Softplus}$. (c) $\varrho = \texttt{GELU}$ ($\mu = 0$, $\sigma = 1$).

(d) $\varrho = \texttt{SiLU}$. (e) $\varrho = \texttt{Swish}$ ($\beta = 0.8$). (f) $\varrho = \texttt{Mish}$.

(g) $\varrho(x) \coloneqq x \cdot \texttt{dSiLU}(x)$. (h) $\varrho(x) \coloneqq x \cdot \left(\frac{\texttt{Softsign}(x)}{2} + \frac{1}{2}\right)$. (i) $\varrho(x) \coloneqq x \cdot \left(\frac{\texttt{Arctan}(x)}{\pi} + \frac{1}{2}\right)$.

Figure 1: Illustrations of how a single active neuron activated by $\varrho \in \widetilde{\mathscr{A}_2}$ is adequate for approximating the `ReLU` activation function.

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 (Bao et al., 2023; Barron, 1993; Bölcskei et al., 2019; Chen et al., 2019; Chui et al., 2018; Cybenko, 1989; Gribonval et al., 2022; Gühring et al., 2020; Hornik et al., 1989; Li et al., 2023; Lu et al., 2021; Montanelli and Yang, 2020; Nakada and Imaizumi, 2020; Shen et al., 2019, 2020, 2022a,b; Suzuki, 2019; Yarotsky, 2017, 2018; Zhang, 2020; Zhou, 2020). 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 neural 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 (Lu et al., 2021; Shen et al., 2019, 2020; Shen et al., 2022; Yarotsky, 2017, 2018; Zhang, 2020; Zhang et al., 2023a). 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`, `CELU`, `SELU`, `Softplus`, `GELU`, `SiLU`, `Swish`, `Mish`, `Sigmoid`, `Tanh`, `Arctan`, `Softsign`, `dSiLU`, and `SRS`. The definitions of these 17 activation functions are presented below. The first 6 activation functions are given by

$$\texttt{ReLU}(x) = \max\{0, x\}, \qquad \texttt{LeakyReLU}(x) = \begin{cases} x & \text{if } x \geq 0, \\ \alpha x & \text{if } x < 0 \end{cases} \quad \text{with } \alpha \in \mathbb{R},$$

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

$$\texttt{CELU}(x) = \begin{cases} x & \text{if } x \geq 0, \\ \alpha(e^{x/\alpha} - 1) & \text{if } x < 0 \end{cases} \quad \text{with } \alpha \in (0, \infty),$$

and

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

where e is the base of the natural logarithm. For the last 6 activation functions, `Arctan` is the inverse tangent function and the other 5 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 5 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 2 for visual representations of all these activation functions.

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, 7, 8, and 9. 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, 7, 8, and 9. 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, 7, 8, and 9 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 $\mathbb{N}^+ = \mathbb{N} \backslash \{0\} = \{1, 2, 3, \cdots\}$.
- 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., $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

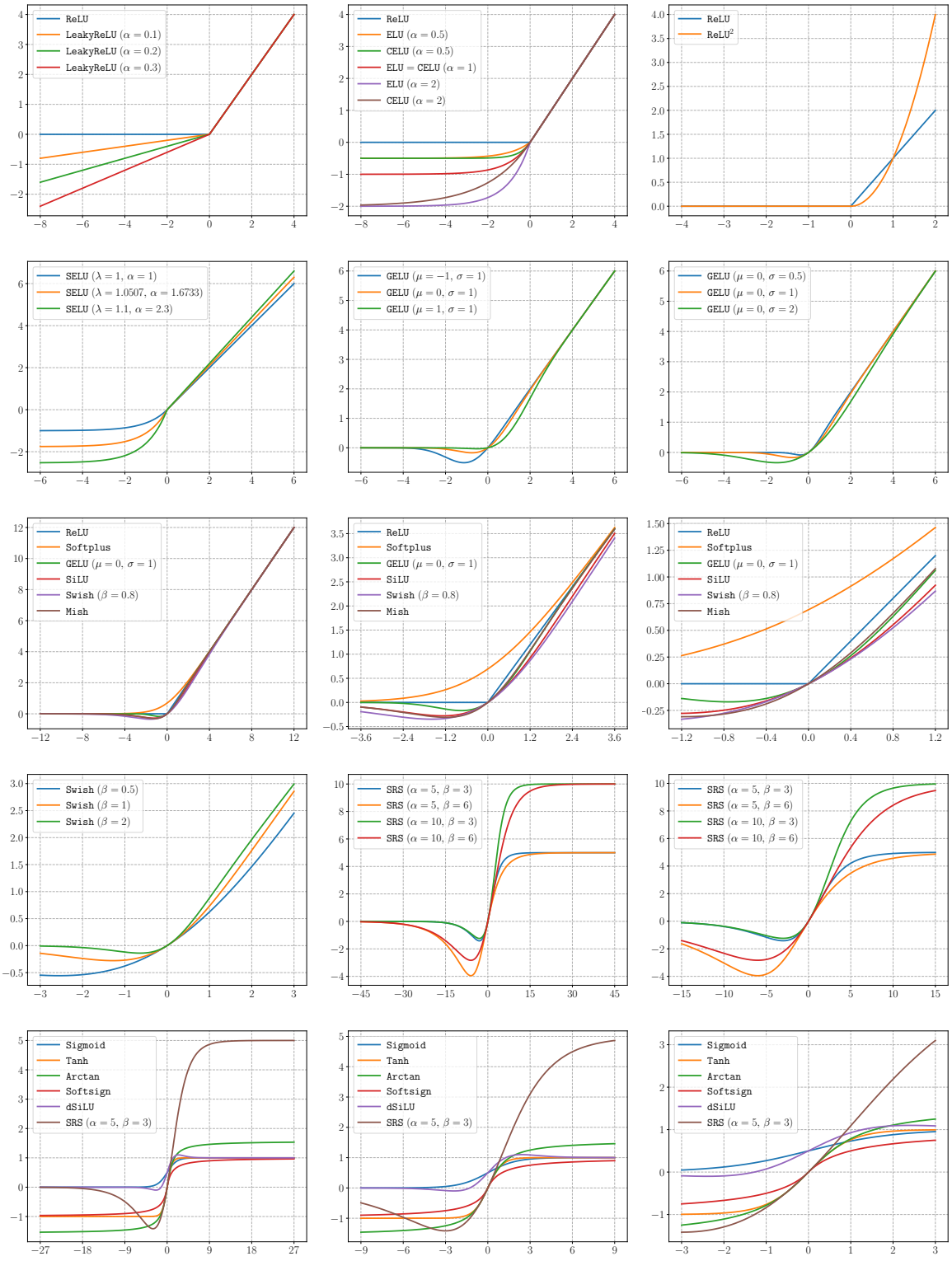


Figure 2: Illustrations of `ReLU`, `LeakyReLU`, `ReLU`^{2}, `ELU`, `CELU`, `SELU`, `Softplus`, `GELU`, `SiLU`, `Swish`, `Mish`, `Sigmoid`, `Tanh`, `Arctan`, `Softsign`, `dSiLU`, and `SRS`.

- 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 $\boldsymbol{A}^\intercal$ 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

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

- 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 $N_0 = d$, $N_1, N_2, \cdots, N_L \in \mathbb{N}^+$, $N_{L+1} = n$, and $\boldsymbol{\mathcal{L}}_i$ is an affine linear map given by $\boldsymbol{\mathcal{L}}_i : \boldsymbol{x} \mapsto \boldsymbol{W}_i\boldsymbol{x} + \boldsymbol{b}_i$ with $\boldsymbol{W}_i \in \mathbb{R}^{N_{i+1}\times N_i}$ and $\boldsymbol{b}_i \in \mathbb{R}^{N_{i+1}}$ being the weight matrix and the bias vector, respectively, for $i = 0, 1, \cdots, L$. Here

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

  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 3 for an illustration.

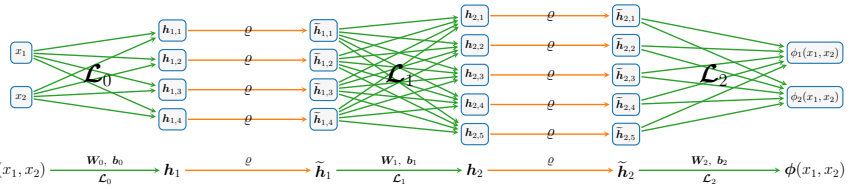


Figure 3: An example of a ϱ-activated network of width 5 and depth 2. The network realizes a vector-valued function $\boldsymbol{\phi} = (\phi_1, \phi_2)$. Here, $\boldsymbol{h}_{i,j}$ (or $\widetilde{\boldsymbol{h}}_{i,j}$) represents the j-th entry of $\boldsymbol{h}_i$ (or $\widetilde{\boldsymbol{h}}_i$) for $(i,j) \in \big\{(1,j) : j = 1,2,3,4\big\} \cup \big\{(2,j) : j = 1,2,3,4,5\big\}$.

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, 7, 8, and 9. These five 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(s) at the cost of a sufficiently small error.

Proposition 10. *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 $\widetilde{\varrho}_\eta \in \mathcal{NN}_\varrho\big\{\widetilde{N},\, \widetilde{L};\, \mathbb{R} \to \mathbb{R}\big\}$ 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 $\boldsymbol{\phi}_{\widetilde{\varrho}} \in \mathcal{NN}_{\widetilde{\varrho}}\big\{N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$, for any $\varepsilon > 0$ and $A > 0$, there exists $\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\big\{\widetilde{N} \cdot N,\ \ \widetilde{L} \cdot L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$ such that

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

The proof of Proposition 10 can be found in Section 4. The utilization of Proposition 10 simplifies our task of proving Theorems 1, 6, 7, 8, and 9. 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 11. *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 12. *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*

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

Proposition 13. *Given any $M > 0$, for each $\varepsilon \in (0,1)$, there exists*

$$\phi_\varepsilon \in \begin{cases} \mathcal{NN}_\varrho\{1,\, 1;\, \mathbb{R} \to \mathbb{R}\} & \text{if } \varrho \in \widetilde{\mathscr{A}_2}, \\ \mathcal{NN}_\varrho\{2,\, 1;\, \mathbb{R} \to \mathbb{R}\} & \text{if } \varrho \in \mathscr{A}_2, \\ \mathcal{NN}_\varrho\{3,\, 2;\, \mathbb{R} \to \mathbb{R}\} & \text{if } \varrho \in \mathscr{A}_3 \end{cases}$$

such that

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

Propositions 11, 12, and 13 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 11 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 14 later. With the uniform continuity of $f^{(n)}$ on a closed interval, Proposition 11 follows straightforwardly. See more details in Section 5.

The proof of Proposition 12 can be divided into two main steps. The first step involves demonstrating that

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

where τ can be used to generate $\mathtt{ReLU}$ 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 11 to uniformly approximate $\varrho^{(k)}$ using a ϱ-activated network. By combining these two steps, we can construct a ϱ-activated network that effectively approximates $\mathtt{ReLU}$. For further details, refer to Section 6.

The core of proving Proposition 13 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 Based on Propositions

The proof of Theorem 1 can be easily demonstrated by using Propositions 10, 12, and 13.

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

We first consider the case $\varrho \in \mathscr{A}_{1,0} \cup \mathscr{A}_{1,1}$, i.e., $\varrho \in \mathscr{A}_{1,k}$ for $k = 0, 1$. By Proposition 12, for any $M>0$, there exists $\widetilde{\varrho}_\eta \in \mathcal{NN}_\varrho\{k+2,\, 1;\, \mathbb{R}\to\mathbb{R}\} \subseteq \mathcal{NN}_\varrho\{3,\, 1;\, \mathbb{R}\to\mathbb{R}\}$ 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 10 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

$$\boldsymbol{\phi}_\varrho \in \mathcal{NN}_\varrho\big\{3N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\} \subseteq \mathcal{NN}_\varrho\big\{3N,\, 2L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$$

such that

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

Next, we consider the case $\varrho \in \mathscr{A}_2 \cup \mathscr{A}_3$. By Proposition 13, for any $M > 0$, there exists $\widetilde{\varrho}_\eta \in \mathcal{NN}_{\varrho}\{3,\, 2;\, \mathbb{R} \to \mathbb{R}\}$ 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 10 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

$$\boldsymbol{\phi}_{\varrho} \in \mathcal{NN}_{\varrho}\big\{3N,\, 2L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$$

such that

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

Thus, we finish the proof of Theorem 1. ∎

3.4 Proofs of Theorems in Section 2.1 Based on Propositions

The proofs of Theorems 6, 7, 8, and 9 can be straightforwardly demonstrated by utilizing Propositions 10, 11, 12, and 13.

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 11, 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 10 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

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

So we finish the proof of Theorem 6. ∎

Proof of Theorem 7. By Proposition 12, for any $M > 0$, $k \in \mathbb{N}$, and $\varrho \in \mathscr{A}_{1,k}$, there exists $\widetilde{\varrho}_\eta \in \mathcal{NN}_{\varrho}\{k+2,\, 1;\, \mathbb{R} \to \mathbb{R}\}$ 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 10 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. ■

Proof of Theorem 8. By Proposition 13, for any $M > 0$ and $\varrho \in \mathscr{A}_2$, there exists $\widetilde{\varrho}_\eta \in \mathcal{NN}_{\varrho}\{2,\, 1;\, \mathbb{R} \to \mathbb{R}\}$ 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 10 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 $\boldsymbol{\phi}_{\varrho} \in \mathcal{NN}_{\varrho}\big\{2N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$ such that

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

So we finish the proof of Theorem 8. ■

Proof of Theorem 9. By Proposition 13, for any $M > 0$ and $\varrho \in \widetilde{\mathscr{A}_2}$, there exists $\widetilde{\varrho}_\eta \in \mathcal{NN}_{\varrho}\{1,\, 1;\, \mathbb{R} \to \mathbb{R}\}$ 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 10 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 $\boldsymbol{\phi}_{\varrho} \in \mathcal{NN}_{\varrho}\big\{N,\, L;\, \mathbb{R}^d \to \mathbb{R}^n\big\}$ such that

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

So we finish the proof of Theorem 9. ■

4. Proof of Proposition 10

We will prove Proposition 10 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 10. 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 $\widetilde{\varrho}$-activated network with $\widehat{L}$ hidden layers, where $L \geq \widehat{L} \in \mathbb{N}^+$. 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 $\max\{N_1, N_2, \cdots, N_L\} \leq N$, $N_{L+1} = n$, $\boldsymbol{W}_\ell \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ 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

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

Moreover, 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 mappings 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

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

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

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

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

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

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

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

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

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

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

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

So we finish the proof of Proposition 10. ■

5. Proof of Proposition 11

In this section, our goal is to prove Proposition 11. 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 11

Lemma 14. *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

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

By assuming the validity of the claim, we have

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

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

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

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

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

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

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 14. ■

5.2 Proof of Proposition 11 Based on Lemma 14

Equipped with Lemma 14, we are prepared to demonstrate the proof of Proposition 11.

Proof of Proposition 11. 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 14.

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 14, 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 $\varepsilon > 0$ and $x \in [a, b]$ are 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 11. ■

6. Proof of Proposition 12

The objective of this section is to provide the proof of Proposition 12. 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 12

Lemma 15. *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. By Taylor's theorem with Peano's form of remainder, there exists $h:\mathbb{R}\to\mathbb{R}$ such that $\lim_{\eta\to 0} h(x_0+\eta)=0$ and

$$f(z)=f(x_0)+f'(x_0)(z-x_0)+h(z)(z-x_0) \quad \text{for any } z\in\mathbb{R}.$$

By substituting z with $x_0+\varepsilon x$ in the above equation, for any $x\in[-M,M]$ and $\varepsilon>0$, we obtain

$$\frac{f(x_0+\varepsilon x)-f(x_0)}{\varepsilon f'(x_0)}=\frac{f(x_0)+f'(x_0)(\varepsilon x)+h(x_0+\varepsilon x)(\varepsilon x)-f(x_0)}{\varepsilon f'(x_0)}=x+\frac{h(x_0+\varepsilon x)x}{f'(x_0)}.$$

It follows from $\lim_{\eta\to 0} h(x_0+\eta)=0$ that

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

from which we deduce

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

So we finish the proof of Lemma 15. ■

6.2 Proof of Proposition 12 Based on Lemma 15

With Lemma 15 in hand, we are ready to present the proof of Proposition 12.

Proof of Proposition 12. 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

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

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

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

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

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

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

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

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

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

from which we deduce

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

To construct a ϱ-activated network to approximate $\mathtt{ReLU}$ well, we only need to construct ϱ-activated networks to effectively approximate $\psi(x)$ and x for any $x \in [-M, M]$. When $k \ge 1$ and $\varrho'(x_1) \ne 0$, we have the option to employ $\frac{\varrho(x_1+\eta x)-\varrho(x_1)}{\eta\varrho'(x_1)}$ for a sufficiently accurate approximation of x when η is small. However, in the scenario where $k = 0$, this approach is not applicable. As a result, we will split the remainder of the proof into two cases: one where $k = 0$ and the other where $k \ge 1$.

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

$$\begin{aligned}\phi_\varepsilon(x) &\coloneqq \tfrac{1}{L_2-L_1}\psi_\varepsilon(x) - \tfrac{1}{L_2-L_1}\Big(\psi_\varepsilon(x-M)+L_1M\Big)\\ &= \tfrac{1}{L_2-L_1}\tfrac{\varrho(x_0+\varepsilon x)-\varrho(x_0)}{\varepsilon} - \tfrac{1}{L_2-L_1}\Big(\tfrac{\varrho(x_0+\varepsilon(x-M))-\varrho(x_0)}{\varepsilon}+L_1M\Big)\end{aligned}$$

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 $\varepsilon\in\big(0,\frac{\delta_\varepsilon}{2M}\big)$ and any $x\in[-M,M]$, we have

$$\begin{aligned}|\phi_\varepsilon(x)-\texttt{ReLU}(x)| &= \Big|\underbrace{\tfrac{1}{L_2-L_1}\psi_\varepsilon(x)-\tfrac{1}{L_2-L_1}\Big(\psi_\varepsilon(x-M)+L_1M\Big)}_{\phi_\varepsilon}-\Big(\underbrace{\tfrac{1}{L_2-L_1}\psi(x)-\tfrac{L_1}{L_2-L_1}x}_{\texttt{ReLU}}\Big)\Big|\\ &\le \big|\tfrac{1}{L_2-L_1}\big|\cdot\big|\psi_\varepsilon(x)-\psi(x)\big|+\big|\tfrac{1}{L_2-L_1}\big|\cdot\Big|\Big(\psi_\varepsilon(x-M)+L_1M\Big)-L_1x\Big|\\ &< K\cdot\tfrac{\varepsilon}{2K}+K\cdot\tfrac{\varepsilon}{2K}=\varepsilon,\end{aligned}$$

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

$$\phi_\varepsilon(x)\rightrightarrows\texttt{ReLU}(x)\quad\text{as }\varepsilon\to0^+\quad\text{for any }x\in[-M,M].$$

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

Case 2: $k\ge1$.

Next, let us consider the case of $k\ge1$. 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 $\varepsilon\in\big(0,\frac{\delta_\varepsilon}{2M}\big)$ and any $x\in[-M,M]\subseteq[-2M,2M]$, we have

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

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 $\varepsilon\in\big(0,\frac{\delta_\varepsilon}{2M}\big)$. 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 $\varepsilon\in\big(0,\frac{\delta_\varepsilon}{2M}\big)$.

Recall that $\varrho\in C^k\big((a_0,b_0)\big)\backslash C^{k+1}\big((a_0,b_0)\big)$ with $k\ge1$. Then there exists $x_1\in(a_0,b_0)$ such that $\varrho'(x_1)\neq0$. 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 15,

$$g_\eta(x)=\frac{\varrho(x_1+\eta x)-\varrho(x_1)}{\eta\varrho'(x_1)}\rightrightarrows x\quad\text{as }\eta\to0^+\quad\text{for any }x\in[-M,M].$$

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

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

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 11,

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

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

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

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

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

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

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

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

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

So we finish the proof of Proposition 12. ■

7. Proof of Proposition 13

We will prove Proposition 13 in this section. To this end, we first establish two lemmas in Section 7.1, which play important roles in proving Proposition 13. Next, we give the detailed proof of Proposition 13 in Section 7.2 based on these two lemmas.

7.1 Lemmas for Proving Proposition 13

Lemma 16. *Given any $A > 0$ and a function $\varrho : \mathbb{R} \to \mathbb{R}$, suppose there exists $x_0 \in \mathbb{R}$ satisfying $\varrho''(x_0) \neq 0$. Then there exists*

$$\phi_\varepsilon \in \mathcal{NN}_\varrho\{3,\, 1;\, \mathbb{R}^2 \to \mathbb{R}\} \quad \text{for each } \varepsilon \in (0,1)$$

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 Taylor's theorem with Peano's form of remainder, there exists $h : \mathbb{R} \to \mathbb{R}$ such that $\lim_{\eta \to 0} h(x_0 + \eta) = 0$ and

$$\varrho(z) = \varrho(x_0) + \varrho'(x_0)(z - x_0) + \tfrac{\varrho''(x_0)}{2}(z - x_0)^2 + h(z)(z - x_0)^2 \quad \text{for any } z \in \mathbb{R}. \tag{6}$$

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

$$\phi_\varepsilon(x,y) \coloneqq \frac{\varrho(x_0 + \varepsilon x + \varepsilon y) - \varrho(x_0 + \varepsilon y) - \varrho(x_0 + \varepsilon x) + \varrho(x_0)}{\varepsilon^2 \varrho''(x_0)} \quad \text{for any } x, y \in \mathbb{R}.$$

Clearly, $\phi_\varepsilon \in \mathcal{NN}_\varrho\{3,\, 1;\, \mathbb{R}^2 \to \mathbb{R}\}$. Moreover, for any $x, y \in [-A, A]$, applying Equation (6) with z taken as $x_0 + \varepsilon x + \varepsilon y$, $x_0 + \varepsilon x$, and $x_0 + \varepsilon y$ therein, we obtain

$$\begin{aligned}
&\varepsilon^2 \varrho''(x_0)\phi_\varepsilon(x,y) = \varrho(x_0 + \varepsilon x + \varepsilon y) - \varrho(x_0 + \varepsilon x) - \varrho(x_0 + \varepsilon y) + \varrho(x_0) \\
&= \varrho(x_0) + \varrho'(x_0)(\varepsilon x + \varepsilon y) + \tfrac{\varrho''(x_0)}{2}(\varepsilon x + \varepsilon y)^2 + h(x_0 + \varepsilon x + \varepsilon y)(\varepsilon x + \varepsilon y)^2 \\
&\quad - \Big(\varrho(x_0) + \varrho'(x_0)(\varepsilon x) + \tfrac{\varrho''(x_0)}{2}(\varepsilon x)^2 + h(x_0 + \varepsilon x)(\varepsilon x)^2\Big) \\
&\quad - \Big(\varrho(x_0) + \varrho'(x_0)(\varepsilon y) + \tfrac{\varrho''(x_0)}{2}(\varepsilon y)^2 + h(x_0 + \varepsilon y)(\varepsilon y)^2\Big) + \varrho(x_0) \\
&= \varrho''(x_0)\varepsilon^2 xy + \Big(h(x_0 + \varepsilon x + \varepsilon y)(\varepsilon x + \varepsilon y)^2 + h(x_0 + \varepsilon x)(\varepsilon x)^2 + h(x_0 + \varepsilon y)(\varepsilon y)^2\Big),
\end{aligned}$$

from which we deduce

$$\begin{aligned}
\phi_\varepsilon(x,y) &= xy + \frac{h(x_0 + \varepsilon x + \varepsilon y)(\varepsilon x + \varepsilon y)^2 + h(x_0 + \varepsilon x)(\varepsilon x)^2 + h(x_0 + \varepsilon y)(\varepsilon y)^2}{\varepsilon^2 \varrho''(x_0)} \\
&= xy + \frac{h(x_0 + \varepsilon x + \varepsilon y)(x + y)^2 + h(x_0 + \varepsilon x)x^2 + h(x_0 + \varepsilon y)y^2}{\varrho''(x_0)}.
\end{aligned}$$

It follows from $\lim_{\eta \to 0} h(x_0 + \eta) = 0$ that

$$h(x_0 + \varepsilon x + \varepsilon y)(x + y)^2 \rightrightarrows 0, \quad h(x_0 + \varepsilon x)x^2 \rightrightarrows 0, \quad h(x_0 + \varepsilon y)y^2 \rightrightarrows 0$$

as $\varepsilon \to 0^+$ for any $x, y \in [-A, A]$. Consequently, we get

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

which means we complete the proof of Lemma 16. ∎

Lemma 17. *Given any $M > 0$ and two functions $g_1, g_{2,\delta} : \mathbb{R} \to \mathbb{R}$ for each $\delta \in (0,1)$, suppose*

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

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

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

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

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

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

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

We 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

$$\begin{aligned}
\big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x) - \texttt{ReLU}(x)\big| &= \big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x) - x \cdot \mathbb{1}_{\{x>0\}}\big| \\
&\le \big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x) - xg_1(K_\varepsilon x)\big| + \big|xg_1(K_\varepsilon x) - x \cdot \mathbb{1}_{\{x>0\}}\big| \\
&\le \big|g_1(K_\varepsilon x)\big| \cdot \big|g_{2,\delta_\varepsilon}(x) - x\big| + |x| \cdot \big|g_1(K_\varepsilon x) - \mathbb{1}_{\{x>0\}}\big| \\
&\le K_0 \cdot \varepsilon_2 + |x| \cdot \big|g_1(K_0K_1x/\varepsilon) - \mathbb{1}_{\{x>0\}}\big| \\
&= \varepsilon/3 + |x| \cdot \big|g_1(K_0K_1x/\varepsilon) - \mathbb{1}_{\{x>0\}}\big|.
\end{aligned}$$

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

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

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

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

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

$$\big|g_1(K_\varepsilon x)g_{2,\delta_\varepsilon}(x) - \texttt{ReLU}(x)\big| < \varepsilon,$$

which means we finish the proof of Lemma 17. ∎

7.2 Proof of Proposition 13 Based on Lemmas 16 and 17

Having established Lemmas 16 and 17 in Section 7.1, we are now prepared to prove Proposition 13.

Proof of Proposition 13. For any $\varepsilon \in (0,1)$, our goal is to construct

$$\phi_\varepsilon \in \begin{cases} \mathcal{NN}_\varrho\{1,\, 1;\, \mathbb{R}\to\mathbb{R}\} & \text{if } \varrho \in \widetilde{\mathscr{A}_2}, \\ \mathcal{NN}_\varrho\{2,\, 1;\, \mathbb{R}\to\mathbb{R}\} & \text{if } \varrho \in \mathscr{A}_2, \\ \mathcal{NN}_\varrho\{3,\, 2;\, \mathbb{R}\to\mathbb{R}\} & \text{if } \varrho \in \mathscr{A}_3 \end{cases}$$

to approximate `ReLU` well on $[-M, M]$. We divide the proof into three cases: $\varrho \in \widetilde{\mathscr{A}_2}$, $\varrho \in \mathscr{A}_2$, and $\varrho \in \mathscr{A}_3$.

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

Let us first consider the case of $\varrho \in \widetilde{\mathscr{A}_2}$. The fact $\varrho \in \widetilde{\mathscr{A}_2}$ implies that there exist $\widetilde{b}_0, \widetilde{b}_1 \in \mathbb{R}$ and $\widetilde{h} : \mathbb{R} \to \mathbb{R}$ such that

$$\sup_{x\in\mathbb{R}} |\widetilde{h}(x)| < \infty, \quad L_1 = \lim_{x\to-\infty} \widetilde{h}(x) \neq L_2 = \lim_{x\to\infty} \widetilde{h}(x), \quad L_1 \cdot L_2 = 0,$$

and

$$\varrho(x) = (x + \widetilde{b}_0) \cdot \widetilde{h}(x) + \widetilde{b}_1 \quad \text{for any } x \in \mathbb{R}.$$

The equality $L_1 \cdot L_2 = 0$ indicates that at least one of the values, either L_1 or L_2, must be zero. Set $w_0 = L_1 + L_2$, $b_0 = (L_1 + L_2)\widetilde{b}_0$, $b_1 = \widetilde{b}_1$, and

$$w_1 = \mathbb{1}_{\{L_1=0\}} - \mathbb{1}_{\{L_1\neq 0\}} = \begin{cases} 1 & \text{if } L_1 = 0, \\ -1 & \text{if } L_1 \neq 0 \end{cases} = \begin{cases} 1 & \text{if } L_1 = 0, \\ -1 & \text{if } L_2 = 0. \end{cases}$$

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

$$\begin{aligned} \varrho(x) = (x + \widetilde{b}_0) \cdot \widetilde{h}(x) + \widetilde{b}_1 &= \Big((L_1 + L_2)x + (L_1 + L_2)\widetilde{b}_0\Big) \cdot \frac{\widetilde{h}(w_1^2 x)}{L_1 + L_2} + \widetilde{b}_1 \\ &= (w_0 x + b_0) \cdot h(w_1 x) + b_1, \end{aligned}$$

where $h : \mathbb{R} \to \mathbb{R}$ is defined via

$$h(x) \coloneqq \frac{\widetilde{h}(w_1 x)}{L_1 + L_2} \quad \text{for any } x \in \mathbb{R}.$$

It is easy to verify that $\sup_{x\in\mathbb{R}} |h(x)| < \infty$,

$$\lim_{x\to-\infty} h(x) = \lim_{x\to-\infty} \frac{\widetilde{h}(w_1 x)}{L_1 + L_2} = \begin{cases} \lim\limits_{x\to-\infty} \frac{\widetilde{h}(x)}{L_1+L_2} = \frac{L_1}{L_1+L_2} = 0 & \text{if } L_1 = 0, \\ \lim\limits_{x\to-\infty} \frac{\widetilde{h}(-x)}{L_1+L_2} = \frac{L_2}{L_1+L_2} = 0 & \text{if } L_2 = 0, \end{cases}$$

and

$$\lim_{x\to\infty} h(x) = \lim_{x\to\infty} \frac{\widetilde{h}(w_1 x)}{L_1 + L_2} = \begin{cases} \lim\limits_{x\to\infty} \frac{\widetilde{h}(x)}{L_1+L_2} = \frac{L_2}{L_1+L_2} = 1 & \text{if } L_1 = 0, \\ \lim\limits_{x\to\infty} \frac{\widetilde{h}(-x)}{L_1+L_2} = \frac{L_1}{L_1+L_2} = 1 & \text{if } L_2 = 0. \end{cases}$$

By defining an affine linear map $\mathcal{L}(x) \coloneqq \frac{w_1}{|w_0 w_1|}x - \frac{b_0}{w_0}$ for any $x \in \mathbb{R}$, we have

$$\varrho \circ \mathcal{L}(x) = (w_0\mathcal{L}(x) + b_0) \cdot h\big(w_1\mathcal{L}(x)\big) + b_1 = \tfrac{w_0 w_1}{|w_0 w_1|}x \cdot h\big(w_1\mathcal{L}(x)\big) + b_1 \tag{9}$$

for any $x \in \mathbb{R}$. To make use of Lemma 17, we define

$$g_1(x) \coloneqq h\big(w_1\mathcal{L}(x)\big) = h\Big(\tfrac{w_1^2}{|w_0 w_1|}x - \tfrac{w_1 b_0}{w_0}\Big) \quad \text{for any } x \in \mathbb{R}$$

and

$$g_{2,\delta}(x) \coloneqq x \quad \text{for any } x \in \mathbb{R} \text{ and each } \delta \in (0,1).$$

It is worth noting that $\frac{w_1^2}{|w_0 w_1|} > 0$. Consequently, we can deduce that

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

According to Lemma 17, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0,1)$ such that

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

This means

$$\Big|h\big(w_1\mathcal{L}(K_\varepsilon x)\big) \cdot x - \mathtt{ReLU}(x)\Big| < \varepsilon \quad \text{for any } x \in [-M, M].$$

Define

$$\phi_\varepsilon(x) \coloneqq \tfrac{|w_0 w_1|}{w_0 w_1}\tfrac{1}{K_\varepsilon}\Big(\varrho \circ \mathcal{L}(K_\varepsilon x) - b_1\Big) = \tfrac{|w_0 w_1|}{w_0 w_1}\tfrac{1}{K_\varepsilon}\varrho\Big(\tfrac{w_1 K_\varepsilon}{|w_0 w_1|}x - \tfrac{b_0}{w_0}\Big) - \tfrac{|w_0 w_1|}{w_0 w_1}\tfrac{1}{K_\varepsilon}b_1$$

for any $x \in \mathbb{R}$. Clearly, $\phi_\varepsilon \in \mathcal{NN}_\varrho\{1,\, 1;\, \mathbb{R} \to \mathbb{R}\}$. Furthermore, based on Equation (9), for any $x \in [-M, M]$, we have

$$\begin{aligned}
\big|\phi_\varepsilon(x) - \mathtt{ReLU}(x)\big| &= \Big|\tfrac{|w_0 w_1|}{w_0 w_1}\tfrac{1}{K_\varepsilon}\Big(\varrho \circ \mathcal{L}(K_\varepsilon x) - b_1\Big) - \mathtt{ReLU}(x)\Big| \\
&= \Big|\tfrac{|w_0 w_1|}{w_0 w_1}\tfrac{1}{K_\varepsilon}\overbrace{\tfrac{w_0 w_1}{|w_0 w_1|}(K_\varepsilon x) \cdot h\big(w_1\mathcal{L}(K_\varepsilon x)\big)}^{=\varrho\circ\mathcal{L}(K_\varepsilon x) - b_1 \text{ by (9)}} - \mathtt{ReLU}(x)\Big| \\
&= \Big|x \cdot h\big(w_1\mathcal{L}(K_\varepsilon x)\big) - \mathtt{ReLU}(x)\Big| < \varepsilon.
\end{aligned}$$

Case 2: $\varrho \in \mathscr{A}_2$.

Next, let us consider the case of $\varrho \in \mathscr{A}_2$. The fact $\varrho \in \mathscr{A}_2$ implies that there exist $b_0, b_1 \in \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ such that

$$\sup_{x\in\mathbb{R}} |h(x)| < \infty, \quad L_1 = \lim_{x\to-\infty} h(x) \neq L_2 = \lim_{x\to\infty} h(x),$$

and

$$\varrho(x) = (x + b_0) \cdot h(x) + b_1 \quad \text{for any } x \in \mathbb{R}.$$

By defining an affine linear map $\mathcal{L}(x) \coloneqq x - b_0$ for any $x \in \mathbb{R}$, we have

$$\varrho \circ \mathcal{L}(x) = (\mathcal{L}(x) + b_0) \cdot h\big(\mathcal{L}(x)\big) + b_1 = x \cdot h\big(\mathcal{L}(x)\big) + b_1 \tag{10}$$

for any $x \in \mathbb{R}$. To make use of Lemma 17, we define

$$g_1(x) \coloneqq \frac{h\big(\mathcal{L}(x)\big) - L_1}{L_2 - L_1} = \frac{h(x - b_0) - L_1}{L_2 - L_1} \quad \text{for any } x \in \mathbb{R}$$

and

$$g_{2,\delta}(x) \coloneqq x \quad \text{for any } x \in \mathbb{R} \text{ and each } \delta \in (0,1).$$

Consequently, we can deduce that

$$\sup_{x\in\mathbb{R}} |g_1(x)| < \infty, \quad \lim_{x\to-\infty} g_1(x) = \tfrac{L_1-L_1}{L_2-L_1} = 0, \quad \text{and} \quad \lim_{x\to\infty} g_1(x) = \tfrac{L_2-L_1}{L_2-L_1} = 1.$$

By Lemma 17 and setting $\widetilde{M} = 2M > 0$, there exist $K_\varepsilon > 0$ and $\delta_\varepsilon \in (0,1)$ such that

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

This means

$$\left|\frac{h\big(\mathcal{L}(K_\varepsilon x)\big) - L_1}{L_2 - L_1} \cdot x - \mathtt{ReLU}(x)\right| < \varepsilon/2 \quad \text{for any } x \in [-2M, 2M]. \tag{11}$$

Define

$$\psi_\varepsilon(x) \coloneqq \tfrac{1}{L_2-L_1}\tfrac{1}{K_\varepsilon}\Big(\varrho \circ \mathcal{L}(K_\varepsilon x) - b_1\Big) = \tfrac{1}{L_2-L_1}\tfrac{1}{K_\varepsilon}\varrho(K_\varepsilon x - b_0) - \tfrac{1}{L_2-L_1}\tfrac{1}{K_\varepsilon} b_1$$

and

$$\psi(x) = \tfrac{L_1}{L_2-L_1} \cdot x + \mathtt{ReLU}(x) \quad \text{for any } x \in \mathbb{R}.$$

Clearly, $\psi_\varepsilon \in \mathcal{NN}_\varrho\{1, 1; \mathbb{R} \to \mathbb{R}\}$. Furthermore, based on Equation (10), for any $x \in [-2M, 2M]$, we have

$$\begin{aligned}
\big|\psi_\varepsilon(x) - \psi(x)\big| &= \left|\tfrac{1}{L_2-L_1}\tfrac{1}{K_\varepsilon}\Big(\varrho \circ \mathcal{L}(K_\varepsilon x) - b_1\Big) - \psi(x)\right| \\
&= \left|\tfrac{1}{L_2-L_1}\tfrac{1}{K_\varepsilon}\overbrace{(K_\varepsilon x) \cdot h\big(\mathcal{L}(K_\varepsilon x)\big)}^{=\varrho\circ\mathcal{L}(K_\varepsilon x) - b_1 \text{ by (10)}} - \psi(x)\right| \\
&= \left|\tfrac{x \cdot h(\mathcal{L}(K_\varepsilon x))}{L_2-L_1} - \big(\tfrac{L_1}{L_2-L_1} \cdot x + \mathtt{ReLU}(x)\big)\right| \\
&= \left|\tfrac{h(\mathcal{L}(K_\varepsilon x)) - L_1}{L_2-L_1} \cdot x - \mathtt{ReLU}(x)\right| < \varepsilon/2,
\end{aligned}$$

where the last equality comes from Equation (11). Then for any $x \in [-M, M]$, we have $x - M \in [-2M, 0]$ and hence $\mathtt{ReLU}(x - M) = 0$, from which we deduce

$$\begin{aligned}
\varepsilon/2 > \Big|\psi_\varepsilon(x - M) - \psi(x - M)\Big| &= \Big|\psi_\varepsilon(x - M) - \Big(\tfrac{L_1}{L_2-L_1} \cdot (x - M) + \mathtt{ReLU}(x - M)\Big)\Big| \\
= \Big|\psi_\varepsilon(x - M) - \tfrac{L_1}{L_2-L_1} \cdot (x - M)\Big| &= \Big|\psi_\varepsilon(x - M) + \tfrac{L_1 M}{L_2-L_1} - \tfrac{L_1}{L_2-L_1} \cdot x\Big|.
\end{aligned}$$

Define

$$\phi_\varepsilon(x) \coloneqq \psi_\varepsilon(x) - \Big(\psi_\varepsilon(x - M) + \tfrac{L_1 M}{L_2-L_1}\Big) \quad \text{for any } x \in \mathbb{R}.$$

It follows from $\psi_\varepsilon \in \mathcal{NN}_\varrho\{1,\, 1;\, \mathbb{R}\to\mathbb{R}\}$ that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{2,\, 1;\, \mathbb{R}\to\mathbb{R}\}$. Moreover, we have

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

Case 3: $\varrho \in \mathscr{A}_3$.

Finally, 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

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

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

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

By Lemma 15,

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

By Lemma 17, 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 16, there exists

$$\Gamma_\eta \in \mathcal{NN}_\varrho\{3,\, 1;\, \mathbb{R}^2 \to \mathbb{R}\} \quad \text{for each } \eta \in (0,1)$$

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

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

Define

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

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

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

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

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

It remains to show $\phi_\varepsilon \in \mathcal{NN}_\varrho\{3,\, 2;\, \mathbb{R} \to \mathbb{R}\}$. By defining

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

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

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

for any $x \in \mathbb{R}$. Recall that $\Gamma_{\eta_\varepsilon} \in \mathcal{NN}_\varrho\{3,\, 1;\, \mathbb{R}^2 \to \mathbb{R}\}$. Hence, we can conclude that $\phi_\varepsilon \in \mathcal{NN}_\varrho\{3,\, 2;\, \mathbb{R} \to \mathbb{R}\}$. This result completes the proof of Proposition 13. ■

Acknowledgments

Jianfeng Lu was partially supported by NSF grants CCF-1910571 and DMS-2012286. Hongkai Zhao was partially supported by NSF grants 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: 10.4208/eajam.2022-270.070123.

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. DOI: 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. DOI: 10.48550/arXiv.1809.03090.

Jonathan T. Barron. Continuously differentiable exponential linear units. *arXiv e-prints*, art. arXiv:1704.07483, April 2017. DOI: 10.48550/arXiv.1704.07483.

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.

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_files/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.

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. DOI: 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.

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: 10.1016/j.neunet.2017.12.012. 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.

Ion Victor Gosea and Athanasios C. Antoulas. Rational approximation of the absolute value function from measurements: a numerical study of recent methods. *arXiv e-prints*, art. arXiv:2005.02736, May 2020. DOI: 10.48550/arXiv.2005.02736.

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation spaces of deep neural networks. *Constructive Approximation*, 55:259–367, 2022. DOI: 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.

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: 10.1016/0893-6080(91)90009-T.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. DOI: 10.1016/0893-6080(89)90020-8.

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.

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. DOI: 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. DOI: 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. URL: https://www.semanticscholar.org/paper/Rectifier-Nonlinearities-Improve-Neural-Network-Maas/367f2c63a6f6a10b3b64b8729d601e69337ee3cc.

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: 10.1016/j.neunet.2019.12.013.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, Haifa, Israel, June 2010. Omnipress. URL: https://icml.cc/Conferences/2010/papers/432.pdf.

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: 10.1016/j.neunet.2019.07.011.

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. DOI: 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: 10.1016/j.matpur.2021.07.009.

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: 10.1016/j.acha.2021.12.005.

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. URL: https://aclanthology.org/N09-2062.

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: 10.1016/j.neunet.2017.07.002.

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: 10.1016/j.acha.2019.06.004.