# Simultaneous Neural Network Approximation for Smooth Functions

Sean Hon

*Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong*

Haizhao Yang

*Department of Mathematics, Purdue University, IN 47907, USA*



**Abstract**

We establish in this work approximation results of deep neural networks for smooth functions measured in Sobolev norms, motivated by recent development of numerical solvers for partial differential equations using deep neural networks. Our approximation results are nonasymptotic in the sense that the error bounds are explicitly characterized in terms of both the width and depth of the networks simultaneously with all involved constants explicitly determined. Namely, for $f \in C^s([0,1]^d)$, we show that deep ReLU networks of width $\mathcal{O}(N \log N)$ and of depth $\mathcal{O}(L \log L)$ can achieve a nonasymptotic approximation rate of $\mathcal{O}(N^{-2(s-1)/d} L^{-2(s-1)/d})$ with respect to the $\mathcal{W}^{1,p}([0,1]^d)$ norm for $p \in [1, \infty)$. If either the ReLU function or its square is applied as activation functions to construct deep neural networks of width $\mathcal{O}(N \log N)$ and of depth $\mathcal{O}(L \log L)$ to approximate $f \in C^s([0,1]^d)$, the approximation rate is $\mathcal{O}(N^{-2(s-n)/d} L^{-2(s-n)/d})$ with respect to the $\mathcal{W}^{n,p}([0,1]^d)$ norm for $p \in [1, \infty)$. An extension of similar approximation results is also provided for target functions in the Hölder space.

*Keywords:* Deep neural networks, Sobolev norm, ReLU$^k$ activation functions, approximation theory, Hölder spaces

## 1. Introduction

Over the past decades, deep neural networks have made remarkable impacts in various areas of science and engineering. With the aid of high-performance computing equipment and abundance of high quality data, neural network based methods outperform traditional machine learning methods in a wide range of applications, including image classification, object detection, speech recognition, to name just a few. The success of neural networks also motivates its applications in scientific computing including the recovery of governing equations for mathematical modeling and prediction [52, 12, 40, 28, 22, 33] and solving partial differential equations (PDEs) [41, 13, 21, 26, 15, 16, 25, 18, 6, 27].

Neural network based methods have evoked many open problems in mathematical theory, notwithstanding their success in practice. In a typical supervised learning algorithm, a potential high-dimensional target function $f(x)$ defined on a domain $\Omega$ is to be learned from

*Email addresses:* seanyshon@hkbu.edu.hk (Sean Hon), haizhao@purdue.edu (Haizhao Yang)

a finite set of data samples $\{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^n$. When a deep network is integrated in the learning process, one needs to identify a deep network $\phi(\mathbf{x};\theta_{\mathbf{S}})$ with $\theta_{\mathbf{S}}$ being the hyperparameter to determine $f(\mathbf{x})$ for unseen data samples $\mathbf{x}$, namely the following optimization problem arises

$$\theta_{\mathbf{S}} = \arg\min_{\theta} R_S(\theta) := \arg\min_{\theta} \tfrac{1}{n}\sum_{i=1}^{n} l(\phi(x_i;\theta), f(x_i)) \tag{1}$$

where $l$ denotes a loss function.

Now, let us inspect the overall inference error which is estimated by $R_D(\theta_S)$, where $R_D(\theta) := E_{\mathbf{x}\sim U(\Omega)}[l(\phi(\mathbf{x};\theta), f(\mathbf{x}))]$. In reality, $U(\Omega)$ is unknown and only finitely many samples from this distribution are available. Hence, the empirical loss $R_S(\theta)$ is minimized hoping to obtain $\phi(\mathbf{x};\theta_{\mathbf{S}})$, instead of minimizing the population loss $R_D(\theta)$ to obtain $\phi(\mathbf{x};\theta_{\mathbf{D}})$, where $\theta_{\mathbf{D}} = \arg\min_\theta R_D(\theta)$. In practice, a numerical optimization method to solve (1) may result in a numerical solution (denoted as $\theta_{\mathcal{N}}$) that may not be a global minimizer $\theta_{\mathbf{S}}$. Therefore, the actually learned neural network to infer $f(x)$ is $\phi(\mathbf{x};\theta_{\mathcal{N}})$ and the corresponding inference error is measured by $R_D(\theta_{\mathcal{N}})$. By the discussion just above, it is crucial to quantify $R_D(\theta_{\mathcal{N}})$ to see how good the learned neural network $\phi(\mathbf{x};\theta_{\mathcal{N}})$ is, since $R_D(\theta_{\mathcal{N}})$ is the expected inference error overall possible data samples. Note that

$$R_D(\theta_{\mathcal{N}}) \le \underbrace{R_D(\theta_{\mathbf{D}})}_{\text{approximation}} + \underbrace{[R_S(\theta_{\mathcal{N}}) - R_S(\theta_{\mathbf{S}})]}_{\text{optimization}} + \underbrace{[R_D(\theta_{\mathcal{N}}) - R_S(\theta_{\mathcal{N}})] + [R_S(\theta_{\mathbf{D}}) - R_D(\theta_{\mathbf{D}})]}_{\text{generalization}}, \tag{2}$$

where the inequality comes from the fact that $[R_S(\theta_{\mathbf{S}})) - R_S(\theta_{\mathbf{D}})] \le 0$ since $\theta_{\mathbf{S}}$ is a global minimizer of $R_S(\theta)$.

Synergies from approximation theory [51, 42, 4, 43, 37, 36, 31, 17, 39, 50, 49, 14, 47, 23, 8, 7], optimization theory [24, 35, 34, 2, 11, 53, 1, 10, 48], and generalization theory [24, 5, 49, 3, 32, 30, 29] have led to many recent advances in the mathematical investigation for deep learning, regarding the error bound (2). All of these areas, having different emphases and angles, have fostered many separate research directions.

Providing an estimate of the first error term of (2), $R_D(\theta_{\mathbf{D}})$, belongs to the regime of approximation theory and is of the main concerns in this paper. The results established in this work provide an upper bound of $R_D(\theta_{\mathbf{D}})$, estimated explicitly in terms of the size of the network, e.g. its width and depth, with a nonasymptotic approximation rate for functions in the Sobolev spaces. Our approximation results are nonasymptotic in the sense that the all involved constants are explicitly determined. We summarize our first main result in the following on the networks with rectified linear unit (ReLU) $\sigma_1$ as the activation function:

**Theorem 1.1.** *Suppose that $f \in C^s([0,1]^d)$ with $s > 1 \in \mathbb{N}^+$ satisfies $\|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} < 1$ for any $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \le s$. For any $N, L \in \mathbb{N}^+$ and $p \in [1,\infty)$, there exists a $\sigma_1$-NN $\phi$ with width $16 s^{d+1} d(N+2)\log_2(8N)$ and depth $27 s^2 (L+2)\log_2(4L)$ such that*

$$\|f - \phi\|_{\mathcal{W}^{1,p}([0,1]^d)} \le 85(s+1)^d 8^s N^{-2(s-1)/d} L^{-2(s-1)/d}.$$

For smooth functions, we show in Theorem 1.1 that deep ReLU networks of width $\mathcal{O}(N\log N)$ and of depth $\mathcal{O}(L\log L)$ can achieve an approximation rate of $\mathcal{O}(N^{-2(s-1)/d}L^{-2(s-1)/d})$ with respect to the $\mathcal{W}^{1,p}([0,1]^d)$ norm. Note that ReLU networks are in fact piecewise linear

functions, so they have at most nonzero first (weak) derivatives. The above approximation rate in the $L^p$ norm estimated in terms of $N$ and $L$ was already covered in [31]. When the $\mathcal{W}^{n,p}$ norm is considered, where $0 < n < 1$ is not an integer, the interpolation technique used in [19] can be combined together with our method here to develop new approximation rates, which is left as future work.

To achieve rates truly in terms of width and depth without the logarithm terms, we also have the following corollary by setting $\widetilde{N} = \mathcal{O}(N \log N)$ and $\widetilde{L} = \mathcal{O}(L \log L)$ and making use of

$$(N \ln N)^{-2(s-1)/d} (L \ln L)^{-2(s-1)/d} \leq \mathcal{O}(N^{-2(s-\rho)/d} L^{-2(s-\rho)/d})$$

for $\rho \in (1, s)$.

**Corollary 1.2.** *Suppose that $f \in C^s([0,1]^d)$ with $s > 1 \in \mathbb{N}^+$ satisfies $\|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} < 1$ for any $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq s$. For any $N, L \in \mathbb{N}^+$, $\rho \in (1, s)$, and $p \in [1, \infty)$, there exist $C_1(s,d)$, $C_2(s,d)$, $C_3(s,d,\rho)$, and a $\sigma_1$-NN $\phi$ with width $C_1 N$ and depth $C_1 L$ such that*

$$\|f - \phi\|_{\mathcal{W}^{1,p}([0,1]^d)} \leq C_3 N^{-2(s-\rho)/d} L^{-2(s-\rho)/d}.$$

Note that the constants $C_1, C_2$ and $C_3$ in Corollary 1.2 can be explicitly determined and we leave it to readers.

To obtain approximation results in terms of the number of parameters in neural networks, the following corollary is followed by setting $N = \mathcal{O}(1)$ and $\varepsilon = \mathcal{O}(L^{-2(s-1)/d})$ from Theorem 1.1.

**Corollary 1.3.** *Suppose that $f \in C^s([0,1]^d)$ with $s > 1 \in \mathbb{N}^+$ satisfies $\|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} < 1$ for any $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq s$. Given any $\varepsilon > 0$ and $p \in [1, \infty)$, there exists a $\sigma_1$-NN $\phi$ with $\mathcal{O}(\varepsilon^{-d/(2(s-1))} \ln \frac{1}{\varepsilon})$ parameters such that*

$$\|f - \phi\|_{\mathcal{W}^{1,p}([0,1]^d)} \leq \varepsilon.$$

Corollary 1.3 partially recovers [19, Corollary 4.1] in which the parameters are of order $\mathcal{O}(\varepsilon^{-d/(s-1)} \ln^2(\varepsilon^{-s/(s-1)}))$.

Considering a smoother neural network in which either the ReLU function $\sigma_1$ or its square $\sigma_1^2$ is applied, we provide similar results in the following Theorem 1.4. Namely, these networks of width $\mathcal{O}(N \log N)$ and of depth $\mathcal{O}(L \log L)$ can achieve the approximation rate of $\mathcal{O}(N^{-2(s-n)/d} L^{-2(s-n)/d})$ with respect to the $\mathcal{W}^{n,p}([0,1]^d)$ norm. It is worth noting that the confinement of space is now relaxed to $\mathcal{W}^{n,p}([0,1]^d)$ from $\mathcal{W}^{1,p}([0,1]^d)$ with such smoother networks.

**Theorem 1.4.** *Suppose that $f \in C^s([0,1]^d)$ with $s \in \mathbb{N}^+$ satisfies $\|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} < 1$ for any $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq s$. For any $N, L \in \mathbb{N}^+$ satisfying $(L - 2 - \log_2 N)N \geq s$, and $p \in [1, \infty)$, there exists a $\sigma_2$-NN $\phi$ with width $16 s^{d+1} d(N+2) \log_2(8N)$ and depth $10(L+2)\log_2(4L)$ such that*

$$\|f - \phi\|_{\mathcal{W}^{n,p}([0,1]^d)} \leq 3(s+1)^d 8^{s-n} N^{-2(s-n)/d} L^{-2(s-n)/d},$$

*where $n < s$ is a positive integer.*

Our work developed here focuses particularly on the Sobolev spaces which are suitable for studying partial differential equations. The above results concern the $\mathrm{ReLU}^k$ network approximation of functions with respect to Sobolev norms. Along this research direction, the work in [19] provides asymptotic upper bounds with unknown parameters with respect to $\mathcal{W}^{n,p}$ norm, where $n$ is a fraction between zero and one, based on the means of localized polynomials. Making use of the same polynomials, asymptotic approximation results measured with respect to high order $\mathcal{W}^{n,p}$ norm were provided in [20]. These results were developed for neural networks with smoother activation functions, in addition to ReLU networks. Measured in $\mathcal{W}^{1,p}$ norm, it was shown in [38] that ReLU networks can achieve essentially the same approximation rates as free-knot spline approximations. In contrast with these existing results, our approximation results are nonasymptotic with known pre-factors, established in the spirit of explicit error characterization methodology developed in [31, 44].

Recall from [9, Theorem 4.2] by DeVore et al., we have the following ostensible negative result on the continuity of the weight selection. Suppose that there is a continuous map $\Sigma$ from the unit ball of Sobolev space with smoothness $n$, i.e. $\mathcal{W}^{n,p}$, to $\mathbb{R}^W$ such that $\|f - g(\Sigma(f))\|_{L^p} \leq \varepsilon$ for all $f \in \mathcal{W}^{n,p}$, where $W$ denotes a fixed number of parameters and $g$ is a map realizing a deep neural network from a given set of parameters in $\mathbb{R}^W$ to the unit ball in $\mathcal{W}^{n,p}$, then $W \geq C\varepsilon^{-n/d}$ with some constant $C$ depending only on $n$. This in a way means any such constructive approximation of ReLU networks cannot have a continuous weight selection property if the approximation rate is better than $C\varepsilon^{-n/d}$, and hence a stable numerical implementation with such an error rate does not exist. It must, however, note that [9, Theorem 4.2] is basically a min-max criterion for evaluating continuous weight selection maps, describing the worst case. That is, the approximation result is obtained by minimizing over all continuous maps $\Sigma$ and network realizations $g$ and maximizing over all target functions. For most smooth functions practically encountered in applications, the theorem does not eliminate the possible cases in which they might still enjoy a continuous weight selection. Thus, there could be a stable numerical algorithm that can achieve our derived approximation rate. In other words, there is a special subclass of functions which arise in practice for which a continuous assignment of the weights exists. It is interesting future work to characterize such a subclass. Finally, to the best of our knowledge, there is not any efficient numerical algorithm to achieve the approximation rate with continuous weight selection especially when the dimension is large. Therefore, approximation results with or without continuous weight selection both have the difficulty of numerical implementations. Designing numerical algorithms to achieve these rates has been an active research field recently.

It is remarked that providing error estimate in the context of high-dimensional PDE problem using the newly introduced Floor-ReLU networks [44], which are fully connected neural networks with either Floor $\lfloor x \rfloor$ or ReLU activation function in each neuron, will be an intriguing option for future work, since these networks conquer the curse of dimensionality in terms of approximation theory. Other neural networks can also be considered. Along this line of work, the novel Floor-Exponential-Step networks with merely three hidden layers and width of order $\mathcal{O}(N)$ constructed in [46] can approximate $d$-dimensional Lipschitz continuous functions with an exponentially small error rate of order $\mathcal{O}(\sqrt{d}2^{-N})$. In [45], neural networks with a simple and computable continuous activation function and a fixed finite number of neurons were developed, which achieve the universal approximation property for all high-

dimensional continuous functions.

This paper is organized as follows. In Section 2, Sobolev spaces are briefly introduced and a number of useful existing results on ReLU network approximation are given. We provide our main approximation results for ReLU networks in Section 3. Several auxiliary results concerning approximating high order polynomials using ReLU networks are first provided in Section 3.1, and our main results on ReLU network are given in Section 3.2. In addition to ReLU networks, we show similar results for the smoother neural networks with ReLU square in Section 3.3. In Section 4, an extension of similar approximation results is provided for functions in the Hölder space.

## 2. Preliminaries

We provide in this section some useful preliminaries of notations and basic approximation results.

### *2.1. Deep Neural Networks*

Let us summarize all basic notations used in deep neural networks as follows.

1. Matrices are denoted by bold uppercase letters. For instance, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a real matrix of size $m \times n$, and $\boldsymbol{A}^T$ denotes the transpose of $\boldsymbol{A}$.

2. Vectors are denoted as bold lowercase letters. For example, $\boldsymbol{v} \in \mathbb{R}^n$ is a column vector of size $n$. Correspondingly, $\boldsymbol{v}(i)$ is the $i$-th element of $\boldsymbol{v}$. $\boldsymbol{v} = [v_1, \cdots, v_n]^T = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ are vectors consisting of numbers $\{v_i\}$ with $\boldsymbol{v}(i) = v_i$.

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

   (a) $|\boldsymbol{\alpha}| = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_d|$;

   (b) $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$, where $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T$;

   (c) $\boldsymbol{\alpha}! = \alpha_1!\alpha_2! \cdots \alpha_d!$;

4. Let $B_{r,|\cdot|}(\boldsymbol{x}) \subseteq \mathbb{R}^d$ be the closed ball with a center $\boldsymbol{x} \subseteq \mathbb{R}^d$ and a radius $r$ measured by the Euclidean distance. Similarly, $B_{r,\|\cdot\|_{\ell^\infty}}(\boldsymbol{x}) \subseteq \mathbb{R}^d$ is a ball measured by the discrete $\ell^\infty$-norm of a vector.

5. Assume $\boldsymbol{n} \in \mathbb{N}^n$, then $f(\boldsymbol{n}) = \mathcal{O}(g(\boldsymbol{n}))$ means that there exists positive $C$ independent of $\boldsymbol{n}$, $f$, and $g$ such that $f(\boldsymbol{n}) \leq Cg(\boldsymbol{n})$ when all entries of $\boldsymbol{n}$ go to $+\infty$.

6. We use $\sigma$ to denote an activation function. Let $\sigma_1 : \mathbb{R} \to \mathbb{R}$ denote the rectified linear unit (ReLU), i.e. $\sigma_1(x) = \max\{0, x\}$. With the abuse of notations, we define $\sigma_1 : \mathbb{R}^d \to \mathbb{R}^d$ as $\sigma_1(\boldsymbol{x}) = \begin{bmatrix} \max\{0, x_1\} \\ \vdots \\ \max\{0, x_d\} \end{bmatrix}$ for any $\boldsymbol{x} = [x_1, \cdots, x_d]^T \in \mathbb{R}^d$.

7. Furthermore, we let the activation function $\sigma_2 : \mathbb{R} \to \mathbb{R}$ be either $\sigma_1$ or $\sigma_1^2$. Similar to $\sigma_1$, we define the action of $\sigma_2$ on a vector $\boldsymbol{x}$.

8. We will use NN as a neural network for short and $\sigma_r$-NN **to specify an** NN **with activation functions** $\sigma_t$ **with** $t \leq r$. We will also use Python-type notations to specify a class of NN's, e.g., $\sigma_1$-NN($\mathrm{c}_1$; $\mathrm{c}_2$; $\cdots$ ; $\mathrm{c}_m$) is a set of ReLU FNNs satisfying $m$ conditions given by $\{\mathrm{c}_i\}_{1\leq i\leq m}$, each of which may specify the number of inputs (#input), the total number of nodes in all hidden layers (#node), the number of hidden layers (#layer), the number of total parameters (#parameter), and the width in each hidden layer (widthvec), the maximum width of all hidden layers (maxwidth), etc. For example, if $\phi \in \sigma_1$-NN(#input $= 2$; widthvec $= [100, 100]$), then $\phi$ satisfies

   (a) $\phi$ maps from $\mathbb{R}^2$ to $\mathbb{R}$.

   (b) $\phi$ has two hidden layers and the number of nodes in each hidden layer is 100.

9. $[n]^L$ is short for $[n, n, \cdots, n] \in \mathbb{N}^L$. For example,

$$\mathrm{NN}(\#\text{input} = d;\ \text{widthvec} = [100, 100]) = \mathrm{NN}(\#\text{input} = d;\ \text{widthvec} = [100]^2).$$

10. For $\phi \in \sigma$-NN(#input $= d$; widthvec $= [N_1, N_2, \cdots, N_L]$), if we define $N_0 = d$ and $N_{L+1} = 1$, then the architecture of $\phi$ can be briefly described as follows:

$$\boldsymbol{x} = \tilde{\boldsymbol{h}}_0 \xrightarrow{\boldsymbol{W}_1,\ \boldsymbol{b}_1} \boldsymbol{h}_1 \xrightarrow{\sigma} \tilde{\boldsymbol{h}}_1 \cdots \xrightarrow{\boldsymbol{W}_L,\ \boldsymbol{b}_L} \boldsymbol{h}_L \xrightarrow{\sigma} \tilde{\boldsymbol{h}}_L \xrightarrow{\boldsymbol{W}_{L+1},\ \boldsymbol{b}_{L+1}} \phi(\boldsymbol{x}) = \boldsymbol{h}_{L+1},$$

    where $\boldsymbol{W}_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the $i$-th linear transform in $\phi$, respectively, i.e.,

$$\boldsymbol{h}_i := \boldsymbol{W}_i \tilde{\boldsymbol{h}}_{i-1} + \boldsymbol{b}_i, \quad \text{for } i = 1, \ldots, L+1,$$

    and

$$\tilde{\boldsymbol{h}}_i = \sigma(\boldsymbol{h}_i), \quad \text{for } i = 1, \ldots, L.$$

    $L$ in this paper is also called the number of hidden layers in the literature.

11. The expression, an FNN with width $N$ and depth $L$, means

   (a) The maximum width of this FNN for all hidden layers less than or equal to $N$.

   (b) The number of hidden layers of this FNN less than or equal to $L$.

*2.2. Sobolev Spaces*

We will use $D$ to denote the weak derivative of a single variable function and $D^{\boldsymbol{\alpha}}$ to denote the partial derivative $D_1^{\alpha_1} D_2^{\alpha_2} \ldots D_d^{\alpha_d}$ of a $d$-dimensional function with $\alpha_i$ as the order of derivative $D_i$ in the $i$-th variable and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^T$. Let $\Omega$ denote an open subset of $\mathbb{R}^d$ and $L^p(\Omega)$ be the standard Lebesgue space on $\Omega$ for $p \in [1, \infty]$. We write $\nabla f := [D_1 f, \ldots, D_d f]^T$. $\partial\Omega$ is the boundary of $\Omega$. Let $\mu(\cdot)$ be the Lebesgue measure. For $f(x) \in \mathcal{W}^{n,p}(\Omega)$, we use the notation

$$\|f\|_{\mathcal{W}^{n,p}(\Omega)} = \|f\|_{\mathcal{W}^{n,p}} = \|f(x)\|_{\mathcal{W}^{n,p}(\Omega,\mu)},$$

if the domain is clear from the context and we use the Lebesgue measure.

**Definition 2.1.** *(Sobolev Space) Let $n \in \mathbb{N}_0$ and $1 \leq p \leq \infty$. Then we define the Sobolev space*

$$\mathcal{W}^{n,p}(\Omega) := \{f \in L^p(\Omega) : D^{\boldsymbol{\alpha}} f \in L^p(\Omega) \text{ for all } \boldsymbol{\alpha} \in \mathbb{N}_0^d \text{ with } |\boldsymbol{\alpha}| \leq n\}$$

*with a norm*

$$\|f\|_{\mathcal{W}^{n,p}(\Omega)} := \left( \sum_{0 \leq |\boldsymbol{\alpha}| \leq n} \|D^{\boldsymbol{\alpha}} f\|_{L^p(\Omega)}^p \right)^{1/p},$$

*if $p < \infty$, and*

$$\|f\|_{\mathcal{W}^{n,\infty}(\Omega)} := \max_{0 \leq |\boldsymbol{\alpha}| \leq n} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)}.$$

*2.3. Auxiliary Neutral Network Approximation Results*

We first give the following useful lemmas on several ReLU networks approximation results with explicit error characterization measured in the $L^\infty$ norm for polynomials.

**Lemma 2.1.** *The followings lemmas are satisfied by $\sigma_1$-NNs.*

*(i) Any one-dimensional continuous piecewise linear function with $N$ breakpoints can be exactly realized by a one-dimensional $\sigma_1$-NN with one-hidden layer and $N$ neurons.*

*(ii) Any identity map in $\mathbb{R}^d$ can be exactly realized by a $d$-dimensional $\sigma_1$-NN with one hidden layer and $2d$ neurons.*

*(iii) ([31, Lemma 5.1]) For any $N, L \in \mathbb{N}^+$, there exists a $\sigma_1$-NN $\phi$ with width $3N$ and depth $L$ such that*

$$\|\phi(x) - x^2\|_{L^\infty([0,1])} \leq N^{-L}.$$

*(iv) ([31, Lemma 4.2]) For any $N, L \in \mathbb{N}^+$ and $a, b \in \mathbb{R}$ with $a < b$, there exists a $\sigma_1$-NN $\phi$ with width $9N + 1$ and depth $L$ such that*

$$\|\phi(x, y) - xy\|_{L^\infty([a,b]^2)} \leq 6(b-a)^2 N^{-L}.$$

*(v) ([31, Theorem 4.1]) Assume $P(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $\|\boldsymbol{\alpha}\|_1 \leq k \in \mathbb{N}^+$. For any $N, L \in \mathbb{N}^+$, there exists a $\sigma_1$-NN $\phi$ with width $9(N+1)+k-1$ and depth $7k^2L$ such that*

$$\|\phi(\boldsymbol{x}) - P(\boldsymbol{x})\|_{L^\infty([0,1]^d)} \leq 9k(N+1)^{-7kL}.$$

The following two propositions will also be used in proving our main results.

**Proposition 2.2.** *(Step function approximations [31, Proposition 4.3]) For any $N, L, d \in \mathbb{N}^+$ and $\delta \in (0, \frac{1}{3K}]$ with $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$, there exists a one-dimensional $\sigma_1$-NN $\phi$ with width $4\lfloor N^{1/d} \rfloor + 3$ and depth $4L + 5$ such that*

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

*for $k = 0, 1, \ldots, K-1$.*

**Proposition 2.3.** *(Point matching [31, Proposition 4.4]) Given any $N, L, s \in \mathbb{N}^+$ and $\xi_i \in [0,1]$ for $i = 0, 1, \ldots, N^2L^2 - 1$, there exists a $\sigma_1$-NN $\phi$ with width $16s(N+1)\log_2(8N)$ and depth $(5L+2)\log_2(4L)$ such that*

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

## 3. Proof of Our Main Results

In this section, the proofs of our main results are given. We first define the following notion of subset of $[0,1]^d$, before presenting our main approximation results. Given any $K \in N^+$ and $\delta \in (0, \frac{1}{K})$, define a trifling region $\Omega([0,1]^d, K, \delta, d)$ of $[0,1]^d$ as

$$\Omega([0,1]^d, K, \delta, d) := \cup_{i=1}^d \{\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T \in [0,1]^d : x_i \in \cup_{k=1}^{K-1} (\tfrac{k}{K} - \delta, \tfrac{k}{K})\}. \quad (3)$$

*3.1. Approximations Using ReLU Neural Networks*

We first present the following Theorem 3.1, which concerns the approximation result by ReLU networks outside the trifling region $\Omega([0,1]^d, K, \delta, d)$.

**Theorem 3.1.** *Suppose that $f \in C^s([0,1]^d)$ with an integer $s > 1$ satisfies $\|\partial^{\boldsymbol{\alpha}} f\|_{L^\infty([0,1]^d)} < 1$ for any $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq s$. For any $N, L \in \mathbb{N}^+$, there exists a $\sigma_1$-NN $\phi$ with width $16 s^{d+1} d(N+2)\log_2(8N)$ and depth $27 s^2 (L+2)\log_2(4L)$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \leq 432 s^d$ and*

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

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

We will now show our main result, Theorem 1.1, provided Theorem 3.1 holds true. The proof of Theorem 3.1 will be given in Section 3.2.

**Proof of Theorem 1.1** When $f$ is a constant function, the statement is trivial. By Theorem 3.1, there exists a $\sigma_1$-NN $\phi$ with width $16 s^{d+1} d(N+2)\log_2(8N)$ and depth $27 s^2 (L+2)\log_2(4L)$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \leq 432 s^d$ and

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

Now, we set $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$ and choose a small $\delta$ such that

$$Kd\delta 432^p \leq (N^{-2(s-1)/d} L^{-2(s-1)/d})^p.$$

Then, we have

$$\begin{aligned}
\|f - \phi\|^p_{\mathcal{W}^{1,p}([0,1]^d)} &= \|f - \phi\|^p_{\mathcal{W}^{1,p}(\Omega([0,1]^d, K, \delta, d))} + \|f - \phi\|^p_{\mathcal{W}^{1,p}([0,1]^d \setminus \Omega([0,1]^d, K, \delta, d))} \\
&\leq Kd\delta (432 s^d)^p + (84(s+1)^d 8^s N^{-2(s-1)/d} L^{-2(s-1)/d})^p \\
&\leq \left(s^d N^{-2(s-1)/d} L^{-2(s-1)/d}\right)^p + \left(84(s+1)^d 8^s N^{-2(s-1)/d} L^{-2(s-1)/d}\right)^p \\
&\leq (85(s+1)^d 8^s N^{-2(s-1)/d} L^{-2(s-1)/d})^p.
\end{aligned}$$

Hence, we have

$$\|f - \phi\|_{\mathcal{W}^{1,p}([0,1]^d)} \leq 85(s+1)^d 8^s N^{-2(s-1)/d} L^{-2(s-1)/d}.$$

□

Before showing Theorem 3.1 directly, we present the following lemmas which will be helpful to understand the different steps illustrated in the proof. Similar to showing [31, Theorem 2.2] where only the $L^\infty$ norm is considered, in order to prove Theorem 3.1 we will first construct ReLU networks to approximate multivariate polynomials and provide an error bound measured in the $\mathcal{W}^{1,\infty}$ norm that is explicitly characterized in terms of the layer and depth of the underlying ReLU networks, via the following steps:

1. We approximate $f(x) = x^2$ by the compositions of the "sawtooth functions", which was first proposed in [51].
2. We approximate $f(x, y) = xy$ using the ReLU network constructed in the previous step based on the identity $xy = 2\Big(\big(\frac{|x+y|}{2}\big)^2 - \big(\frac{|x|}{2}\big)^2 - \big(\frac{|y|}{2}\big)^2\Big)$.
3. We approximate $f(x_1, x_2, \cdots, x_k) = x_1 x_2 \cdots x_k$ for $k \geq 2$ repeatedly using the ReLU networks constructed in the previous step.
4. We approximate a general polynomial $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq k \in \mathbb{N}^+$. Since any one term polynomial of degree than or equal to $k$ can be written as $Cz_1 z_2 \cdots z_k$, where $C$ is a constant, we can obtain an error bound based on the previous approximation result.

We begin with giving an approximation result for the simple function of $x^2$.

**Lemma 3.2.** *For any $N, L \in \mathbb{N}^+$, there exists a $\sigma_1$-NN $\phi$ with width $3N$ and depth $2L$ such that $\|\phi(x)\|_{\mathcal{W}^{1,\infty}([0,1])} \leq 2$ and*

$$||\phi(x) - x^2||_{\mathcal{W}^{1,\infty}((0,1))} \leq N^{-L}.$$

**Proof** As in the proof of Lemma 2.1 (iii), we begin by defining the piecewise linear function $f_s : [0, 1] \to [0, 1]$ for $s \in \mathbb{N}^+$, $s = 1, 2, \ldots$, satisfying the following properties

1. $f_s(x) = x^2$ at a set of break points $\big\{\frac{j}{2^s} : j = 0, 1, 2, \ldots, 2^s\big\}$.
2. $f_s(x)$ is linear between any two adjacent break points.

For any integer $N$, let $k \in \mathbb{N}^+$ be the unique number such that $(k-1)2^{k-1} + 1 \leq N \leq k2^k$. For such an $N$, $k$, and any $L'$, let $s = L'k$. It is shown in the proof of [31, Lemma 5.1] that there exists a $\sigma_1$-NN $\phi(x) = f_s(x) = f_{L'k}(x)$ with width $3N$ and depth $L'$ such that

$$\|\phi(x) - x^2\|_{L^\infty([0,1])} = \|f_{L'k} - x^2\|_{L^\infty([0,1])} \leq 2^{-2(L'k+1)} \leq 2^{-2L'k} \leq N^{-L'}. \tag{4}$$

We now show that the approximation error of the first order (weak) derivative can be measured in a similar fashion.

Note that for all $x \in \big(\frac{j}{2^s}, \frac{j+1}{2^s}\big)$,

$$\phi(x) = \Big(\frac{(j+1)^2}{2^s} - \frac{j^2}{2^s}\Big)\Big(x - \frac{j}{2^s}\Big) + \Big(\frac{j}{2^s}\Big)^2. \tag{5}$$

From (5), we have for each $j = 0, 1, \ldots, 2^s - 1$,

$$\begin{aligned}
|\phi(x) - x^2|_{\mathcal{W}^{1,\infty}\left(\left(\frac{j}{2^s}, \frac{j+1}{2^s}\right)\right)} &= \left\| \tfrac{(j+1)^2}{2^s} - \tfrac{j^2}{2^s} - 2x \right\|_{L^\infty\left(\left(\frac{j}{2^s}, \frac{j+1}{2^s}\right)\right)} \\
&= \left\| \tfrac{2j+1}{2^s} - 2x \right\|_{L^\infty\left(\left(\frac{j}{2^s}, \frac{j+1}{2^s}\right)\right)} \\
&= \max\left\{ \left| \tfrac{2j+1}{2^s} - 2\left(\tfrac{j}{2^s}\right) \right|, \left| \tfrac{2j+1}{2^s} - 2\left(\tfrac{j+1}{2^s}\right) \right| \right\} \\
&= 2^{-s} \\
&= 2^{-L'k}.
\end{aligned}$$

Clearly, from (4) & (5) we have

$$\|\phi(x) - x^2\|_{\mathcal{W}^{1,\infty}((0,1))} = \|f_{L'k} - x^2\|_{\mathcal{W}^{1,\infty}((0,1))} \leq 2^{-L'k} \leq N^{-L'/2}.$$

Finally, we have

$$\begin{aligned}
\|\phi(x)\|_{\mathcal{W}^{1,\infty}([0,1])} \leq \left\| \tfrac{(j+1)^2 - j^2}{2^s} \right\|_{L^\infty\left(\left(\frac{j}{2^s}, \frac{j+1}{2^s}\right)\right)} = \tfrac{2j+1}{2^s} &\leq \tfrac{2(2^s-1)+1}{2^s} \\
&= 2 - \tfrac{1}{2^{L'k}} \\
&\leq 2.
\end{aligned}$$

Setting $L' = 2L$, we have the desired $\phi(x)$ and hence the proof is finished. □

**Lemma 3.3.** *For any $N, L \in \mathbb{N}^+$, there exists a $\sigma_1$-NN $\phi$ with width $9N$ and depth $2L$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \leq 12$ and*

$$\|\phi(x, y) - xy\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \leq 6N^{-L}.$$

**Proof** By Lemma 3.2, there exists a $\sigma_1$-NN $\psi$ with width $3N$ and depth $2L$ such that $\|\psi\|_{\mathcal{W}^{1,\infty}((0,1))} \leq 2$ and

$$\|z^2 - \psi(z)\|_{\mathcal{W}^{1,\infty}((0,1))} \leq N^{-L}.$$

Combining the above inequality and the fact that for any $x, y \in \mathbb{R}$

$$xy = 2\Big(\big(\tfrac{|x+y|}{2}\big)^2 - \big(\tfrac{|x|}{2}\big)^2 - \big(\tfrac{|y|}{2}\big)^2\Big),$$

we construct the following network $\phi$

$$\phi(x, y) = 2\Big(\psi\big(\tfrac{|x+y|}{2}\big) - \psi\big(\tfrac{|x|}{2}\big) - \psi\big(\tfrac{|y|}{2}\big)\Big).$$

We have

$$\begin{aligned}
& \|\phi(x, y) - xy\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \\
\leq\ & 2\big\|\psi\big(\tfrac{|x+y|}{2}\big) - \big(\tfrac{|x+y|}{2}\big)^2\big\|_{\mathcal{W}^{1,\infty}((0,1)^2)} + 2\big\|\psi\big(\tfrac{|x|}{2}\big) - \big(\tfrac{|x|}{2}\big)^2\big\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \\
& +2\big\|\psi\big(\tfrac{|y|}{2}\big) - \big(\tfrac{|y|}{2}\big)^2\big\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \\
\leq\ & 2N^{-L} + 2N^{-L} + 2N^{-L} \\
=\ & 6N^{-L}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}\|\phi(x,y)\|_{\mathcal{W}^{1,\infty}((0,1)^2)} &= \Big\|2\Big(\psi\big(\tfrac{|x+y|}{2}\big)-\psi\big(\tfrac{|x|}{2}\big)-\psi\big(\tfrac{|y|}{2}\big)\Big)\Big\|_{\mathcal{W}^{1,\infty}((0,1)^2)}\\ &\le 12.\end{aligned}$$

□

By rescaling, we have the following modification of Lemma 3.3.

**Lemma 3.4.** *For any $N, L \in \mathbb{N}^+$ and $a, b \in \mathbb{R}$ with $a < b$, there exists a $\sigma_1$-NN $\phi$ with width $9N+1$ and depth $2L$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}((a,b)^2)} \le 12(b-a)^2$ and*

$$\|\phi(x,y)-xy\|_{\mathcal{W}^{1,\infty}((a,b)^2)} \le 6(b-a)^2N^{-L}.$$

**Proof** By Lemma 3.3, there exists a $\sigma_1$-NN $\psi$ with width $9N$ and depth $2L$ such that $\|\psi\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \le 12$ and

$$\|\psi(\widetilde{x},\widetilde{y})-\widetilde{x}\widetilde{y}\|_{\mathcal{W}^{1,\infty}((0,1)^2)} \le 6N^{-L}.$$

By setting $x = a+(b-a)\widetilde{x}$ and $y = a+(b-a)\widetilde{y}$ for any $\widetilde{x}, \widetilde{y} \in (0,1)$, we define the following network $\phi$

$$\phi(x,y) = (b-a)^2\psi\big(\tfrac{x-a}{b-a}, \tfrac{y-a}{b-a}\big) + a(x-a) + a(y-a) + a^2.$$

Note that $a(x-a)+a(y-a)$ is positive. Hence, the width of $\phi$ can be as small as $9N+1$. Thus, by $xy = (b-a)^2(\frac{x-a}{b-a}\cdot\frac{y-a}{b-a}) + a(x-a)+a(y-a)+a^2$, we have

$$\begin{aligned}\|\phi(x,y)-xy\|_{\mathcal{W}^{1,\infty}((a,b)^2)} &= (b-a)^2\big\|\psi\big(\tfrac{x-a}{b-a},\tfrac{y-a}{b-a}\big)-\big(\tfrac{x-a}{b-a}\cdot\tfrac{y-a}{b-a}\big)\big\|_{\mathcal{W}^{1,\infty}((a,b)^2)}\\ &\le 6(b-a)^2N^{-L}.\end{aligned}$$

Finally, we have

$$\begin{aligned}\|\phi(x,y)\|_{\mathcal{W}^{1,\infty}((a,b)^2)} &= \Big\|(b-a)^2\psi\big(\tfrac{x-a}{b-a},\tfrac{y-a}{b-a}\big)+a(x-a)+a(y-a)+a^2\Big\|_{\mathcal{W}^{1,\infty}((a,b)^2)}\\ &\le 12(b-a)^2.\end{aligned}$$

□

**Lemma 3.5.** *For any $N, L, k \in \mathbb{N}^+$ with $k \ge 2$, there exists a $\sigma_1$-NN $\phi$ with width $9(N+1)+k-1$ and depth $14k(k-1)L$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}((0,1)^k)} \le 18$ and*

$$\|\phi(\boldsymbol{x})-x_1x_2\cdots x_k\|_{\mathcal{W}^{1,\infty}((0,1)^k)} \le 10(k-1)(N+1)^{-7kL}.$$

**Proof** By Lemma 3.4, there exists a $\sigma_1$-NN $\phi_1$ with width $9(N+1)+1$ and depth $14kL$ such that $\|\phi_1\|_{\mathcal{W}^{1,\infty}((-0.1,1.1)^2))} \le 18$ and

$$\begin{aligned}\|\phi_1(x,y)-xy\|_{\mathcal{W}^{1,\infty}((-0.1,1.1)^2)} &\le 6(1.2)^2(N+1)^{-7kL}\\ &\le 9(N+1)^{-7kL}.\end{aligned}$$

Now, our goal is to construct via induction that for $i = 1, 2, \ldots, k-1$ there exists a $\sigma_1$-NN $\phi_i$ with width $9(N+1)+i$ and depth $14kiL$ such that

$$\|\phi_i(x_1, \cdots, x_{i+1}) - x_1 x_2 \cdots x_{i+1}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+1})} \leq 10i(N+1)^{-7kL},$$

for any $[x_1, x_2, \cdots, x_{i+1}]^T \in (0,1)^{i+1}$.

When $d = 1$, $\phi_1$ satisfies the condition.

Assuming now for any $i \in \{1, 2, \ldots, d-1\}$ there exists a $\sigma_1$-NN $\phi_i$ such that the both conditions hold, we define $\phi_{i+1}$ as follows:

$$\phi_{i+1}(x_1, \cdots, x_{i+2}) = \phi_1(\phi_i(x_1, \cdots, x_{i+1}), \sigma(x_{i+2}))$$

for any $x_1, \cdots, x_{k+2} \in \mathbb{R}$. We can shift $x_{i+2}$ to obtain a nonnegative number $x_{i+2} + 0.1$, which can be copied via a ReLU network of width 1 to the input of $\phi_1$. Before inputting $x_{i+2} + 0.1$ to $\phi_1$, we can shift it back to $x_{i+2}$. Therefore, $\phi_{i+1}$ can be implemented via a $\sigma_1$-NN with width $9(N+1)+i+1$ and depth $14kiL + 14kL = 14k(i+1)L$.

By the induction assumption, we have

$$\|\phi_i(x_1, x_2, \cdots, x_{i+1}) - x_1 x_2 \cdots x_{i+1}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+1})} \leq 10i(N+1)^{-7kL}.$$

Note that $10i(N+1)^{-7kL} \leq 10k2^{-7k} < 10k\frac{1}{100k} = 0.1$ for any $N, L, k \geq d \in \mathbb{N}^+$ and $i \in \{1, 2, \cdots, d-1\}$. Thus, we have

$$\phi_i(x_1, x_2, \cdots, x_{i+1}) \in (-0.1, 1.1)$$

and

$$\frac{\partial \phi_i}{\partial x_1} \in (-0.1, 1.1)$$

for any $x_1, x_2, \cdots, x_{i+1} \in (0,1)$.

Hence, for any $x_1, x_2, \cdots, x_{i+2} \in (0,1)$, we have

$$\begin{aligned}
& \|\phi_{i+1}(x_1, \cdots, x_{i+2}) - x_1 \cdots x_{i+2}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})} \\
= \; & \|\phi_1(\phi_i(x_1, \cdots, x_{i+1}), \sigma(x_{i+2})) - x_1 \cdots x_{i+2}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})} \\
\leq \; & \|\phi_1(\phi_i(x_1, \cdots, x_{i+1}), x_{i+2}) - \phi_i(x_1, \cdots, x_{i+1}) x_{i+2}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})} \\
& + \|\phi_i(x_1, \cdots, x_{i+1}) x_{i+2} - x_1 \cdots x_{i+2}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})} \\
\leq \; & 10(N+1)^{-7kL} + 10i(N+1)^{-7kL} \\
= \; & 10(i+1)(N+1)^{-7kL}, \qquad (6)
\end{aligned}$$

where the second inequality is obtained since we have

$$\begin{aligned}
\left| \frac{\partial(\phi_1(\phi_i(x_1, \cdots, x_{i+1}), x_{i+2}) - \phi_i(x_1, \cdots, x_{i+1}) x_{i+2})}{\partial x_1} \right| &= \left| \frac{\partial \phi_1(\phi_i(x_1, \cdots, x_{i+1}), x_{i+2})}{\partial \phi_i} \cdot \frac{\partial \phi_i}{\partial x_1} - x_{i+2} \cdot \frac{\partial \phi_i}{\partial x_1} \right| \\
&= \underbrace{\left| \frac{\partial \phi_1(\phi_i(x_1, \cdots, x_{i+1}), x_{i+2})}{\partial \phi_i} - x_{i+2} \right|}_{\leq 9(N+1)^{-7kL}} \underbrace{\left| \frac{\partial \phi_i}{\partial x_1} \right|}_{<=1.1} \\
&\leq 10(N+1)^{-7kL}
\end{aligned}$$

and

$$\begin{aligned}\left|\frac{\partial(\phi_1(\phi_i(x_1,\cdots,x_{i+1}),x_{i+2})-\phi_i(x_1,\cdots,x_{i+1})x_{i+2})}{\partial x_{i+2}}\right| &= \left|\frac{\partial\phi_1(\phi_i(x_1,\cdots,x_{i+1}),x_{i+2})}{\partial x_{i+2}}-\phi_i\right| \\ &\leq 9(N+1)^{-7kL} \\ &\leq 10(N+1)^{-7kL}.\end{aligned}$$

Thus,

$$|\phi_i(x_1,\cdots,x_{i+1})x_{i+2}-x_1\cdots x_{i+2}|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})}\leq 10(N+1)^{-7kL}.$$

Also, using similar arguments or see the proof of [31, Lemma 5.1], it can be shown that

$$|\phi_i(x_1,\cdots,x_{i+1})x_{i+2}-x_1\cdots x_{i+2}|_{L^\infty((0,1)^{i+2})}\leq 10(N+1)^{-14kL}.$$

Hence, we have shown

$$\|\phi_1(\phi_i(x_1,\cdots,x_{i+1}),x_{i+2})-\phi_i(x_1,\cdots,x_{i+1})x_{i+2}\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})}\leq 10(N+1)^{-7kL},$$

which was used in showing (6).

Also, for any $x_1,x_2,\cdots,x_{i+2}\in(0,1)$, we have

$$\begin{aligned}\|\phi_{i+1}(x_1,\cdots,x_{i+2})\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})} &= \|\phi_1(\phi_i(x_1,\cdots,x_{i+1}),\sigma(x_{i+2}))\|_{\mathcal{W}^{1,\infty}((0,1)^{i+2})} \\ &\leq 18.\end{aligned}$$

At last, setting $\phi=\phi_{d-1}$ in (6), we have finished the proof by induction. □

**Proposition 3.6.** *Suppose $\boldsymbol{x}^{\boldsymbol{\alpha}}=x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha}\in\mathbb{N}^d$ with $|\boldsymbol{\alpha}|\leq k\in\mathbb{N}^+$. For any $N,L\in\mathbb{N}^+$, there exists a $\sigma_1$-NN $\phi$ with width $9(N+1)+k-1$ and depth $14k^2L$ such that $\|\phi\|_{\mathcal{W}^{1,\infty}([0,1]^d)}\leq 18$ and*

$$\|\phi(\boldsymbol{x})-\boldsymbol{x}^{\boldsymbol{\alpha}}\|_{\mathcal{W}^{1,\infty}([0,1]^d)}\leq 10k(N+1)^{-7kL}.$$

**Proof** This proof is similar to that of [31, Proposition 4.1], but for readability we provide a detailed proof as follows.

The case when $k=1$ is trivial. When $k\geq 2$, we set $\widetilde{k}=|\boldsymbol{\alpha}|\leq k$, $\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\ldots,\alpha_d]^T\in\mathbb{R}^d$, and let $[z_1,z_2,\cdots,z_{\widetilde{k}}]^T\in\mathbb{R}^{\widetilde{k}}$ be the vector such that

$$z_l=x_j,\quad \text{if } \sum_{i=1}^{j-1}\alpha_i<l\leq\sum_{i=1}^{j}\leq\alpha_j,\quad \text{for } j=1,2,\cdots,d.$$

In other words, we have

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

We now construct the target deep ReLU neural network. First, there exists a linear map $\mathcal{L}:\mathbb{R}^d\to\mathbb{R}^k$ $\mathcal{L}(\boldsymbol{x})=[z_1,z_2,\cdots,z_{\widetilde{k}},1,\cdots,1]$, which copies $\boldsymbol{x}$ to form a new vector

$[z_1, z_2, \cdots, z_{\widetilde{k}}, 1, \cdots, 1] \in \mathbb{R}^k$. Second, there exists by Lemma 3.5 a function $\psi : \mathbb{R}^k \to \mathbb{R}$ implemented by a ReLU network with width $9(N+1)+k-1$ and depth $14k(k-1)L$ such that $\|\psi\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \leq 18$ and $\psi$ maps $[z_1, z_2, \cdots, z_{\widetilde{k}}, 1, \cdots, 1]$ to $z_1 z_2 \cdots z_{\widetilde{k}}$ within an error of $10(k-1)(N+1)^{-7kL}$. Thus, we can construct the network $\phi = \psi \circ \mathcal{L}$. Then, $\phi$ can be implemented by a ReLU with width $9(N+1)+k-1$ and depth $14k(k-1)L \leq 14k^2L$, and

$$\begin{aligned}
\|\phi(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{\mathcal{W}^{1,\infty}([0,1]^d)} &= \|\psi \circ \mathcal{L}(\boldsymbol{x}) - x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&= \|\psi(z_1, z_2, \cdots, z_{\widetilde{k}}, 1, \cdots, 1) - z_1 z_2 \cdots z_{\widetilde{k}}\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&\leq 10(k-1)(N+1)^{-7kL} \\
&\leq 10k(N+1)^{-7kL}.
\end{aligned}$$

Also, we have

$$\begin{aligned}
\|\phi(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}([0,1]^d)} &= \|\psi \circ \mathcal{L}(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&= \|\psi(z_1, z_2, \cdots, z_{\widetilde{k}}, 1, \cdots, 1)\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&\leq 18.
\end{aligned}$$

This proof is then finished. □

We are now ready to show our main approximation result of Theorem 3.1.

*3.2. Proof of Theorem 3.1*

**Proof** We set $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$ and let $\Omega([0,1]^d, K, \delta, d)$ defined by (3) partition $[0,1]^d$ into $K^d$ cubes $Q_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta} \in \{0, 1, \cdots, K-1\}^d$ such that

$$[0,1]^d = \Omega([0,1]^d, K, \delta, d) \bigcup \big( \cup_{\boldsymbol{\beta} \in \{0,1,\cdots,K-1\}^d} Q_{\boldsymbol{\beta}} \big).$$

For each $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_d]^T \in \{0, 1, \cdots, K-1\}^d$, we define

$$Q_{\boldsymbol{\beta}} = \left\{ \boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T : x_i \in \left[ \tfrac{\beta_i}{K}, \tfrac{\beta_i+1}{K} - \delta \cdot 1_{\{\beta_i \leq K-2\}}(\beta_i) \right] \text{ for } i = 1, 2, \ldots, d \right\},$$

where $1_{\{\beta_i \leq K-2\}}(\beta_i)$ is the indicator function of the set $\{\beta_i \leq K-2\}$.

By Proposition 2.2, there exists a $\sigma_1$-NN $\psi$ with width $4N+3$ and depth $4L+5$ such that

$$\psi(x) = k, \quad \text{if } x \in \left[ \tfrac{k}{K}, \tfrac{k+1}{K} - \delta \cdot 1_{\{k \leq K-2\}} \right] \text{ for } k = 0, 1, \ldots, K-1.$$

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

Define

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

then

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

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

$$f(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}|\le s-1} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} + \sum_{|\boldsymbol{\alpha}|=s} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x})+\xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}},$$

where $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{\psi}(\boldsymbol{x})$.

By Lemma 3.4, there exists a $\sigma_1$-NN $\widetilde{\phi}$ with width $9(N+1)+1$ and depth $4s(L+1)$ such that $\|\widetilde{\phi}\|_{\mathcal{W}^{1,\infty}((-3,3)^2)} \le 432$ and

$$\begin{aligned}\|\widetilde{\phi}(x,y) - xy\|_{\mathcal{W}^{1,\infty}((-3,3)^2))} &\le 6(6)^2(N+1)^{-2s(L+1)} \\ &= 216(N+1)^{-2s(L+1)} =: \mathcal{E}_1.\end{aligned}$$

For each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \le s$, by Proposition 3.6 there exist $\sigma_1$-NN$s$ $P_{\boldsymbol{\alpha}}(\boldsymbol{x})$ with width $9(N+1)+s-1$ and depth $14s^2L$ such that

$$\|P_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \le 10s(N+1)^{-7sL} =: \mathcal{E}_2. \tag{7}$$

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

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

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

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

Note that $K^d = (\lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor)^d \le N^2L^2$ and $\xi_{\boldsymbol{\alpha},i} \in [0,1]$ for $i = 0, 1, \cdots, K^d - 1$. By Proposition 2.3, there exists a $\sigma_1$-NN $\widetilde{\phi}_{\boldsymbol{\alpha}}$ of width $16s(N+1)\log_2(8N)$ and depth $5(L+2)\log_2(4L)$ such that

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

For each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \le s-1$, we define

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

It can be seen that $\phi_{\boldsymbol{\alpha}}$ is also of width $16s(N+1)\log_2(8N)$ and depth $5(L+2)\log_2(4L)$.

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

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

From $\boldsymbol{\psi}(\boldsymbol{x}) = \frac{\boldsymbol{\beta}}{K}$ for $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$, it follows that

$$\begin{aligned}\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} &= \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))\|_{L^\infty(Q_{\boldsymbol{\beta}})} \\ &= \big|\phi_{\boldsymbol{\alpha}}(\tfrac{\boldsymbol{\beta}}{K}) - \partial^{\boldsymbol{\alpha}} f(\tfrac{\boldsymbol{\beta}}{K})\big| \\ &\le 2N^{-2s}L^{-2s} =: \mathcal{E}_3. \end{aligned} \tag{8}$$

Note that since $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))$ for $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$ is constant, its weak derivative is zero, which has given us the first equality of (8).

Define

$$\phi(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \le s-1} \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) \tag{9}$$

for any $\boldsymbol{x} \in \mathbb{R}^d$. Let us now estimate the error for any $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$.

$$\begin{aligned} & \|\phi(\boldsymbol{x}) - f(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\theta}})} \\ \le & \underbrace{\sum_{|\boldsymbol{\alpha}| \le s-1} \Big\| \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_1} \\ & + \underbrace{\sum_{|\boldsymbol{\alpha}| = s} \Big\| \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}} \boldsymbol{h})}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_2}, \end{aligned}$$

where $E_1$ is further decomposed into two parts via

$$\begin{aligned} E_1 &= \sum_{|\boldsymbol{\alpha}| \le s-1} \Big\| \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} \\ &\le \underbrace{\sum_{|\boldsymbol{\alpha}| \le s-1} \Big\| \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \widetilde{\phi}\Big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) \Big\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_{1,1}} \\ &\quad + \underbrace{\sum_{|\boldsymbol{\alpha}| \le s-1} \Big\| \widetilde{\phi}\Big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_{1,2}}. \end{aligned}$$

For each $\boldsymbol{\alpha} \in \mathbb{R}^d$ with $|\boldsymbol{\alpha}| \le s-1$ and $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$, since $\mathcal{E}_3 \in [0,2]$ and $\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x})) \in (-1,1)$ according to (8), we have $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) \in (-3,3)$ and hence $\frac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \in (-3,3)$.

Also, we have $P_{\boldsymbol{\alpha}}(\boldsymbol{x}) \in (-2,3) \subseteq (-3,3)$ and $\|P_{\boldsymbol{\alpha}}(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \le 3$ for any $\boldsymbol{x} \in [0,1]^d$ and $|\boldsymbol{\alpha}| \le s-1$, since $\mathcal{E}_2 < 2$ from (7).

Hence, we can now measure $E_{1,1}$, $E_{1,2}$ and $E_2$:

$$
\begin{aligned}
E_{1,1} &= \sum_{|\boldsymbol{\alpha}|\le s-1} \left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le \sum_{|\boldsymbol{\alpha}|\le s-1} \Bigg( \underbrace{\left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{\le \mathcal{E}_1} \\
&\quad + \underbrace{\left\| \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{\le \mathcal{E}_1} \\
&\quad + \underbrace{\left\| \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{\le 3\mathcal{E}_3} \Bigg) \\
&\le \sum_{|\boldsymbol{\alpha}|\le s-1} (2\mathcal{E}_1 + 3\mathcal{E}_3) \\
&\le s^d (2\mathcal{E}_1 + 3\mathcal{E}_3).
\end{aligned}
$$

Note that the last inequality is followed by the fact that $\sum_{|\boldsymbol{\alpha}|\le s-1} 1 \le \sum_{i=0}^{s-1} (i+1)^{d-1} \le s^d$. Thus,

$$
\begin{aligned}
E_{1,2} &= \sum_{|\boldsymbol{\alpha}|\le s-1} \left\| \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le \sum_{|\boldsymbol{\alpha}|\le s-1} \underbrace{\left\| \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{\le \mathcal{E}_1} \\
&\quad + \sum_{|\boldsymbol{\alpha}|\le s-1} \underbrace{\left\| \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})}}_{\le \mathcal{E}_2} \\
&\le \sum_{|\boldsymbol{\alpha}|\le s-1} (\mathcal{E}_1 + \mathcal{E}_2) \\
&\le s^d (\mathcal{E}_1 + \mathcal{E}_2)
\end{aligned}
$$

and

$$
\begin{aligned}
E_2 &= \sum_{|\boldsymbol{\alpha}|= s} \left\| \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}} \boldsymbol{h})}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le \sum_{|\boldsymbol{\alpha}|= s} \left\| \tfrac{1}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le (s+1)^{d-1} K^{-(s-1)}.
\end{aligned}
$$

Note that the last inequality is followed by the fact that $\sum_{|\boldsymbol{\alpha}|= s} 1 \le (s+1)^{d-1}$.

Using $(N+1)^{-7s(L+1)} \leq (N+1)^{-2s(L+1)} \leq (N+1)^{-2s}2^{-2sL} \leq N^{-2s}L^{-2s}$ and $K = \lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor \geq \frac{N^{2/d}L^{2/d}}{8}$, we have

$$\begin{aligned}
&\|\phi(\boldsymbol{x}) - f(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}(Q_{\boldsymbol{\beta}})} \\
\leq\ & E_{1,1} + E_{1,2} + E_2 \\
=\ & s^d(2\mathcal{E}_1 + 3\mathcal{E}_3) + s^d(\mathcal{E}_1 + \mathcal{E}_2) + (s+1)^{d-1}K^{-(s-1)} \\
\leq\ & (s+1)^d(K^{-(s-1)} + 3\mathcal{E}_1 + \mathcal{E}_2 + 3\mathcal{E}_3) \\
\leq\ & (s+1)^d\big(K^{-(s-1)} + 648(N+1)^{-2s(L+1)} + 10s(N+1)^{-7s(L+1)} + 6N^{-2s}L^{-2s}\big) \\
\leq\ & (s+1)^d\big(8^{s-1}N^{-2(s-1)/d}L^{-2(s-1)/d} + (654+10s)N^{-2s}L^{-2s}\big) \\
\leq\ & (s+1)^d(8^{s-1} + 654 + 10s)N^{-2(s-1)/d}L^{-2(s-1)/d} \\
\leq\ & 84(s+1)^d 8^s N^{-2(s-1)/d}L^{-2(s-1)/d}.
\end{aligned}$$

Since $\boldsymbol{\beta} \in \{0,1,2,\cdots,K-1\}^d$ is arbitrary and the fact that $[0,1]^d\backslash\Omega([0,1]^d,K,\delta,d) \subseteq \cup_{\boldsymbol{\beta}\in\{0,1,\cdots,K-1\}^d}Q_{\boldsymbol{\beta}}$, we have

$$\|\phi(\boldsymbol{x}) - f(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}([0,1]^d\backslash\Omega([0,1]^d,K,\delta,d))} \leq 84(s+1)^d 8^s N^{-2(s-1)/d}L^{-2(s-1)/d}.$$

Furthermore, we have

$$\begin{aligned}
\|\phi(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}([0,1]^d)} &= \Big\|\sum_{|\boldsymbol{\alpha}|\leq s-1}\widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big)\Big\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&\leq \sum_{|\boldsymbol{\alpha}|\leq s-1}\|\widetilde{\phi}\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&\leq 432s^d.
\end{aligned}$$

As last, we finish the proof by estimating the width and depth of the network implementing $\phi(\boldsymbol{x})$. From (9), we know that $\phi(\boldsymbol{x})$ consists of the following subnetworks:

1. $\boldsymbol{\psi} \in \mathrm{NN}(\text{width} \leq d(4N+3); \text{depth} \leq 4L+5)$;
2. $\phi_{\boldsymbol{\alpha}} \in \mathrm{NN}(\text{width} \leq 16s(N+1)\log_2(8N); \text{depth} \leq 5(L+2)\log_2(4L))$;
3. $P_{\boldsymbol{\alpha}} \in \mathrm{NN}(\text{width} \leq 9(N+1)+s-1; \text{depth} \leq 14s^2L)$;
4. $\widetilde{\phi} \in \mathrm{NN}(\text{width} \leq 9(N+1)+1; \text{depth} \leq 4s(L+1))$.

Thus, we can infer that the $\phi$ can be implemented by a ReLU network with width $16s^{d+1}d(N+2)\log_2(8N)$ and depth

$$(4L+5) + 4s(L+1) + 14s^2L + 5(L+2)\log_2(4L) + 3 \leq 27s^2(L+2)\log_2(4L).$$

Hence, we have finished the proof. □

*3.3. Approximation Using $\sigma_2$-Neural Networks*

In this subsection, we provide approximation results for smoothing functions measured in the $\mathcal{W}^{n,\infty}$ norm, where $n$ is a positive integer, using the smoother $\sigma_2$-NNs. First, we list a few basic lemmas of $\sigma_2$-NNs repeatedly applied in our main analysis.

**Lemma 3.7.** *The following basic lemmas of $\sigma_2$-NNs hold:*

*(i)* *$\sigma_1$-NNs are $\sigma_2$-NNs.*

*(ii)* *Any identity map in $\mathbb{R}^d$ can be realized exactly by a $\sigma_2$-NN with one hidden layer and $2d$ neurons.*

*(iii)* *$f(x) = x^2$ can be realized exactly by $\sigma_2$-NN with one hidden layer and two neurons.*

*(iv)* *$f(x, y) = xy = \frac{(x+y)^2-(x-y)^2}{4}$ can be realized exactly by $\sigma_2$-NN with one hidden layer and four neurons.*

*(v)* *Assume $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ for $\boldsymbol{\alpha} \in \mathbb{N}^d$. For any $N, L \in \mathbb{N}^+$ such that $NL + 2^{\lfloor \log_2 N \rfloor} \geq |\boldsymbol{\alpha}|$, there exists a $\sigma_2$-NN $\phi$ with width $4N + 2d$ and depth $L + \lceil \log_2 N \rceil$ such that*

$$\phi(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}} \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^d.$$

*(vi)* *Assume $P(\boldsymbol{x}) = \sum_{j=1}^{J} c_j \boldsymbol{x}^{\boldsymbol{\alpha}_j}$ for $\boldsymbol{\alpha}_j \in \mathbb{N}^d$. For any $N, L, a, b \in \mathbb{N}^+$ such that $ab \geq J$ and $(L-2b-b\log_2 N)N \geq b \max_j |\boldsymbol{\alpha}_j|$, there exists a $\sigma_2$-NN $\phi$ with width $4Na+2d+2$ and depth $L$ such that*

$$\phi(\boldsymbol{x}) = P(\boldsymbol{x}) \quad \text{for any } \boldsymbol{x} \in \mathbb{R}^d.$$

**Proof** Showing (i) to (iv) is trivial. We will only prove (v) and (vi) in the following.

Part (v): In the case of $|\boldsymbol{\alpha}| = k \leq 1$, the proof is simple and left for the reader. When $|\boldsymbol{\alpha}| = k \geq 2$, the main idea of the proof of (v) can be summarized in Figure 1. We apply $\sigma_1$-NNs to implement a $d$-dimensional identity map as in Lemma 2.1 (iii). These identity maps maintain necessary entries of $\boldsymbol{x}$ to be multiplied together. We apply $\sigma_2$-NNs to implement the multiplication function in Lemma 3.7 (iii) and carry out the multiplication $N$ times per layer. After $L$ layers, there are $k - NL \leq N$ multiplication to be implemented. Finally, these at most $N$ multiplications can be carried out with a small $\sigma_2$-NNs in a dyadic tree structure.

Part (vi): The main idea of the proof is to apply Part (v) $J$ times to construct $J$ $\sigma_2$-NNs, $\{\phi_j(\boldsymbol{x})\}_{j=1}^{J}$, to represent $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$ and arrange these $\sigma_2$-NNs as sub-NN blocks to form a larger $\sigma_2$-NN $\tilde{\phi}(\boldsymbol{x})$ with $ab$ blocks as shown in Figure 2, where each red rectangle represents one $\sigma_2$-NN $\phi_j(\boldsymbol{x})$ and each blue rectangle represents one $\sigma_1$-NN of width 2 as an identity map of $\mathbb{R}$. There are $ab$ red blocks with $a$ rows and $b$ columns. When $ab \geq J$, these sub-NN blocks can carry out all monomials $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$. In each column, the results of the multiplications of $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$ are added up to the input of the narrow $\sigma_1$-NN, which can carry the sum over to the next column. After the calculation of $b$ columns, $J$ additions of the monomials $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$ have been implemented, resulting in the output $P(\boldsymbol{x})$.

By Part (v), for any $N \in \mathbb{N}^+$, there exists a $\sigma_2$-NN $\phi_j(\boldsymbol{x})$ of width $2d + 4N$ and depth $L_j = \lceil \frac{|\boldsymbol{\alpha}_j|}{N} \rceil + \lceil \log_2 N \rceil$ to implement $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$. Note that $b \max_j L_j \leq b \left( \frac{\max_j |\boldsymbol{\alpha}_j|}{N} + 2 + \log_2 N \right)$.

Hence, there exists a $\sigma_2$-NN $\tilde{\phi}(\boldsymbol{x})$ of width $2da+4Na+2$ and depth $b\left(\frac{\max_j|\boldsymbol{\alpha}_j|}{N}+2+\log_2 N\right)$ to implement $P(\boldsymbol{x})$ as in Figure 2. Note that the total width of each column of blocks is $2ad+4Na+2$ but in fact this width can be reduced to $2d+4Na+2$, since the red blocks in each column can share the same identity map of $\mathbb{R}^d$ (the blue part of Figure 1).

Note that $b\left(\frac{\max_j|\boldsymbol{\alpha}_j|}{N}+2+\log_2 N\right)\leq L$ is equivalent to $(L-2b-b\log_2 N)N\geq b\max_j|\boldsymbol{\alpha}_j|$. Hence, for any $N,L,a,b\in\mathbb{N}^+$ such that $ab\geq J$ and $(L-2b-b\log_2 N)N\geq b\max_j|\boldsymbol{\alpha}_j|$, there exists a $\sigma_2$-NN $\phi(\boldsymbol{x})$ with width $4Na+2d+2$ and depth $L$ such that $\tilde{\phi}(\boldsymbol{x})$ is a sub-NN of $\phi(\boldsymbol{x})$ in the sense of $\phi(\boldsymbol{x})=\mathrm{Id}\circ\tilde{\phi}(\boldsymbol{x})$ with Id as an identify map of $\mathbb{R}$, which means that $\phi(\boldsymbol{x})=\tilde{\phi}(\boldsymbol{x})=P(\boldsymbol{x})$. The proof of Part (vi) is completed. □

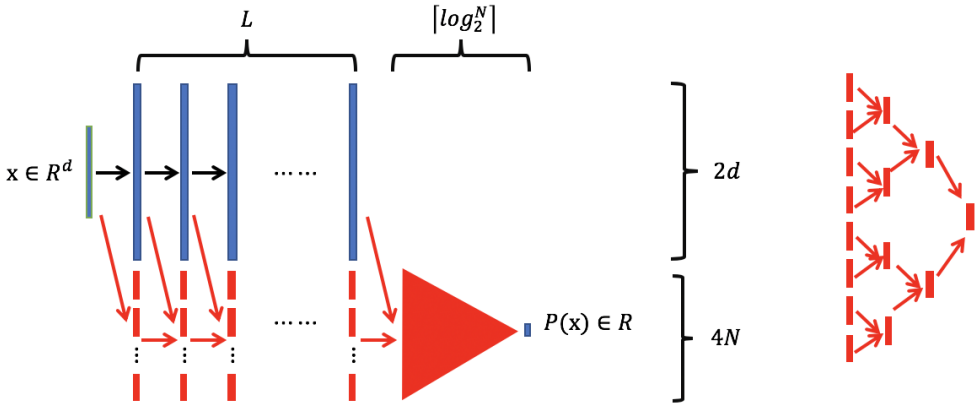


Figure 1: Left: An illustration of the proof of Lemma 3.7 (v). Green vectors represent the input and output of the $\sigma_2$-NN carrying out $P(\boldsymbol{x})$. Blue vectors represent the $\sigma_1$-NN that implements a $d$-dimensional identity map in Lemma 2.1 (iii), which was repeatedly applied for $L$ times. Black arrows represent the data flow for carrying out the identity maps. Red vectors represent the $\sigma_2$-NNs implementing the multiplication function in Lemma 3.7 (iii) and there $NL$ such red vectors. Red arrows represent the data flow for carrying out the multiplications. Finally, a red triangle represent a $\sigma_2$-NN of width at most $4N$ and depth at most $\lceil\log_2^N\rceil$ carrying out the rest of the multiplications. Right: An example of the red triangle is given on the right when it consists of 15 red vectors carrying out 15 multiplications.

Similar to the ReLU network case, we first present the following approximation result before showing our main theorem - Theorem 1.4.

**Theorem 3.8.** *Suppose that $f\in C^s([0,1]^d)$ with $s\in\mathbb{N}^+$ satisfies $\|\partial^{\boldsymbol{\alpha}}f\|_{L^\infty([0,1]^d)}<1$ for any $\boldsymbol{\alpha}\in\mathbb{N}^d$ with $|\boldsymbol{\alpha}|\leq s$. For any $N,L\in\mathbb{N}^+$ satisfying $(L-2-\log_2 N)N\geq s$, there exists a $\sigma_2$-NN $\phi$ with width $16s^{d+1}d(N+2)\log_2(8N)$ and depth $10(L+2)\log_2(4L)$ such that $\|\phi(x)\|_{\mathcal{W}^{n,\infty}([0,1])}\leq s^d$ and*

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

*where $n<s$ is a positive integer, $K=\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor$ and $0\leq\delta\leq\frac{1}{3K}$.*

**Proof** Similar to the proof of Theorem 3.1, we set $K=\lfloor N^{1/d}\rfloor^2\lfloor L^{2/d}\rfloor$ and let $\Omega([0,1]^d,K,\delta,d)$ partition $[0,1]^d$ into $K^d$ cubes $Q_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}\in\{0,1,\cdots,K-1\}^d$ such that

$$[0,1]^d=\Omega([0,1]^d,K,\delta,d)\bigcup\left(\cup_{\boldsymbol{\beta}\in\{0,1,\cdots,K-1\}^d}Q_{\boldsymbol{\beta}}\right).$$

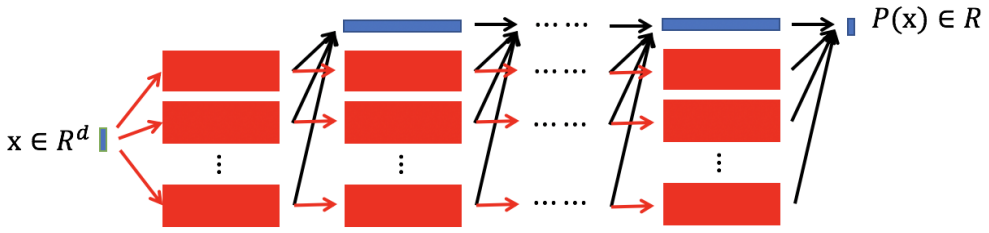


Figure 2: An illustration of the proof of Lemma 3.7 (vi). Green vectors represent the input and output of the $\sigma_2$-NN $\tilde{\phi}(\boldsymbol{x})$ carrying out $P(\boldsymbol{x})$. Each red rectangle represents one $\sigma_2$-NN $\phi_j(\boldsymbol{x})$ and each blue rectangle represents one $\sigma_1$-NN of width 2 as an identity map of $\mathbb{R}$. There are $ab \geq J$ red blocks with $a$ rows and $b$ columns. When $ab \geq J$, these sub-NN blocks can carry out all monomials $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$. In each column, the results of the multiplications of $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$ are added up to (indicated by black arrows) the input of the narrow $\sigma_1$-NN, which can carry the sum over to the next column. Each red arrow passes $\boldsymbol{x}$ to the next red block. After the calculation of $b$ columns, $J$ additions of the monomials $\boldsymbol{x}^{\boldsymbol{\alpha}_j}$ have been implemented, resulting in the output $P(\boldsymbol{x})$.

For each $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_d]^T \in \{0, 1, \cdots, K-1\}^d$, we define

$$Q_{\boldsymbol{\beta}} = \left\{ \boldsymbol{x} = [x_1, x_2, \cdots, x_d]^T : x_i \in \left[ \tfrac{\beta_i}{K}, \tfrac{\beta_i+1}{K} - \delta \cdot 1_{\{\beta_i \leq K-2\}} \right] \text{ for } i = 1, 2, \ldots, d \right\}.$$

By Proposition 2.2, there exists a $\sigma_2$-NN $\psi$ with width $4N+3$ and depth $4L+5$ such that

$$\psi(x) = k, \quad \text{if } x \in \left[ \tfrac{k}{K}, \tfrac{k+1}{K} - \delta \cdot 1_{\{k \leq K-2\}} \right] \text{ for } k = 0, 1, \ldots, K-1.$$

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

Define

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

then

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

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

$$f(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \leq s-1} \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} + \sum_{|\boldsymbol{\alpha}| = s} \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}) + \xi_{\boldsymbol{x}} \boldsymbol{h})}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}},$$

where $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{\psi}(\boldsymbol{x})$.

By Lemma 3.7 (iv), there exists a $\sigma_2$-NN $\widetilde{\phi}$ with width 4 and depth 1 such that

$$\widetilde{\phi}(x, y) = xy$$

for any $x, y \in (-3, 3)$.

Note that it is trivial to construct $\sigma_2$-NNs $P_{\boldsymbol{\alpha}}(\boldsymbol{x})$ for $\boldsymbol{x}^{\boldsymbol{\alpha}}$ when $|\boldsymbol{\alpha}| \leq 1$. Thus, for each $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| \leq s-1$ and for any $N, L \in \mathbb{N}^+$ satisfying $(L - 2 - \log_2 N)N \geq s$, by Lemma 3.7 (vi) with $a = b = J = 1$, there exists a $\sigma_2$-NN $P_{\boldsymbol{\alpha}}$ with width $4N + 2d + 2$ and depth $L$ such that

$$P_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{\alpha}}$$

for any $\boldsymbol{x} \in \mathbb{R}^d$.

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

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

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

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

Note that $K^d = (\lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor)^d \leq N^2 L^2$ and $\xi_{\boldsymbol{\alpha},i} \in [0,1]$ for $i = 0, 1, \cdots, K^d - 1$. By Proposition 2.3, there exists a $\sigma_1$-NN $\widetilde{\phi}_{\boldsymbol{\alpha}}$, which is also a $\sigma_2$-NN, of width $16s(N+1)\log_2(8N)$ and depth $5(L+2)\log_2(4L)$ such that

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

Define

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

For each $|\boldsymbol{\alpha}| \leq s-1$, we know that $\phi_{\boldsymbol{\alpha}}$ is also of width $16s(N+1)\log_2(8N)$ and depth $5(L+2)\log_2(4L)$.

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

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

From $\boldsymbol{\psi}(\boldsymbol{x}) = \frac{\boldsymbol{\beta}}{K}$ for $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$, it follows that

$$\begin{aligned}
\|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} &= \|\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))\|_{L^\infty(Q_{\boldsymbol{\beta}})} \\
&= \big|\phi_{\boldsymbol{\alpha}}(\tfrac{\boldsymbol{\beta}}{K}) - \partial^{\boldsymbol{\alpha}} f(\tfrac{\boldsymbol{\beta}}{K})\big| \\
&\leq 2N^{-2s}L^{-2s} =: \mathcal{E}_3.
\end{aligned}$$

Note that since $\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x})) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))$ for $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$ is constant, its weak derivative is zero, which has given the above inequality.

Define

$$\phi(\boldsymbol{x}) = \sum_{|\boldsymbol{\alpha}| \leq s-1} \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) \tag{10}$$

for any $\boldsymbol{x} \in \mathbb{R}^d$.

Let us now estimate the error for any $\boldsymbol{x} \in Q_{\boldsymbol{\beta}}$.

$$\begin{aligned}
&\|\phi(\boldsymbol{x}) - f(\boldsymbol{x})\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} \\
\leq\ & \underbrace{\sum_{|\boldsymbol{\alpha}|\leq s-1} \left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_1} \\
&+ \underbrace{\sum_{|\boldsymbol{\alpha}|=s} \left\| \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x})+\xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_2}.
\end{aligned}$$

$$\begin{aligned}
E_1 \ =\ & \sum_{|\boldsymbol{\alpha}|\leq s-1} \left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} \\
\leq\ & \underbrace{\sum_{|\boldsymbol{\alpha}|\leq s-1} \left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_{1,1}} \\
&+ \underbrace{\sum_{|\boldsymbol{\alpha}|\leq s-1} \left\| \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} \boldsymbol{h}^{\boldsymbol{\alpha}} \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=:E_{1,2}}.
\end{aligned}$$

Hence, we can now measure $E_{1,1}$, $E_{1,2}$, and $E_2$:

$$\begin{aligned}
E_{1,1} \ =\ & \sum_{|\boldsymbol{\alpha}|\leq s-1} \left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} \\
\leq\ & \sum_{|\boldsymbol{\alpha}|\leq s-1} \Bigg( \underbrace{\left\| \widetilde{\phi}\Big( \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=0} \\
&+ \underbrace{\left\| \widetilde{\phi}\Big( \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=0} \\
&+ \underbrace{\left\| \tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!} P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \right\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{\leq \mathcal{E}_3} \Bigg) \\
\leq\ & \sum_{|\boldsymbol{\alpha}|\leq s-1} \mathcal{E}_3 \\
\leq\ & s^d \mathcal{E}_3.
\end{aligned}$$

Note that the last inequality is followed by the fact that $\sum_{|\boldsymbol{\alpha}|\leq s-1} 1 = \sum_{i=1}^{s-1} (i+1)^{d-1} \leq s^d$.

Similarly,

$$\begin{aligned}
E_{1,2} &= \sum_{|\boldsymbol{\alpha}|\le s-1} \Big\| \widetilde{\phi}\Big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le \sum_{|\boldsymbol{\alpha}|\le s-1} \underbrace{\Big\| \widetilde{\phi}\Big(\tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}P_{\boldsymbol{\alpha}}(\boldsymbol{h}) \Big\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=0} \\
&\quad + \sum_{|\boldsymbol{\alpha}|\le s-1} \underbrace{\Big\| \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}P_{\boldsymbol{\alpha}}(\boldsymbol{h}) - \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})}}_{=0} \\
&\le 0.
\end{aligned}$$

$$\begin{aligned}
E_2 &= \sum_{|\boldsymbol{\alpha}|= s} \Big\| \tfrac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{\psi}(\boldsymbol{x})+\xi_{\boldsymbol{x}}\boldsymbol{h})}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le \sum_{|\boldsymbol{\alpha}|= s} \Big\| \tfrac{1}{\boldsymbol{\alpha}!}\boldsymbol{h}^{\boldsymbol{\alpha}} \Big\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} \\
&\le (s+1)^{d-1}K^{-(s-n)}.
\end{aligned}$$

Note that the last inequality is followed by the fact that $\sum_{|\boldsymbol{\alpha}|=s} 1 \le (s+1)^{d-1}$. Using $K = \lfloor N^{1/d}\rfloor^2 \lfloor L^{2/d}\rfloor \ge \frac{N^{2/d}L^{2/d}}{8}$, we have

$$\begin{aligned}
\|\phi(\boldsymbol{x}) - f(\boldsymbol{x})\|_{\mathcal{W}^{n,\infty}(Q_{\boldsymbol{\beta}})} &\le E_{1,1} + E_{1,2} + E_2 \\
&= s^d\mathcal{E}_3 + (s+1)^{d-1}K^{-(s-n)} \\
&\le (s+1)^d(K^{-(s-n)} + \mathcal{E}_3) \\
&\le (s+1)^d\big(K^{-(s-n)} + 2N^{-2s}L^{-2s}\big) \\
&\le (s+1)^d\big(8^{s-n}N^{-2(s-n)/d}L^{-2(s-n)/d} + 2N^{-2s}L^{-2s}\big) \\
&\le (s+1)^d(8^{s-n}+2)N^{-2(s-n)/d}L^{-2(s-n)/d} \\
&\le 2(s+1)^d 8^{s-n}N^{-2(s-n)/d}L^{-2(s-n)/d}.
\end{aligned}$$

Since $\boldsymbol{\beta} \in \{0,1,2,\cdots,K-1\}^d$ is arbitrary and the fact that $[0,1]^d\backslash\Omega([0,1]^d,K,\delta,d) \subseteq \cup_{\boldsymbol{\beta}\in\{0,1,\cdots,K-1\}^d} Q_{\boldsymbol{\beta}}$, we have

$$\|\phi(\boldsymbol{x}) - f(\boldsymbol{x})\|_{\mathcal{W}^{n,\infty}([0,1]^d\backslash\Omega([0,1]^d,K,\delta,d))} \le 2(s+1)^d 8^{s-n}N^{-2(s-n)/d}L^{-2(s-n)/d}.$$

Furthermore, we have

$$\begin{aligned}
\|\phi(\boldsymbol{x})\|_{\mathcal{W}^{1,\infty}([0,1]^d)} &= \Big\| \sum_{|\boldsymbol{\alpha}|\le s-1} \widetilde{\phi}\Big(\tfrac{\phi_{\boldsymbol{\alpha}}(\boldsymbol{\psi}(\boldsymbol{x}))}{\boldsymbol{\alpha}!}, P_{\boldsymbol{\alpha}}(\boldsymbol{h})\Big) \Big\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&\le \sum_{|\boldsymbol{\alpha}|\le s-1} \|\widetilde{\phi}\|_{\mathcal{W}^{1,\infty}([0,1]^d)} \\
&\le s^d.
\end{aligned}$$

As last, we finish the proof by estimating the width and depth of the network implementing $\phi(\boldsymbol{x})$. From (10), assuming for any $N, L \in \mathbb{N}^+$ satisfying $(L-2-\log_2 N)N \geq s$, we know that $\phi(\boldsymbol{x})$ consists of the following subnetworks:

1. $\boldsymbol{\psi} \in \mathrm{NN}(\text{width} \leq d(4N+3); \text{depth} \leq 4L+5)$;
2. $\phi_{\boldsymbol{\alpha}} \in \mathrm{NN}(\text{width} \leq 16s(N+1)\log_2(8N); \text{depth} \leq 5(L+2)\log_2(4L))$;
3. $P_{\boldsymbol{\alpha}} \in \mathrm{NN}(\text{width} \leq 4N+2d+2; \text{depth} \leq L)$;
4. $\widetilde{\phi} \in \mathrm{NN}(\text{width} \leq 4; \text{depth} \leq 1)$.

Thus, we can infer that the $\phi$ can be implemented by a ReLU network with width $16s^{d+1}d(N+2)\log_2(8N)$ and depth

$$(4L+5)+1+L+5(L+2)\log_2(4L)+3 \leq 10(L+2)\log_2(4L).$$

Hence, we have finished the proof. □

We can now prove Theorem 1.4 using Theorem 3.8.

**Proof of Theorem 1.4** When $f$ is a constant function, the statement is trivial. By Theorem 3.8, there exists a $\sigma_2$-NN $\phi$ with width $16s^{d+1}d(N+2)\log_2(8N)$ and depth $10(L+2)\log_2(4L)$ such that $\|\phi\|_{\mathcal{W}^{n,\infty}([0,1]^d)} \leq s^d$ and

$$\|f-\phi\|_{\mathcal{W}^{n,p}([0,1]^d)} \leq 2(s+1)^d 8^{s-n} N^{-2(s-n)/d} L^{-2(s-n)/d},$$

Now, we set $K = \lfloor N^{1/d} \rfloor^2 \lfloor L^{2/d} \rfloor$ and choose a small $\delta$ such that

$$Kd\delta \leq \left(N^{-2(s-n)/d} L^{-2(s-n)/d}\right)^p.$$

Then, we have

$$\begin{aligned}
\|f-\phi\|^p_{\mathcal{W}^{n,p}([0,1]^d)} &= \|f-\phi\|^p_{\mathcal{W}^{n,p}(\Omega([0,1]^d,K,\delta,d))} + \|f-\phi\|^p_{\mathcal{W}^{n,p}([0,1]^d\setminus\Omega([0,1]^d,K,\delta,d))} \\
&\leq Kd\delta(s^d)^p + \left(2(s+1)^d 8^{s-n} N^{-2(s-n)/d} L^{-2(s-n)/d}\right)^p \\
&\leq \left(s^d N^{-2(s-1)/d} L^{-2(s-1)/d}\right)^p + \left(2(s+1)^d 8^{s-n} N^{-2(s-n)/d} L^{-2(s-n)/d}\right)^p \\
&\leq \left(3(s+1)^d 8^{s-n} N^{-2(s-n)/d} L^{-2(s-n)/d}\right)^p.
\end{aligned}$$

Hence, we have

$$\|f-\phi\|_{\mathcal{W}^{1,p}([0,1]^d)} \leq 3(s+1)^d 8^{s-n} N^{-2(s-n)/d} L^{-2(s-n)/d}.$$

□

## 4. Extension to approximation for functions in the Hölder space

In this section, we indicate that our approximation results derived for smooth functions can be further generalized to functions in the Hölder space. The latter is a more suitable function space in the context of solving PDEs since the derivatives of a typical solution to an equation in its weak formulation are often not as smooth.

Let $\beta = s + r > 0, r \in (0, 1]$ and $s = \lfloor \beta \rfloor \in \mathbb{N}_0$, where $\mathbb{N}_0$ denotes the set of nonnegative integers. For a finite constant $B_0 > 0$, the Hölder space of functions $\mathcal{H}^\beta([0,1]^d, B_0)$ is defined by

$$\mathcal{H}^\beta([0,1]^d, B_0) = \Big\{ f : [0,1]^d \to \mathbb{R}, \max_{\|\alpha\|_1 \le s} \|\partial^\alpha f\|_\infty \le B_0, \max_{\|\alpha\|_1 = s} \sup_{x \ne y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x-y\|_2^r} \le B_0 \Big\}.$$

Following similar augments presented in Section 3 where Taylor polynomials are used as approximants, we can first construct ReLU/ReLU$^2$ subnetworks to realize these polynomials and hence implement the desired target functions in the Hölder space by composing the subnetworks.

We have the following results with proofs left for the reader:

**Theorem 4.1.** *Suppose that $f \in \mathcal{H}^\beta([0,1]^d, B_0)$ with $\beta = s + r, r \in (0, 1]$ and $s > 1$. For any $N, L \in \mathbb{N}^+$ and $p \in [1, \infty)$, there exists a $\sigma_1$-NN $\phi$ with width $C_1(B_0, \beta, d) N \log_2(N)$ and depth $C_2(B_0, \beta, d) L \log_2(L)$ such that*

$$\|f - \phi\|_{\mathcal{W}^{1,p}([0,1]^d)} \le C_3(B_0, \beta, d) N^{-2(\beta-1)/d} L^{-2(\beta-1)/d},$$

*where $C_1, C_2, C_3 > 0$ are constants depending on $B_0, \beta, d$ which can be explicitly determined.*

**Theorem 4.2.** *Suppose that $f \in \mathcal{H}^\beta([0,1]^d, B_0)$ with $\beta = s + r, r \in (0, 1]$ and $s \in \mathbb{N}_0$. For any $N, L \in \mathbb{N}^+$ satisfying $(L - 2 - \log_2 N)N \ge s$, and $p \in [1, \infty)$, there exists a $\sigma_2$-NN $\phi$ with width $C_1(B_0, \beta, d) N \log_2(N)$ and depth $C_2(B_0, \beta, d) L \log_2(L)$ such that*

$$\|f - \phi\|_{\mathcal{W}^{n,p}([0,1]^d)} \le C_3(B_0, \beta, d) N^{-2(\beta-n)/d} L^{-2(\beta-n)/d},$$

*where $n < \beta$ is a positive integer and $C_1, C_2, C_3 > 0$ are constants depending on $B_0, \beta, d$ which can be explicitly determined.*

## 5. Conclusions

We have given a number of theoretical results on explicit error characterization for approximating smooth functions using deep ReLU networks and their smoother variants. Our results measured in Sobolev norms are well-suited for studying solving high-dimensional PDEs. Further generalizing our analysis to other neutral networks such as the Floor-ReLU networks will be an interesting direction for research work. Numerical investigation of our findings in the setting for solving parametric PDEs will also be left as future work.

**Acknowledgements**

S. Hon was partially supported by the Hong Kong RGC under Grant 22300921, a start-up allowance from the Croucher Foundation, and a Tier 2 Start-up Grant from the Hong Kong Baptist University. H. Yang was partially supported by the US National Science Foundation under award DMS-1945029.

## References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A Convergence Theory for Deep Learning via Over-Parameterization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 242–252. PMLR, 2019.

[2] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 322–332. PMLR, 2019.

[3] Julius Berner, Philipp Grohs, and Arnulf Jentzen. Analysis of the Generalization Error: Empirical Risk Minimization over Deep Artificial Neural Networks Overcomes the Curse of Dimensionality in the Numerical Approximation of Black–Scholes Partial Differential Equations. *SIAM Journal on Mathematics of Data Science*, 2(3):631–657, 2020.

[4] Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal Approximation with Sparsely Connected Deep Neural Networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

[5] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[6] Fan Chen, Jianguo Huang, Chunmei Wang, and Haizhao Yang. Friedrichs Learning: Weak Solutions of Partial Differential Equations via Deep Learning. *arXiv e-prints*, page arXiv:2012.08023, 2020.

[7] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear Approximation and (Deep) ReLU Networks. *Constructive Approximation*, 2021.

[8] Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

[9] Ronald A. DeVore, Ralph Howard, and Charles Micchelli. Optimal nonlinear approximation. *manuscripta mathematica*, 63:469–478, 1989.

[10] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient Descent Finds Global Minima of Deep Neural Networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685. PMLR, 2019.

[11] Simon S. Du, Barnabás Poczós, Xiyu Zhai, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–19, 2019.

[12] Weinan E. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

[13] Weinan E, Jiequn Han, and Arnulf Jentzen. Deep Learning-Based Numerical Methods for High-Dimensional Parabolic Partial Differential Equations and Backward Stochastic Differential Equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

[14] Weinan E, Chao Ma, and Lei Wu. The Barron Space and the Flow-Induced Function Spaces for Neural Network Models. *Constructive Approximation*, 2021.

[15] Dennis Elbrächter, Philipp Grohs, Arnulf Jentzen, and Christoph Schwab. DNN Expression Rate Analysis of High-Dimensional PDEs: Application to Option Pricing. *Constructive Approximation*, 2021.

[16] Moritz Geist, Philipp Petersen, Mones Raslan, Reinhold Schneider, and Gitta Kutyniok. Numerical Solution of the Parametric Diffusion Equation by Deep Neural Networks. *Journal of Scientific Computing*, 88(1):22, 2021.

[17] Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, and Felix Voigtlaender. Approximation Spaces of Deep Neural Networks. *Constructive Approximation*, 2021.

[18] Yiqi Gu, Haizhao Yang, and Chao Zhou. SelectNet: Self-paced learning for high-dimensional partial differential equations. *Journal of Computational Physics*, 441:110444, 2021.

[19] 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(5):803–859, 2020.

[20] Ingo Gühring and Mones Raslan. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021.

[21] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[22] John Harlim, Shixiao W Jiang, Senwei Liang, and Haizhao Yang. Machine learning for prediction with missing dynamics. *Journal of Computational Physics*, 428:109922, 2021.

[23] Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. ReLU Deep Neural Networks and Linear Finite Elements. *Journal of Computational Mathematics*, 38(3):502–527, 2020.

[24] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 8571–8580, 2018.

[25] Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. Solving parametric PDE problems with artificial neural networks. *European Journal of Applied Mathematics*, 32(3):421–435, 2021.

[26] Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A Theoretical Analysis of Deep Neural Networks and Parametric PDEs. *Constructive Approximation*, 2021.

[27] Senwei Liang, Shixiao W Jiang, John Harlim, and Haizhao Yang. Solving PDEs on Unknown Manifolds with Machine Learning. *arXiv e-prints*, page arXiv:2106.06682, 2021.

[28] Zichao Long, Yiping Lu, and Bin Dong. PDE-Net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019.

[29] Jianfeng Lu and Yulong Lu. A Priori Generalization Error Analysis of Two-Layer Neural Networks for Solving High Dimensional Schrödinger Eigenvalue Problems. *arXiv e-prints*, page arXiv:2105.01228, 2021.

[30] Jianfeng Lu, Yulong Lu, and Min Wang. A Priori Generalization Analysis of the Deep Ritz Method for Solving High Dimensional Elliptic Equations. *Proceedings of Thirty Fourth Conference on Learning Theory*, PMLR, 134:3196-3241, 2021.

[31] 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.

[32] Tao Luo and Haizhao Yang. Two-Layer Neural Networks for Partial Differential Equations: Optimization and Generalization Theory, 2020.

[33] Chao Ma, Jianchun Wang, and Weinan E. Model Reduction with Memory and the Machine Learning of Dynamical Systems. *Communications in Computational Physics*, 25(4):947–962, 2019.

[34] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 2019.

[35] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665—-E7671, 2018.

[36] Hadrien Montanelli and Qiang Du. New Error Bounds for Deep ReLU Networks Using Sparse Grids. *SIAM Journal on Mathematics of Data Science*, 1(1):78–92, 2019.

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

[38] Joost A A Opschoor, Philipp C Petersen, and Christoph Schwab. Deep ReLU networks and high-order finite element methods. *Analysis and Applications*, 18(05):715–770, 2020.

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

[40] Tong Qin, Kailiang Wu, and Dongbin Xiu. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, 395:620–635, 2019.

[41] M Raissi, P Perdikaris, and G E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[42] Christoph Schwab and Jakob Zech. Deep learning in high dimension: Neural network expression rates for generalized polynomial chaos expansions in UQ. *Analysis and Applications*, 17(1), 2019.

[43] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

[44] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

[45] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep Network Approximation: Achieving Arbitrary Accuracy with Fixed Number of Neurons. *arXiv e-prints*, page arXiv:2107.02397, 2021.

[46] Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 141:160–173, 2021.

[47] Jonathan W Siegel and Jinchao Xu. Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321, 2020.

[48] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical Insights Into the Optimization Landscape of Over-Parameterized Shallow Neural Networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019.

[49] E. Weinan, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425, 2019.

[50] E. Weinan and Qingcan Wang. Exponential convergence of the deep neural network approximation for analytic functions. *Science China Mathematics*, 61(10):1733–1740, 2018.

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

[52] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.*, 120(14):143001, 2018.

[53] Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery Guarantees for One-hidden-layer Neural Networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4140–4149. PMLR, 2017.