Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2023-0264

Vol. **36**, No. 1, pp. 71-103
July 2024

Convergence Analysis for Over-Parameterized Deep Learning

Yuling Jiao[1], Xiliang Lu[1], Peiying Wu[1] and Jerry Zhijian Yang[1,*]

[1] *School of Mathematics and Statistics, Wuhan University, Wuhan 430072, P.R. China.*

Received 7 October 2023; Accepted (in revised version) 22 April 2024

Abstract. The success of deep learning in various applications has generated a growing interest in understanding its theoretical foundations. This paper presents a theoretical framework that explains why over-parameterized neural networks can perform well. Our analysis begins from the perspective of approximation theory and argues that over-parameterized deep neural networks with bounded norms can effectively approximate the target. Additionally, we demonstrate that the metric entropy of such networks is independent of the number of network parameters. We utilize these findings to derive consistency results for over-parameterized deep regression and the deep Ritz method, respectively. Furthermore, we prove convergence rates when the target has higher regularity, which, to our knowledge, represents the first convergence rate for over-parameterized deep learning.

AMS subject classifications: 65M15, 65N15, 65Y20

Key words: Over-parameterization, convergence rate, approximation, generalization.

1 Introduction

The success of deep learning in various applications has spurred a growing interest in understanding its theoretical foundations. One of the most crucial questions is why over-parameterized neural networks can perform well. The current literature [40] suggests that the generalization error of neural networks generally increases with the increasing complexity of the network function space, making it theoretically difficult for over-parameterized neural networks to converge in terms of generalization error. However, in practice, training over-parameterized deep neural networks is widely used since it makes model training more computationally convenient. Moreover, recent studies have shown

*Corresponding author. *Email addresses:* `yulingjiaomath@whu.edu.cn` (Y. Jiao), `xllv.math@whu.edu.cn` (X. Lu), `peiyingwu@whu.edu.cn` (P. Wu), `zjyang.math@whu.edu.cn` (J. Z. Yang)

that (stochastic) gradient descent with randomized initialization and small step-size converges linearly in over-parameterized regimes, even though the optimization problem in deep learning is highly non-convex, see [2,12,13,25,34,55] and the references therein. All of these indicate a conflict between existing theory and practice, and a new perspective is urgently needed to resolve this dilemma.

To address this dilemma, significant effort has been devoted to developing over-parameterized deep learning theory [4,6,10]. Belkin et al. proposed the double descent curve in [6] to describe the limitations of classical analysis, but did not provide explanations. Currently, the main perspective on understanding over-parameterization for linear and kernel models is benign overfitting due to the double descent phenomenon for estimators interpolating data with minimum norm [3,4,6–9,33,43,50]. However, [29] provides a negative result that the empirical risk minimization estimator can be inconsistent in nonparametric least squares regression with over-parameterized deep neural networks. In this work, we introduce a new theoretical framework based on function space theory and establish the consistency of norm-bounded over-parameterized deep learning. We demonstrate that the complexity of a neural network can be controlled by the metric entropy of the balls in certain metric space, which is independent of the number of parameters. This provides a novel perspective for understanding the good generalization ability of over-parameterized neural networks. We illustrate our approach with two representative examples: the regression model and the deep Ritz method. The main contributions of this work are summarized as follows.

- We establish a new bound for the approximation error of *ReLU* deep neural networks in the Sobolev space, which may be of independent interest.

- We provide a unified consistency analysis of over-parameterized regression models and deep Ritz methods, which offers a novel perspective for understanding over-parameterized deep learning.

- Our framework is applicable to various activation functions, including *ReLU* and *Sigmoidal* functions. By exploring the smoothness of the target and network, we drive improved convergence rate.

The paper is organized as follows. In Section 2, we give some notations and mathematical background used in this paper. Section 3 provides a brief overview of our main results. In Section 4, we present our proof framework. In Section 5, we summarize our findings and conclude the paper. Some technical detailed proofs are given in Section A.

2 Notations and background

In this section, we provide all the notations we need in this paper. For $k\in\mathbb{R}$, we define $\mathbb{R}_{>k}:=\{x\in\mathbb{R}|x>k\}$ and $\mathbb{N}_0:=\{x\in\mathbb{N}|x\geq 0\}$. If $x\in\mathbb{R}$, $\lfloor x\rfloor:=\max\{k\in\mathbb{Z}:k\leq x\}$ denotes the largest integer strictly smaller than x. $\mathfrak{C}\in\mathbb{R}$ is a positive constant number, and $\mathfrak{C}(d)$

is a polynomial that depends on d. For two function spaces A and B, $A \circlearrowleft B$ means A is embedded into B. We use the usual *multiindex* notation, i.e. for $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ we write $\|\boldsymbol{\alpha}\|_1 := \alpha_1 + \cdots + \alpha_d$ and $\boldsymbol{\alpha}! := \alpha_1! \cdot \cdots \cdot \alpha_d!$.

Let $\Omega \subset \mathbb{R}^d$ be some open set. For a function $f : \Omega \to \mathbb{R}$, we denote by

$$D^{\boldsymbol{\alpha}} f := \frac{\partial^{\|\boldsymbol{\alpha}\|_1} f}{\partial {x_1}^{\alpha_1} \partial {x_2}^{\alpha_2} \cdots \partial {x_d}^{\alpha_d}}$$

its (weak or classical) derivative of order $\boldsymbol{\alpha}$. For $s \in \mathbb{N}_0 \cup \{\infty\}$, we denote by $C^s(\Omega)$ the set of s times continuously differentiable functions on Ω. Additionally, if $\overline{\Omega}$ is compact, we set, for $f \in C^s(\Omega)$

$$\|f\|_{C^s(\overline{\Omega})} := \max_{0 \le \|\boldsymbol{\alpha}\|_1 \le s} \sup_{\boldsymbol{x} \in \Omega} |D^{\boldsymbol{\alpha}} f(\boldsymbol{x})|.$$

2.1 Sobolev space

For any $s \in \mathbb{N}_0$ and $1 \le p < \infty$, we define the *Sobolev space* $W^{s,p}(\Omega)$ by

$$W^{s,p}(\Omega) := \left\{ f \in L^p(\Omega) : D^{\boldsymbol{\alpha}} f \in L^p(\Omega),\ \forall \boldsymbol{\alpha} \in \mathbb{N}_0^d \text{ with } \|\boldsymbol{\alpha}\|_1 \le s \right\}.$$

In particular, when $p = 2$, we define $H^s(\Omega) := W^{s,2}(\Omega)$ for any $s \in \mathbb{N}_0$. Moreover, for any $f \in W^{s,p}(\Omega)$ with $1 \le p < \infty$, we define the Sobolev norm by

$$\|f\|_{W^{s,p}(\Omega)} := \left(\sum_{0 \le \|\boldsymbol{\alpha}\|_1 \le s} \|D^{\boldsymbol{\alpha}} f\|_{L^p(\Omega)}^p \right)^{\frac{1}{p}}.$$

When $p = \infty$, we have

$$\|f\|_{W^{s,\infty}(\Omega)} := \max_{0 \le \|\boldsymbol{\alpha}\|_1 \le s} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)}.$$

2.2 Hölder space

Let $\beta = s + r > 0$, $r \in (0,1]$ and $s = \lfloor \beta \rfloor \in \mathbb{N}_0$. For a finite constant $B > 0$, the bounded *Hölder class* of functions $\mathcal{H}^\beta(\Omega, B)$ is defined as

$$\mathcal{H}^\beta(\Omega, B) = \left\{ f : \Omega \to \mathbb{R}, \max_{\|\boldsymbol{\alpha}\|_1 \le s} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty} \le B,\ \max_{\|\boldsymbol{\alpha}\|_1 = s} \sup_{\boldsymbol{x} \ne \boldsymbol{y}} \frac{|D^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^r} \le B \right\}, \tag{2.1}$$

where the norm $\|f\|_{\mathcal{H}^\beta(\Omega)}$ is defined as

$$\sum_{\|\boldsymbol{\alpha}\|_1 \le s} \|f\|_{C^s(\Omega)} + \sum_{\|\boldsymbol{\alpha}\|_1 = s} \sup_{\boldsymbol{x} \ne \boldsymbol{y}} \frac{|D^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^r}.$$

For any $f \in \mathcal{H}^\beta(\Omega, B)$, all partial derivatives of f up to the $\lfloor \beta \rfloor$-th order exist.

2.3 Neural network

Let $W,L,d,\mathfrak{n}_{\boldsymbol{\theta}}\in\mathbb{N}$, $M\in\mathbb{R}$ and $M>0$, $d>2$. We consider the function $f:\mathbb{R}^d\to\mathbb{R}$ that can be parameterized by a neural network of the form

$$\begin{aligned}
&f_0(\boldsymbol{x})=\boldsymbol{x},\\
&f_\ell(\boldsymbol{x})=\rho(\boldsymbol{A}_\ell f_{\ell-1}(\boldsymbol{x})+\boldsymbol{b}_\ell),\quad \ell=1,\cdots,L-1,\\
&f(\boldsymbol{x})=f_L(\boldsymbol{x})=\boldsymbol{A}_L f_{L-1}(\boldsymbol{x})+\boldsymbol{b}_L.
\end{aligned} \tag{2.2}$$

The network weights A_ℓ are defined as $A_\ell:\mathbb{R}^{N_{\ell-1}}\to\mathbb{R}^{N_\ell}$, where $N_0=d$, $N_L=1$, and N_ℓ represents the width of the ℓ-th layer of the network. The numbers W and L represent the largest width and depth of the network, respectively. $\mathfrak{n}_{\boldsymbol{\theta}}$ denotes the total number of nonzero weights, while ρ represents the activation function. In this paper, we focus on two specific activation functions: *ReLU* and *Sigmoidal*. When the activation function ρ is clear, we use the notation $\mathcal{NN}(W,L)$ to represent a neural network with a width of W, depth of L, and $\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}})$ to represent a neural network with a depth of L and total number of non-zero weights $\mathfrak{n}_{\boldsymbol{\theta}}$.

We introduce the concept of a *norm bounded neural network*, which is defined as a network whose output function $f_{\boldsymbol{\theta}}$ is constrained by a value of M using a specific norm $|\cdot|$. This type of network is denoted as $\mathcal{NN}(W,L,|\cdot|,M)$ or $\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},|\cdot|,M)$. The choice of norm used in the constraint depends on the desired smoothness of the function space. Throughout the paper, we denote $\mathcal{P}$ as a function class consisting of feedforward neural networks parameterized by $\boldsymbol{\theta}:=((\boldsymbol{A}_1,\boldsymbol{b}_1),\cdots,(\boldsymbol{A}_L,\boldsymbol{b}_L))$.

3 Main result

We provide convergence analysis for deep regression model and deep Ritz method [53] for solving elliptic partial differential equation. These two problems are not only representative model problems in deep learning but also of great practical interest in their own right. We analyze *ReLU* and *Sigmoidal* activation functions, which are the two most common types.

3.1 Regression model

Consider a nonparametric regression model

$$Y=f_0(X)+\xi, \tag{3.1}$$

where $Y\in\mathbb{R}$ is a response, $X\in\mathbb{R}^d$ is a d-dimensional vector of predictors, $f_0:[0,1]^d\to\mathbb{R}$ is an unknown regression function, ξ is an error with mean 0 and finite variance V^2, independent of X. A basic problem in statistics and machine learning is to estimate the unknown target regression function f_0 based on a random sample, $S_n=\{(X_i,Y_i)\}_{i=1}^n\subseteq$

$[0,1]^d \times \mathbb{R}$, where n is the sample size, that are independent and identically distributed (i.i.d.) as (X,Y).

A basic paradigm for estimating f_0 is to minimize the mean square error or the L^2 risk. For any (random) function f, let $Z \equiv (X,Y)$ be a random vector, the least-squares estimation is to find a measurable function $f_0 : \mathbb{R}^d \to \mathbb{R}$ satisfying

$$f_0 := \underset{f}{\operatorname{argmin}} \mathcal{L}(f) = \underset{f}{\operatorname{argmin}} \underset{Z}{\mathbb{E}} |Y - f(X)|^2. \tag{3.2}$$

However, in applications, the distribution of (X,Y) is typically unknown and only a random sample $S_n = \{(X_i,Y_i)\}_{i=1}^n$ is available. Let

$$\widehat{\mathcal{L}}(f) := \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - Y_i|^2, \tag{3.3}$$

be the empirical risk of f on the sample S_n. Based on the observed random sample, our primary goal is to construct an estimators of f_0 within a certain class of functions $\mathcal{P}$ by minimizing the empirical risk. Then we have the following convergence rate theorem.

Theorem 3.1 (Informal version of Theorem 4.4)**.** *Let $B > 0$, $d > 2$, $\beta \geq 2$ and $k \geq \mathfrak{C}(B,d,\beta)$. If $f_0 \in \mathcal{H}^\beta([0,1]^d, B)$ is the target function of a regression model, there exists an over-parameterized ReLU network class $\mathcal{P} = \mathcal{NN}(W,L,\|\cdot\|_{W^{1,\infty}},2B)$ with $W,L \geq \mathfrak{C}(n,d,\beta,k,B)$, such that for the empirical risk minimizer $\widehat{f}_{\boldsymbol{\theta}} = \operatorname{argmin}_{f\in\mathcal{P}} \widehat{\mathcal{L}}(f)$,*

$$\underset{S_n}{\mathbb{E}}\left[\|\widehat{f}_{\boldsymbol{\theta}} - f_0\|^2_{L^2(\mu)}\right] \leq \mathfrak{C}(\beta,d,B) n^{-\frac{1}{d}}.$$

In contrast to existing convergence results where the number of parameters of neural networks is required to be smaller than the number of training samples [5,16,17,27,30,42, 44,45,48,49], Theorem 3.1 holds for any sufficiently large width, i.e., Theorem 3.1 applies to over-parameterization schemes.The proof will be given in Section 4.1.

3.2 Deep Ritz method

Since the *ReLU* function is not smooth, it may not be suitable for problems that require high smoothness in the solution. In this section, we present a convergence analysis of the over-parameterized Deep Ritz Method [53] for second-order elliptic equations with Neumann boundary conditions. To ensure sufficient smoothness, we use the *Sigmoidal* activation function, which is infinitely continuously differentiable.

Let $[0,1]^d$ be the unit hypercube on $\mathbb{R}^d$, $\Omega \subset [0,1]^d$ be a convex bounded open set and $\partial\Omega$ be the boundary of Ω. Consider the elliptic equation on Ω equipped with Neumann boundary condition:

$$-\Delta u + wu = f \quad \text{on } \Omega, \quad \frac{\partial u}{\partial \boldsymbol{n}} = g \quad \text{on } \partial\Omega. \tag{3.4}$$

According to the variation method and the integration by parts formula, the energy functional can be defined by

$$\mathcal{L}(u)=\int_\Omega \left(\frac{1}{2}|\nabla u|^2+\frac{1}{2}w|u|^2-fu\right)-\int_{\partial\Omega}(gTu)ds, \tag{3.5}$$

where T is the trace operator. We use Monte Carlo method to discretize the energy functional and rewrite (3.5) as

$$\mathcal{L}(u)=|\Omega| \underset{X\sim U(\Omega)}{\mathbb{E}}\left[\frac{\|\nabla u(X)\|_2^2}{2}+\frac{w(X)u^2(X)}{2}-u(X)f(X)\right]-|\partial\Omega| \underset{Y\sim U(\partial\Omega)}{\mathbb{E}}[Tu(Y)g(Y)], \tag{3.6}$$

where $U(\Omega)$, $U(\partial\Omega)$ are the uniform distribution on Ω and $\partial\Omega$.

We now introduce the discrete version:

$$\widehat{\mathcal{L}}(u)=\frac{|\Omega|}{N_{in}}\sum_{k=1}^{N_{in}}\left[\frac{\|\nabla u(X_k)\|_2^2}{2}+\frac{w(X_k)u^2(X_k)}{2}-u(X_k)f(X_k)\right]-\frac{|\partial\Omega|}{N_b}\sum_{k=1}^{N_b}[u(Y_k)g(Y_k)], \tag{3.7}$$

where $\{X_k\}_{k=1}^{N_{in}}\sim U(\Omega)$, $\{Y_k\}_{k=1}^{N_b}\sim U(\partial\Omega)$ i.i.d.. For the deep Ritz method, we have the following convergence rate theorem.

Theorem 3.2 (Informal version of Theorem 4.8)**.** *Let $B>0$, $p>d>2$, $N_{in}=N_b=n$ and $p,d,n\in\mathbb{N}_0$. Suppose that $u^*\in W^{4,p}(\Omega)$ is the target solution of the elliptic partial differential equation, and it satisfies $\|u^*\|_{W^{4,p}(\Omega)}\le B$. Then, there exists an over-parameterized Sigmoidal network function class $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^2},2B)$ with at least $\mathfrak{n}_{\boldsymbol{\theta}}\ge\mathfrak{C}(d,B,p,n)$ non-zero weights, such that for the empirical risk minimizer $\widehat{u}_{\boldsymbol{\theta}}=\mathrm{argmin}_{u\in\mathcal{P}}\widehat{\mathcal{L}}(u)$,*

$$\underset{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}{\mathbb{E}}[\|\widehat{u}_{\boldsymbol{\theta}}-u^*\|_{H^1(\Omega)}^2]\le\mathfrak{C}(d,B,M)n^{-\frac{1}{d}}.$$

There have been several works attempting to explain the mechanisms of DRM and PINNs from a mathematical perspective [14, 15, 23, 24, 26, 31, 32, 35–39, 41, 46, 47, 51, 52]. However, all the studies conducted so far have assumed a scenario where the number of neural network parameters is less than the number of training samples. Theorem 3.2 demonstrates that over-parameterized *Sigmoidal* neural networks can achieve a convergence rate of $n^{-\frac{1}{d}}$ in the H^1 norm in solving elliptic partial differential equations with Neumann boundary conditions.

3.3 The smoothness of network

We also discuss the impact of the smoothness index on the network's convergence and demonstrate that higher smoothness leads to better convergence. To further explore this topic, we revisit the regression model. Since ReLU networks only have first-order smoothness, we can impose stronger assumptions by replacing the activation function with a Sigmoidal activation function. This allows us to investigate how the smoothness of the network function class affects the convergence rate.

Theorem 3.3 (Informal version of Theorem 4.10). *Let $B>0$, $s=\lfloor\beta\rfloor$, $1\leq t\leq\eta\leq s-1$, $t<d/2$ and $t,\eta,d\in\mathbb{N}_0$. $S_n=\{(X_i,Y_i)\}_{i=1}^n$ is a random sample set, $f_0\in\mathcal{H}^\beta([0,1]^d,B)$ is the target function of a regression model. For each t, there exists an over-parameterized Sigmoidal network function class $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_\theta,\|\cdot\|_{C^t},2B)$ with $\mathfrak{n}_\theta\geq\mathfrak{C}(d,B,s,n,\eta,t)$, such that for the empirical risk minimizer $\widehat{f}_\theta=\mathrm{argmin}_{f\in\mathcal{P}}\widehat{\mathcal{L}}(f)$,*

$$\mathop{\mathbb{E}}_{S_n}\left[\|\widehat{f}_\theta-f_0\|_{L^2(\mu)}^2\right]\leq\mathfrak{C}(s,d,B,t,\eta)n^{-\frac{t}{d}}.$$

From Theorem 3.3, we observe that the upper bound of excess risk is inversely proportional to the smoothness index of the *Sigmoidal* neural network t. Therefore, the higher the smoothness, the better the convergence.

4 Proof sketch

This section presents the proof sketches for Theorems 3.1, 3.2, and 3.3, with each proof consisting of four steps.

4.1 Proof sketch of Theorem 3.1

Assumption 4.1. Assume that the target function f_0 belongs to $\mathcal{H}^\beta\left([0,1]^d,B\right)$ defined in (2.1) for a given $\beta\geq 2$ and a finite constant $B>0$.

For any estimator $\widehat{f}_\theta$, we evaluate its quality via its excess risk, defined as the difference between the L_2 risks of $\widehat{f}_\theta$ and f_0, then

$$\mathcal{L}(\widehat{f}_\theta)-\mathcal{L}(f_0)=\mathop{\mathbb{E}}_{Z}|Y-\widehat{f}_\theta(X)|^2-\mathop{\mathbb{E}}_{Z}|Y-f_0(X)|^2=\mathop{\mathbb{E}}_{X}|\widehat{f}_\theta(X)-f_0(X)|^2=\|\widehat{f}_\theta-f_0\|_{L^2(\mu)}^2, \tag{4.1}$$

where μ denotes the marginal distribution of X. A good estimator $\widehat{f}_\theta$ should have a small excess risk $\|\widehat{f}_\theta-f_0\|_{L^2(\mu)}^2$. Thereafter, we focus on deriving the non-asymptotic upper bounds of the expected excess risk $\mathop{\mathbb{E}}_{S_n}\|\widehat{f}_\theta-f_0\|_{L^2(\mu)}^2$.

Step 1. Error decomposition. We decompose the expected excess risk into approximation error and statistical error and analyze them separately.

Proposition 4.1. Suppose that $\mathcal{P}=\mathcal{NN}(W,L)$ (or $\mathcal{NN}(L,\mathfrak{n}_\theta)$), then

$$\mathop{\mathbb{E}}_{S_n}\left[\|\widehat{f}_\theta-f_0\|_{L^2(\mu)}^2\right]\leq\underbrace{\inf_{\bar{f}\in\mathcal{P}}\|\bar{f}-f_0\|_{L^2(\mu)}^2}_{\mathcal{E}_{app}}+\underbrace{\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]}_{\mathcal{E}_{sta}}. \tag{4.2}$$

The approximation error $\mathcal{E}_{app}$ describes the expressive power of the network class $\mathcal{P}$ in L^2 norm. The statistical error $\mathcal{E}_{sta}$ can also be referred to as the generalization error, which is caused by the Monte Carlo discretization of $\mathcal{L}(\cdot)$ defined in (3.2) with $\widehat{\mathcal{L}}(\cdot)$ in (3.3).

Step 2. Approximation error bound. In this part, we follow the constructions outlined in [28, 54] and [21] to derive a novel approximation error bound for the Hölder smooth functions with smoothness index $\beta \geq 2$ using *ReLU* activated neural networks. We demonstrate that large *ReLU* networks can not only achieve accurate approximation of a target function in terms of function value, but also approximate the first derivative of the target function.

Theorem 4.1. *Consider ReLU activation function. For any $k \in \mathbb{N}_0$ which satisfies $k \geq \log_2 \mathfrak{C}_2(d,s)+1$ and $f \in \mathcal{H}^\beta([0,1]^d,B)$ where $\beta \geq 2$, $s = \lfloor \beta \rfloor$, $d \geq s$ and $d,s \in \mathbb{N}_0$, there exists $\bar{f}_{\boldsymbol{\theta}} \in \mathcal{NN}(W,L)$ where*

$$W = \mathfrak{C}_1(d,s) 2^{k+\frac{kd}{s-1}},$$
$$L = 4\lceil \log_2(d+s-1) \rceil + 2,$$

such that

$$\|f - \bar{f}_{\boldsymbol{\theta}}\|_{W^{1,\infty}([0,1]^d)} \leq 2\mathfrak{C}_2(d,s) 2^{-k} B \leq B, \tag{4.3}$$

where $\mathfrak{C}_1(d,s)$ and $\mathfrak{C}_2(d,s)$ depend on d and s.

While Hieber [44], Suzuki [48] and Chen et al. [11] established approximation error bounds for *ReLU* deep neural networks in the L^∞ space, our paper presents a novel bound for approximation error specifically in the Sobolev space $W^{1,\infty}$.

Since $\mathcal{H}^\beta([0,1]^d)$ is embedded into $W^{1,\infty}([0,1]^d)$ and $f_0 \in \mathcal{H}^\beta([0,1]^d,B)$, by the approximation result and the triangle inequality, we have $\|\bar{f}_{\boldsymbol{\theta}}\|_{W^{1,\infty}([0,1]^d)} \leq \|f_0 - \bar{f}_{\boldsymbol{\theta}}\|_{W^{1,\infty}([0,1]^d)} + \|f_0\|_{W^{1,\infty}([0,1]^d)} \leq 2B$. It is reasonable to add a constraint to the neural network: for any $f_{\boldsymbol{\theta}} \in \mathcal{P}$, $\|f_{\boldsymbol{\theta}}\|_{W^{1,\infty}([0,1]^d)} \leq 2B$, because the network function class $\mathcal{P}$ defined in this way is non-empty and contains the best approximation element $\bar{f}_{\boldsymbol{\theta}}$. According to the definition, we can say that $\mathcal{P} = \mathcal{NN}(W,L,\|\cdot\|_{W^{1,\infty}},2B)$ when the width of the network is W and the depth is L. We summarize the above analysis in Corollary 4.1.

Assumption 4.2. We assume that the *ReLU* neural network function class $\mathcal{P}$ is $W^{1,\infty}$-norm bounded by $2B$, i.e. $\mathcal{P} = \mathcal{NN}(W,L,\|\cdot\|_{W^{1,\infty}},2B)$.

Corollary 4.1. *Let $\beta \geq 2$, $s = \lfloor \beta \rfloor$, $d \geq s$ and $d,s \in \mathbb{N}_0$, ρ be ReLU activation function. Assume that the problem satisfies Assumption* 4.1 *and the neural network function class $\mathcal{P}$ satisfies Assumption* 4.2. *For $k \in \mathbb{N}_0$ which satisfies $k \geq \log_2 \mathfrak{C}_2(d,s)+1$, there exist neural networks $\bar{f}_{\boldsymbol{\theta}} \in \mathcal{P}$ with*

$$W = \mathfrak{C}_1(d,s) 2^{k+\frac{kd}{s-1}},$$
$$L = 4\lceil \log_2(d+s-1) \rceil + 2,$$

such that

$$\|f-\bar{f}_{\boldsymbol{\theta}}\|_{W^{1,\infty}([0,1]^d)}\leq 2\mathfrak{C}_2(d,s)2^{-k}B\leq B, \tag{4.4}$$

where $\mathfrak{C}_1(d,s)$ and $\mathfrak{C}_2(d,s)$ depend on d and s.

Proof. Directly from Theorem 4.1. □

Step 3. Statistical error bound. We provide a statistical error bound for $W^{1,\infty}$-norm bounded *ReLU* neural networks from the perspective of function space. Our primary contribution lies in demonstrating that norm-bounded over-parameterized neural networks can achieve convergence, with their complexity controlled by the metric entropy of the function space, independent of the number of parameters.

We define the Rademacher complexity of function class $\mathcal{F}$ associate with random sample $\{X_k\}_{k=1}^N$ as

$$\mathcal{R}(\mathcal{F})=\mathop{\mathbb{E}}_{\{X_k,\sigma_k\}_{k=1}^N}\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{k=1}^N\sigma_k f(X_k)\right],$$

where $\{\sigma_k\}_{k=1}^N$ are *i.i.d* Rademacher variables with $\mathbb{P}(\sigma_k=1)=\mathbb{P}(\sigma_k=-1)=\frac{1}{2}$. Then we can drive the statistical error bound by calculating the Rademacher complexity of the function space $\mathcal{P}$.

Lemma 4.1.

$$\mathcal{E}_{sta}=\mathop{\mathbb{E}}_{Z}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]\leq 2\mathcal{R}(\mathcal{P}). \tag{4.5}$$

In order to bound the Rademacher complexity of $\mathcal{P}$, we recall the definition of the covering number.

Definition 4.1. Let W be a function class. For any $\epsilon>0$, let V be an $\epsilon-cover$ of W with respect to the distance d_∞, that is, for any $u\in W$, there exists a $v\in V$ such that $d_\infty(u,v)<\epsilon$, where d_∞ is defined by

$$d_\infty(u,v):=||u-v||_{L^\infty}.$$

The covering number $\mathcal{C}(\epsilon,W,d_\infty)$ is defined to be the minimum cardinality among all $\epsilon-cover$ of W with respect to the distance d_∞.

By applying Dudley's Theorem A.5, our objective is to limit the covering number, which can be bounded above using the following theorem.

Theorem 4.2. *Let $\mathcal{F}$ be the norm-ball of radius $2B$ in $W^{1,\infty}([0,1]^d)$. Then for any $\epsilon>0$*

$$\log\mathcal{C}(\epsilon,\mathcal{F},d_\infty)\leq\mathfrak{C}(d)(2B)^d\epsilon^{-d}=\mathfrak{C}(d,B)\epsilon^{-d}. \tag{4.6}$$

Proof. See [19], Theorem 4.3.36. □

The most important result in this section is now the following, in which we can see that the upper bound of $\mathcal{E}_{sta}$ is independent of W and L.

Theorem 4.3. *Let the sample size is n and $d>2$, $d\in\mathbb{N}_0$. For the regression model, if the neural networks satisfy Assumption* 4.2, *i.e.* $\mathcal{P}=\mathcal{NN}(W,L,\|\cdot\|_{W^{1,\infty}},2B)$, *we have*

$$\mathcal{E}_{sta}=\mathop{\mathbb{E}}_{Z}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]\leq\mathfrak{C}(d,B)n^{-\frac{1}{d}}. \tag{4.7}$$

Step 4. Total error bound. By combining Steps 1 to Step 3, we can obtain our main result.

Theorem 4.4. *Let ρ be ReLU function. Let $B>0$, $\beta\geq 2$, $s=\lfloor\beta\rfloor$, $d\geq s$ and $d>2$, $d,s\in\mathbb{N}_0$. For the problem satisfies Assumption* 4.1, *there exists an over-parameterized neural network function class $\mathcal{P}$ that satisfies Assumption* 4.2. *If the sample size is n, when $k\in\mathbb{N}_0$, $k\geq\max\{\log_2\mathfrak{C}_2(d,s)+1,\mathfrak{C}(d,B,s)\log_2 n\}$ and*

$$\begin{aligned}&W=\mathfrak{C}_1(d,s)2^{k+\frac{kd}{s-1}},\\&L=4\lceil\log_2(d+s-1)\rceil+2,\end{aligned}$$

we have that for the estimator $\widehat{f}_{\boldsymbol{\theta}}\in\mathcal{P}$,

$$\mathop{\mathbb{E}}_{S_n}\left[\|\widehat{f}_{\boldsymbol{\theta}}-f_0\|^2_{L^2(\mu)}\right]\leq\mathfrak{C}(s,d,B)n^{-\frac{1}{d}}, \tag{4.8}$$

where $\mathfrak{C}_1(d,s)$ and $\mathfrak{C}_2(d,s)$ depend on d and s, and $\mathfrak{C}(d,B,s)$ depends on d,B,s.

Remark 4.1. While our work is deeply rooted in classical theory, focusing on exploring size-independent generalization bounds, we recognize that our contributions are just the beginning of our exploration into this complex field. The challenges we've encountered in applying constraints to the $W^{1,\infty}([0,1]^d)$ norms within neural networks have not only highlighted certain limitations but also underscored the need for more detailed exploration and refinement in future research. Specifically, these limitations stem from the fact that approaches to effectively regularize the $W^{1,\infty}([0,1]^d)$ norm are not yet well-established in the current literature. Despite these challenges, it's worth noting that our work provides a valuable perspective for understanding common techniques like weight clipping, batch normalization, and spectral normalization, shedding light on their potential effectiveness in regulating neural network regularity.

4.2 Proof sketch of Theorem 3.2

Deep Ritz method have been proven numerically to be efficient in solving partial differential equations. For the second-order elliptic equations with Neumann boundary conditions (3.4), we make the following assumption on the known terms in equation, for $1<p<\infty$: $f\in W^{2,p}(\Omega)$, $g\in W^{3-\frac{1}{p},p}(\partial\Omega)$, $w(x)\in C^2(\Omega)$, and $\|w(x)\|_{C^2(\Omega)}\geq c_1$ where c_1 is some positive constant. Assume that $M=\max\{\|f\|_{W^{2,p}(\Omega)},\|g\|_{W^{3-\frac{1}{p},p}(\partial\Omega)},\|w\|_{C^2(\Omega)}\}$, and $\partial\Omega\in C^4$.

Lemma 4.2. *The unique weak solution $u^*\in H^1(\Omega)$ of (3.4) is the unique minimizer of $\mathcal{L}(u)$ over $H^1(\Omega)$. Moreover, $u^*\in W^{4,p}(\Omega)$.*

Proof. See [1], Theorem 15.2. □

Assumption 4.3. Assume that $p>d$ and $\|u^*\|_{W^{4,p}(\Omega)}\le B$.

Step 1. Error decomposition.

Proposition 4.2.

$$\frac{c_1\wedge 1}{2}\|u-u^*\|^2_{H^1(\Omega)}\le\mathcal{L}(u)-\mathcal{L}(u^*)\le\frac{\|w\|_{L^\infty(\Omega)}\vee 1}{2}\|u-u^*\|^2_{H^1(\Omega)}.$$

The above statement highlights the properties of the loss function. The following proposition plays a crucial role by dividing the total errors into two distinct types.

Proposition 4.3. Suppose that $\mathcal{P}=\mathcal{N}\mathcal{N}(L,\mathfrak{n}_\theta))$, then

$$\begin{aligned}&\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[||\widehat{u}_\theta-u^*||^2_{H^1(\Omega)}\right]\\ \le&\frac{2}{c_1\wedge 1}\Bigg\{\underbrace{\frac{M\vee 1}{2}\inf_{u\in\mathcal{P}}||u-u^*||^2_{H^1(\Omega)}}_{\mathcal{E}_{app}}+2\underbrace{\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]}_{\mathcal{E}_{sta}}\Bigg\}.\end{aligned}$$

Proof. The proof is similar to the proof of Proposition 4.1. □

Step 2. Approximation error bound. Following the results in [20], we show that for an arbitrary accuracy $\epsilon>0$ and $B>0$, every function from the ball of the Sobolev space $W^{s,p}$

$$\mathcal{F}_{s,d,p}:=\left\{f\in W^{s,p}\left([0,1]^d\right):\|f\|_{W^{s,p}\left([0,1]^d\right)}\le B\right\}$$

can be ϵ-approximated in weaker Sobolev norms $W^{\eta,p}$ (with $s,\eta,p\in\mathbb{N}_0$, $s\ge\eta+1$ and $1\le p\le\infty$) by neural networks with *Sigmoidal* activation function.

Theorem 4.5. *Let $s,\eta,d,p\in\mathbb{N}_0$, $s\ge\eta+1$, $1\le p\le\infty$, ρ be tanh function $\frac{e^x-e^{-x}}{e^x+e^{-x}}$ or sigmoid function $\frac{1}{1+e^{-x}}$. For any $\epsilon>0$ and $f\in\mathcal{F}_{s,d,p}$, there exists a neural network $f_{\boldsymbol{\theta}}\in\mathcal{N}\mathcal{N}(L,\mathfrak{n}_{\boldsymbol{\theta}})$ with depth L and total number of nonzero weights $\mathfrak{n}_{\boldsymbol{\theta}}\ge\mathfrak{C}\epsilon^{-\frac{d}{s-\eta-\mu\eta}}$ such that*

$$\|f-f_{\boldsymbol{\theta}}\|_{W^{\eta,p}\left([0,1]^d\right)}\le\epsilon,$$

where μ is an arbitrarily small positive number and $L,\mathfrak{C}$ depend on d,s,p,η,μ,B.

Proof. See [20]. The main idea is based on the common strategy of approximating f by localized polynomials which in turn are approximated by neural networks. □

Since the region $[0,1]^d$ is larger than the region Ω we consider (recalling that we assumed without loss of generality that $\Omega \subset [0,1]^d$ at the beginning), we need the following extension result.

Lemma 4.3. *Let $\eta \in \mathbb{N}_0$, $1 \le p < \infty$. There exists a linear operator E from $W^{\eta,p}(\Omega)$ to $W_0^{\eta,p}([0,1]^d)$ and $Eu = u$ in Ω.*

Proof. See Theorem 7.25 in [18]. □

From Lemma 4.2 we know that our target function $u^* \in W^{4,p}(\Omega)$. Hence we are able to obtain an approximation result in $W^{3,p}(\Omega)$ norm.

Corollary 4.2. *Let $s=4$, $\eta=3$ and $d \in \mathbb{N}_0$, ρ be tanh function $\frac{e^x - e^{-x}}{e^x + e^{-x}}$ or sigmoid function $\frac{1}{1+e^{-x}}$. For any $\epsilon > 0$ and $u^* \in \mathcal{F}_{4,d,p}$, there exists a neural network $\bar{u}_{\boldsymbol{\theta}} \in \mathcal{NN}(L, \mathfrak{n}_{\boldsymbol{\theta}})$ with depth L and total number of nonzero weights $\mathfrak{n}_{\boldsymbol{\theta}} \ge \mathfrak{C}\epsilon^{-\frac{d}{1-3\mu}}$ such that*

$$\|u^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{3,p}(\Omega)} \le \epsilon,$$

where μ is an arbitrarily small positive number and $L, \mathfrak{C}$ depend on d, μ, p, B.

Proof. Set $s=4$ and $\eta=3$ in Theorem 4.5 and use the fact $\|u^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{3,p}(\Omega)} \le \|Eu^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{3,p}([0,1]^d)}$, where E is the extension operator in Lemma 4.3. □

By Sobolev Imbedding Theorem, when $p > d$, $W^{3,p}(\Omega) \circlearrowleft W^{2,\infty}(\Omega)$ and $W^{4,p}(\Omega) \circlearrowleft W^{2,\infty}(\Omega)$. Then $\|u^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{2,\infty}(\Omega)} \le \|u^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{3,p}(\Omega)} \le \epsilon$. In Assumption 4.3, $\|u^*\|_{W^{3,\infty}(\Omega)} \le B$, then by the approximation result and the triangle inequality, we obtain $\|\bar{u}_{\boldsymbol{\theta}}\|_{W^{2,\infty}(\Omega)} \le \|u^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{2,\infty}(\Omega)} + \|u^*\|_{W^{2,\infty}(\Omega)} \le 2B$. From Corollary 4.2, we know that $\bar{u}_{\boldsymbol{\theta}} \in \mathcal{NN}(L, \mathfrak{n}_{\boldsymbol{\theta}})$ is an infinitely continuously differentiable function, then we have $\|\bar{u}_{\boldsymbol{\theta}}\|_{C^2(\Omega)} \le 2B$. It is reasonable to add a constraint to the neural network: for any $f_{\boldsymbol{\theta}} \in \mathcal{P}$, $\|f_{\boldsymbol{\theta}}\|_{C^2(\Omega)} \le 2B$, because the network function class $\mathcal{P}$ defined in this way contains the best approximation element $\bar{u}_{\boldsymbol{\theta}}$ and is non-empty. According to the definition, we can say that $\mathcal{P} = \mathcal{NN}(L, \mathfrak{n}_{\boldsymbol{\theta}}, \|\cdot\|_{C^2}, 2B)$ when the depth of the network is L and the total number of nonzero weights is $\mathfrak{n}_{\boldsymbol{\theta}}$.

Assumption 4.4. For $s \ge 2$, $1 \le t \le s-1$, $t,s \in \mathbb{N}_0$, we assume that the *Sigmoidal* neural network function class $\mathcal{P} = \mathcal{NN}(L, \mathfrak{n}_{\boldsymbol{\theta}})$ is C^t-norm bounded, i.e. $\mathcal{P} = \mathcal{NN}(L, \mathfrak{n}_{\boldsymbol{\theta}}, \|\cdot\|_{C^t}, 2B)$.

Corollary 4.3. *Let $t=2$, $p > d$, $p,d \in \mathbb{N}_0$ and ρ be tanh function $\frac{e^x - e^{-x}}{e^x + e^{-x}}$ or sigmoid function $\frac{1}{1+e^{-x}}$. Assume that the problem satisfies Assumption 4.3 and the neural network function class $\mathcal{P}$ satisfies Assumption 4.4, i.e. $\mathcal{P} = \mathcal{NN}(L, \mathfrak{n}_{\boldsymbol{\theta}}, \|\cdot\|_{C^2}, 2B)$. For any $0 < \epsilon < B$, there exists a neural network $\bar{u}_{\boldsymbol{\theta}} \in \mathcal{P}$ with depth L and total number of nonzero weights $\mathfrak{n}_{\boldsymbol{\theta}} \ge \mathfrak{C}\epsilon^{-\frac{d}{1-3\mu}}$ such that*

$$\|u^* - \bar{u}_{\boldsymbol{\theta}}\|_{W^{2,\infty}(\Omega)} \le \epsilon,$$

where μ is an arbitrarily small positive number and $L, \mathfrak{C}$ depends on d, μ, p, B.

Proof. Similar to the proof of Corollary 4.1, it can be drawn directly from Corollary 4.2. □

Step 3. Statistical error bound.

Lemma 4.4.

$$\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]\le\mathfrak{C}(d,M,|\partial\Omega|,|\Omega|)\sum_{j=1}^{4}\mathcal{R}(\mathcal{F}_j),$$

where

$$\begin{aligned}&\mathcal{F}_1=\{\pm f:\Omega\to\mathbb{R}\,|\,\exists u\in\mathcal{P},\ s.t.\ f(x)=|\nabla u(x)|^2\},\\&\mathcal{F}_2=\{\pm f:\Omega\to\mathbb{R}\,|\,\exists u\in\mathcal{P},\ s.t.\ f(x)=u^2(x)\},\\&\mathcal{F}_3=\mathcal{P},\quad \mathcal{F}_4=\mathcal{P}|_{\partial\Omega}.\end{aligned}$$

Proof. We can prove this lemma using the same method as Lemma 4.1, by combining equations (3.6) and (3.7) and applying Talagrand's lemma. □

Theorem 4.6. *Let $\mathcal{F}$ be the norm-ball of radius $2B$ in $C^t([0,1]^d)$, $t>0$. Then for any $\epsilon>0$,*

$$\log\mathcal{C}(\epsilon,\mathcal{F},d_\infty)\le\mathfrak{C}(d,t)(2B)^{\frac{d}{t}}\epsilon^{-\frac{d}{t}}=\mathfrak{C}(d,B,t)\epsilon^{-\frac{d}{t}}.$$

Proof. See [19], Theorem 4.3.36. □

When the neural network function space $\mathcal{P}$ satisfies the condition in Assumption 4.4, we have the following result. This demonstrates that the upper bound of $\mathcal{E}_{sta}$ is independent on $\mathfrak{n}_{\boldsymbol{\theta}}$ and L.

Theorem 4.7. *Let the sample size is n, $t=2$ and $d>2$, $d\in\mathbb{N}_0$. For the deep Ritz method, if the neural networks satisfy Assumption 4.4, i.e. $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^2},2B)$, we have*

$$\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]\le\mathfrak{C}(d,B,M)n^{-\frac{1}{d}}.$$

Step 4. Total error bound. Combining Corollary 4.3, Theorem 4.7 and Proposition 4.3, we come to the following conclusion.

Theorem 4.8. *Let ρ be tanh function $\frac{e^x-e^{-x}}{e^x+e^{-x}}$ or sigmoid function $\frac{1}{1+e^{-x}}$. Let $B>0$, $t=2$, $p>d>2$, $p,d\in\mathbb{N}_0$ and μ is an arbitrarily small positive number. For the problem satisfies Assumption 4.3, there exists an over-parameterized neural network function class $\mathcal{P}$ that satisfies Assumption 4.4. If the sample size $N_{in}=N_b=n$, when $\mathfrak{n}_{\boldsymbol{\theta}}$ is no less than*

$$\max\{\mathfrak{C}(d,\mu,B,M,p)n^{\frac{1}{2(1-3\mu)}},\mathfrak{C}(d,\mu,B,p)B^{-\frac{d}{1-3\mu}}\}$$

and L depends on d,μ,B,p, we have that for the estimator $\widehat{u}_{\boldsymbol{\theta}}\in\mathcal{P}$

$$\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}[\|\widehat{u}_{\boldsymbol{\theta}}-u^*\|_{H^1(\Omega)}^2]\le\mathfrak{C}(d,\mu,B,M)n^{-\frac{1}{d}}.$$

4.3 Proof sketch of Theorem 3.3

In Theorem 3.1 we have analyzed the error bound under *ReLU* neural networks, however, since the *ReLU* networks only have the first order smoothness, we can only consider the network function class $\mathcal{P}$ satisfies Assumption 4.2. If we replace the activation function with *Sigmoidal* activation function, we can make stronger assumptions (Assumption 4.4) about the network function class and investigate the impact of the smoothness of the neural network function class on the convergence rate.

Consider the regression model (3.1) with Assumption 4.1. Firstly, we can get the approximation error bound in (4.2) by Theorem 4.5.

Corollary 4.4. *Let $s=\lfloor\beta\rfloor$, $1\le t\le\eta\le s-1$, $p=\infty$ and $t,\eta,d\in\mathbb{N}_0$, ρ be tanh function $\frac{e^x-e^{-x}}{e^x+e^{-x}}$ or sigmoid function $\frac{1}{1+e^{-x}}$. Assume that the problem satisfies Assumption* 4.1 *and the neural network function class $\mathcal{P}$ satisfies Assumption* 4.4*. For any $0<\epsilon<B$, there exists a neural network $\bar{f}_{\boldsymbol{\theta}}\in\mathcal{P}$ with depth L and total number of nonzero weights $\mathfrak{n}_{\boldsymbol{\theta}}\ge\mathfrak{C}\epsilon^{-\frac{d}{s-\eta-\mu\eta}}$ such that*

$$\|f_0-\bar{f}_{\boldsymbol{\theta}}\|_{W^{\eta,\infty}\left([0,1]^d\right)}\le\epsilon,$$

where μ is an arbitrarily small positive number and $L,\mathfrak{C}$ depends on d,μ,B,s,η.

Proof. Since $1\le t\le\eta\le s-1$ and $f_0\in\mathcal{H}^\beta\left([0,1]^d,B\right)$, we have that $\|f_0\|_{W^{t,\infty}([0,1]^d)}\le B$. Regardless of the Assumption 4.4, we have $\|\bar{f}_{\boldsymbol{\theta}}\|_{W^{\eta,\infty}([0,1]^d)}\le\epsilon<B$ by the approximation result in Theorem 4.5 with $p=\infty$. Using the triangle inequality, we have $\|\bar{f}_{\boldsymbol{\theta}}\|_{W^{t,\infty}([0,1]^d)}\le\|f_0-\bar{f}_{\boldsymbol{\theta}}\|_{W^{\eta,\infty}([0,1]^d)}+\|f_0\|_{W^{t,\infty}([0,1]^d)}\le 2B$. Since the activation function is infinitely continuously differentiable, we have $\|\bar{f}_{\boldsymbol{\theta}}\|_{C^t([0,1]^d)}\le 2B$. It means that the assumption $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^t},2B)$ is reasonable and does not change the approximation result. The proof is complete. □

Secondly, we can get the statistical error bound in (4.2) by combining Lemma 4.1, Dudley's theorem A.5 and Theorem 4.6.

Theorem 4.9. *Let the sample size is n, $1\le t<d/2$ and $t,d\in\mathbb{N}_0$. For the regression model, if the neural networks satisfy Assumption* 4.4*, i.e. $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^t},2B)$, we have*

$$\mathcal{E}_{sta}=\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]\le\mathfrak{C}(d,B,t)n^{-\frac{t}{d}}.$$

Finally, combining Corollary 4.4, Theorem 4.9 and Proposition 4.1 with $s=\lfloor\beta\rfloor$, $1\le t\le\eta\le s-1$, $p=\infty$, $t<d/2$ and $t,\eta,d\in\mathbb{N}_0$, we come to the total error bound.

Theorem 4.10. *Let ρ be tanh function $\frac{e^x-e^{-x}}{e^x+e^{-x}}$ or sigmoid function $\frac{1}{1+e^{-x}}$. Let $B>0$, $s=\lfloor\beta\rfloor$, $1\le t\le\eta\le s-1$, $p=\infty$, $t<d/2$ and $t,\eta,d\in\mathbb{N}_0$, and μ is an arbitrarily small positive number. Consider the problem satisfies Assumption* 4.1*. For each t, there exists an over-parameterized*

neural network function class $\mathcal{P}$ *that satisfies Assumption* 4.4. *If the sample size is* n, *when* $\mathfrak{n}_{\boldsymbol{\theta}} \geq \max\{\mathfrak{C}(d,\mu,B,s,t,\eta) n^{\frac{t}{2(s-\eta-\mu\eta)}}, \mathfrak{C}B^{-\frac{d}{s-\eta-\mu\eta}}\}$, $\mathfrak{C}$ *and* L *depend on* d,μ,B,s,η, *we have that for the estimator* $\widehat{f}_{\boldsymbol{\theta}} \in \mathcal{P}$,

$$\mathop{\mathbb{E}}_{S_n}\left[\|\widehat{f}_{\boldsymbol{\theta}} - f_0\|_{L^2(\mu)}^2\right] \leq \mathfrak{C}(s,d,B,t,\eta,\mu) n^{-\frac{t}{d}}.$$

5 Conclusion

In this paper, we introduce a new theoretical framework based on function space theory and establish the consistency of norm-bounded over-parameterized deep learning. We demonstrate that the complexity of a neural network can be controlled by the metric entropy of the balls in certain metric space, which is independent of the number of parameters. This provides a novel perspective for understanding the good generalization ability of over-parameterized neural networks.

Further research is needed to explore related issues. One area of focus is finding new methods for estimating approximation error, particularly for target functions with smoothness $\beta < 2$. Another is extending this framework to include other activation functions and to different kinds of problems.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2023YFA1000103), by the National Natural Science Foundation of China (No. 123B2019, No. 12125103, No. 12071362, No. 12371424, No. 12371441), and by the Fundamental Research Funds for the Central Universities.

A Appendix

A.1 Proof of regression model

A.1.1 Proof of Proposition 4.1

For any $\bar{f} \in \mathcal{P}$, we have

$$\begin{aligned} \mathcal{L}(\widehat{f}_{\boldsymbol{\theta}}) - \mathcal{L}(f_0) &= \mathcal{L}(\widehat{f}_{\boldsymbol{\theta}}) - \widehat{\mathcal{L}}(\widehat{f}_{\boldsymbol{\theta}}) + \widehat{\mathcal{L}}(\widehat{f}_{\boldsymbol{\theta}}) - \widehat{\mathcal{L}}(\bar{f}) + \widehat{\mathcal{L}}(\bar{f}) - \mathcal{L}(f_0) \\ &\leq \mathcal{L}(\widehat{f}_{\boldsymbol{\theta}}) - \widehat{\mathcal{L}}(\widehat{f}_{\boldsymbol{\theta}}) + \widehat{\mathcal{L}}(\bar{f}) - \mathcal{L}(f_0), \end{aligned}$$

where the inequality is due to the fact that $\widehat{\mathcal{L}}(\widehat{f}_{\boldsymbol{\theta}})-\widehat{\mathcal{L}}(\bar{f})\leq 0$. Take the expectation of both sides of this equation, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{S_n}\left[\mathcal{L}(\widehat{f}_{\boldsymbol{\theta}})-\mathcal{L}(f_0)\right]&\leq\mathop{\mathbb{E}}_{S_n}\left[\mathcal{L}(\widehat{f}_{\boldsymbol{\theta}})-\widehat{\mathcal{L}}(\widehat{f}_{\boldsymbol{\theta}})+\widehat{\mathcal{L}}(\bar{f})-\mathcal{L}(f_0)\right]\\&\leq\mathop{\mathbb{E}}_{S_n}\left[\widehat{\mathcal{L}}(\bar{f})-\mathcal{L}(f_0)\right]+\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]\\&=\left[\mathcal{L}(\bar{f})-\mathcal{L}(f_0)\right]+\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right].\end{aligned}$$

Since $\bar{f}$ can be any element in $\mathcal{P}$, we take the infimum of $\bar{f}$:

$$\mathop{\mathbb{E}}_{S_n}\left[\mathcal{L}(\widehat{f}_{\boldsymbol{\theta}})-\mathcal{L}(f_0)\right]\leq\inf_{\bar{f}\in\mathcal{P}}\left[\mathcal{L}(\bar{f})-\mathcal{L}(f_0)\right]+\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right].$$

By $\|\bar{f}-f_0\|^2_{L^2(\mu)}=\mathcal{L}(\bar{f})-\mathcal{L}(f_0)$,

$$\mathop{\mathbb{E}}_{S_n}\left[\|\widehat{f}_{\boldsymbol{\theta}}-f_0\|^2_{L^2(\mu)}\right]\leq\inf_{\bar{f}\in\mathcal{P}}\|\bar{f}-f_0\|^2_{L^2(\mu)}+\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right].$$

A.1.2 Proof of Theorem 4.1

The approximation error depends on $\mathcal{P}$ through its parameters and is related to the smoothness of the target. In this section, the function class $\mathcal{P}=\mathcal{NN}(W,L)$ consists of the feed-forward neural networks with the *ReLU* activation function. We show that every function from the unit ball of the Sobolev space $W^{s,\infty}$

$$\mathcal{F}_{s,d,\infty}:=\left\{f\in W^{s,\infty}\left([0,1]^d\right):\|f\|_{W^{s,\infty}\left([0,1]^d\right)}\leq B\right\}$$

can be approximated in weaker Sobolev norms $W^{1,\infty}$(with $s\geq 2$) by neural networks with *ReLU* activation function. The main idea is based on the common strategy of approximating f by localized polynomials which in turn are approximated by neural networks. Following the constructions in [20] and [28], we can build approximate partitions of unity which are compatible with *ReLU* activation function. Firstly, we consider the approximation of the quadratic function $f(x)=x^2$.

Lemma A.1. *Let $f(x)=x^2$, for any $k\in\mathbb{N}$, there exists $\phi_k^0\in\mathcal{NN}\left(2^{k+1},2\right)$ such that $\phi_k^0(0)=0$ and*

$$\|f-\phi_k^0\|_{L^\infty([0,1])}\leq 2^{-2(k+1)},\quad \|f-\phi_k^0\|_{W^{1,\infty}([0,1])}\leq 2^{-k}.$$

Proof. We modify the construction in [28]. We define a set of teeth functions $T_i:\mathbb{R}\to[0,1]$ by

$$T_1(x):=\begin{cases}2x, & 0\leq x\leq\frac{1}{2},\\ 2(1-x), & \frac{1}{2}<x\leq 1,\\ 0, & \text{else,}\end{cases}$$

and $T_{i+1}=T_1\circ T_i$, for $i\in\mathbb{N}$. It is easy to check that T_i has 2^{i-1} teeth and

$$T_i(x)=\sum_{j=0}^{2^{i-1}-1}\sigma\left(1-\left|2^ix-2j-1\right|\right).$$

Since $|x|=\sigma(x)+\sigma(-x)$, the function

$$\begin{aligned}f_{i,j}(x)=&\sigma\left(1-\left|2^ix-2j-1\right|\right)\\=&\sigma\left(1-\sigma\left(2^ix-2j-1\right)-\sigma\left(-2^ix+2j+1\right)\right)\\=&\sigma\left(1-\left(2^i+2j+1\right)\sigma\left(\frac{2^i}{2^i+2j+1}x-\frac{2j+1}{2^i+2j+1}\right)\right.\\&\left.-\left(2^i+2j+1\right)\sigma\left(-\frac{2^i}{2^i+2j+1}x+\frac{2j+1}{2^i+2j+1}\right)\right)\end{aligned}$$

is in $\mathcal{N}\mathcal{N}(2,2)$. Then $T_i\in\mathcal{N}\mathcal{N}(2^i,2)$.

For any $k\in\mathbb{N}$, let $\phi_k^0:[0,1]\to[0,1]$ be the piecewise linear function such that $\phi_k^0(\frac{j}{2^k})=(\frac{j}{2^k})^2$ for $j=0,1,\cdots,2^k$, and ϕ_k^0 is linear on $[\frac{j-1}{2^k},\frac{j}{2^k}]$ for $j=1,2,\cdots,2^k$. Then, for all $x\in[\frac{j-1}{2^k},\frac{j}{2^k}]$,

$$\begin{aligned}&\phi_k^0(x)=\left(\frac{j^2}{2^k}-\frac{(j-1)^2}{2^k}\right)\left(x-\frac{j-1}{2^k}\right)+\left(\frac{j-1}{2^k}\right)^2,\quad j=1,2,\cdots,2^k,\\&D(\phi_k^0(x))=\frac{j^2}{2^k}-\frac{(j-1)^2}{2^k}.\end{aligned}$$

Since $\|f-\phi_k^0\|_{W^{1,\infty}([0,1])}=\max\left\{\|f-\phi_k^0\|_{L^\infty([0,1])},\|Df-D\phi_k^0\|_{L^\infty([0,1])}\right\}$,

$$\begin{aligned}\|f-\phi_k^0\|_{L^\infty([0,1])}=&\left|x^2-\left(\frac{j^2}{2^k}-\frac{(j-1)^2}{2^k}\right)\left(x-\frac{j-1}{2^k}\right)+\left(\frac{j-1}{2^k}\right)^2\right|\\=&\left|\left(x-\frac{2j-1}{2^{k+1}}\right)^2-\left(\frac{1}{2^{k+1}}\right)^2\right|\le\left(\frac{1}{2^{k+1}}\right)^2,\\\|Df-D\phi_k^0\|_{L^\infty([0,1])}=&\max\left\{\left|\frac{2j-1}{2^k}-\frac{2(j-1)}{2^k}\right|,\left|\frac{2j-1}{2^k}-\frac{2j}{2^k}\right|\right\}=2^{-k}.\end{aligned}$$

Therefore,

$$\|f-\phi_k\|_{W^{1,\infty}([0,1])}\le 2^{-k},\quad k\in\mathbb{N}.$$

Furthermore, $\phi_{k-1}^0(x)-\phi_k^0(x)=\frac{T_k(x)}{4^k}$ and $x-\phi_1^0(x)=\frac{T_1(x)}{4}$. Hence,

$$\phi_k^0(x)=x-\left(x-\phi_1^0(x)\right)-\sum_{i=2}^k\left(\phi_{i-1}^0(x)-\phi_i^0(x)\right)=\sigma(x)-\sum_{i=1}^k\frac{T_i(x)}{4^i},\quad x\in[0,1],\quad k\in\mathbb{N}.$$

Since $\sum_{i=1}^k 2^i+1=2^{k+1}-1$, we have $\phi_k^0\in\mathcal{N}\mathcal{N}(2^{k+1}-1,2)$, which completes the proof. □

Using the relation

$$xy=2\left(\left(\frac{|x+y|}{2}\right)^2-\left(\frac{|x|}{2}\right)^2-\left(\frac{|y|}{2}\right)^2\right),$$

we can approximate the product function by neural networks and then further approximate any monomials $x_1\cdots x_d$.

Lemma A.2. *Let $f(x,y)=xy$, for any $k\in\mathbb{N}$, there exists $\phi_k^1\in\mathcal{NN}\left(3\cdot 2^{k+1},4\right)$ such that $\phi_k^1:[-1,1]^2\to[-1,1]$ and*

$$\left\|f-\phi_k^1\right\|_{W^{1,\infty}([-1,1]^2)}\le 6\cdot 2^{-k}.$$

Furthermore, $\phi_k^1(x,y)=0$ if $xy=0$.

Proof. By Lemma A.1, if $\widetilde{f}(x)=x^2$, there exists network $\phi_k^0\in\mathcal{NN}\left(2^{k+1}-1,2\right)$ such that $\left\|\widetilde{f}-\phi_k^0\right\|_{W^{1,\infty}([0,1])}\le 2^{-k}$ and $\phi_k^0(0)=0$. Using

$$xy=2\left(\left(\frac{|x+y|}{2}\right)^2-\left(\frac{|x|}{2}\right)^2-\left(\frac{|y|}{2}\right)^2\right),$$

we consider the function

$$\begin{aligned}\widetilde{\phi}_k^1(x,y):=&2\phi_k^0\left(\frac{1}{2}|x+y|\right)-2\phi_k^0\left(\frac{1}{2}|x|\right)-2\phi_k^0\left(\frac{1}{2}|y|\right)\\=&2\phi_k^0\left(\frac{1}{2}\sigma(x+y)+\frac{1}{2}\sigma(-x-y)\right)\\&-2\phi_k^0\left(\frac{1}{2}\sigma(x)+\frac{1}{2}\sigma(-x)\right)-2\phi_k^0\left(\frac{1}{2}\sigma(y)+\frac{1}{2}\sigma(-y)\right).\end{aligned}$$

Then, $\widetilde{\phi}_k^1(x,y)=0$ if $xy=0$, and, for any $x,y\in[-1,1]$,

$$\begin{aligned}\left\|xy-\widetilde{\phi}_k^1(x,y)\right\|_{W^{1,\infty}([0,1]^2)}\le&2\left\|\left(\frac{|x+y|}{2}\right)^2-\phi_k^0\left(\frac{|x+y|}{2}\right)\right\|_{W^{1,\infty}([0,1]^2)}\\&+2\left\|\left(\frac{|x|}{2}\right)^2-\phi_k^0\left(\frac{|x|}{2}\right)\right\|_{W^{1,\infty}([0,1]^2)}\\&+2\left\|\left(\frac{|y|}{2}\right)^2-\phi_k^0\left(\frac{|y|}{2}\right)\right\|_{W^{1,\infty}([0,1]^2)}\\\le&6\cdot 2^{-k}.\end{aligned}$$

Finally, let $\chi(x)=\sigma(x)-\sigma(-x)-2\sigma\left(\frac{1}{2}x-\frac{1}{2}\right)+2\sigma\left(-\frac{1}{2}x-\frac{1}{2}\right)=(x\vee -1)\wedge 1$, then $\chi\in\mathcal{NN}(4,1)$. We construct the target function as

$$\phi_k^1(x,y)=\chi\left(\widetilde{\phi}_k^1(x,y)\right)=\left(\widetilde{\phi}_k^1(x,y)\vee -1\right)\wedge 1.$$

Then, for any $x,y\in[-1,1]$,

$$\left\|xy-\phi_k^1(x,y)\right\|_{W^{1,\infty}([0,1]^2)}\le\left\|xy-\widetilde{\phi}_k^1(x,y)\right\|_{W^{1,\infty}([0,1]^2)}\le 6\cdot 2^{-k}.$$

Here, $\phi_k^1\in\mathcal{NN}\left(3\cdot 2^{k+1},4\right)$. □

Lemma A.3. *For any $k\in\mathbb{N}$, $a,b\in\mathbb{R}$ with $a<b$, there exists $\phi_k^1\in\mathcal{NN}\left(3\cdot 2^{k+1},4\right)$ such that*

$$\left\|xy-\phi_k^1(x,y)\right\|_{W^{1,\infty}([a,b]^2)}\le 6(b-a)^2 2^{-k}.$$

Proof. See [22]. By Lemma A.1 and Lemma A.2, there exists $\hat{\phi}_k^1\in\mathcal{NN}\left(3\cdot(2^{k+1}-1),4\right)$ such that

$$\left\|\hat{x}\hat{y}-\hat{\phi}_k^1(\hat{x},\hat{y})\right\|_{W^{1,\infty}([0,1]^2)}\le 6\cdot 2^{-k}.$$

By setting $x=a+(b-a)\hat{x}$ and $y=a+(b-a)\hat{y}$ for any $\hat{x},\hat{y}\in[0,1]$, we define the following network ϕ_k^1

$$\phi_k^1=(b-a)^2\hat{\phi}_k^1\left(\frac{x-a}{b-a},\frac{y-a}{b-a}\right)+a(x-a)+a(y-a)+a^2.$$

Note that $a(x-a)+a(y-a)$ is positive. Hence, the width of ϕ_k^1 can be as small as $3\cdot(2^{k+1}-1)+1<3\cdot(2^{k+1})$. Thus, by $xy=(b-a)^2\left(\frac{x-a}{b-a}\cdot\frac{y-a}{b-a}\right)+a(x-a)+a(y-a)+a^2$, we have

$$\begin{aligned}&\left\|\phi_k^1(x,y)-xy\right\|_{W^{1,\infty}([a,b]^2)}\\=&(b-a)^2\left\|\hat{\phi}_k^1\left(\frac{x-a}{b-a},\frac{y-a}{b-a}\right)-\left(\frac{x-a}{b-a}\cdot\frac{y-a}{b-a}\right)\right\|_{W^{1,\infty}([a,b]^2)}\\\le&(b-a)^2 6\cdot 2^{-k}.\end{aligned}$$

This completes the proof. □

Lemma A.4. *For any $d\ge 2$, $m=\lceil\log_2 d\rceil$ and $k\in\mathbb{N}$, $k\ge 2\lceil\log_2 d\rceil+2$, there exists $\phi_k^m\in\mathcal{NN}\left(6d\cdot 2^k,4\lceil\log_2 d\rceil\right)$ such that $\phi_k^m:[-1,1]^d\to[-1,1]$ and*

$$\|x_1\cdots x_d-\phi_k^m(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}\le 3\cdot 2^{2\lceil\log_2 d\rceil-1-k},\quad \boldsymbol{x}=(x_1,\cdots,x_d)^\top.$$

Furthermore, $\phi_k^m(\boldsymbol{x})=0$ if $x_1\cdots x_d=0$.

Proof. $m=\lceil\log_2 d\rceil$. For $m=1$, by Lemma A.2, there exists $\phi_k^1\in\mathcal{NN}\left(3\cdot 2^{k+1},4\right)$ such that $\phi_k^1:[-1,1]^2\to[-1,1]$ and $\|x_1x_2-\phi_k^1(x_1,x_2)\|_{W^{1,\infty}([-1,1]^2)}\le 6\cdot 2^{-k}$ for any $x_1,x_2\in[-1,1]$. We define $\phi_k^m:[-1,1]^{2^m}\to[-1,1]$ inductively by

$$\phi_k^m(x_1,\cdots,x_{2^m})=\phi_k^1\left(\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}}),\phi_k^{m-1}(x_{2^{m-1}+1},\cdots,x_{2^m})\right).$$

Then, $\phi_k^m(x_1,\cdots,x_{2^m})=0$ if $x_1\cdots x_{2^m}=0$ because the equation is true for $m=1$. Next, we inductively show that $\phi_k^m\in\mathcal{NN}\left(3\cdot 2^{k+m},4m\right)$ and

$$\|x_1\cdots x_{2^m}-\phi_k^m(x_1,\cdots,x_{2^m})\|_{W^{1,\infty}([-1,1]^{2^m})}\le 3\cdot 2^{2m-1-k}=4^{m-1}\cdot 6\cdot 2^{-k}.$$

It is obvious that the assertion is true for $m=1$ by construction. Assume that the assertion is true for some $m-1\in\mathbb{N}$, we will prove that it is true for m. By the assumption, $\phi_k^{m-1}\in\mathcal{NN}(3\cdot 2^{m+k-1},4m-4)$ and

$$\left\|x_1\cdots x_{2^{m-1}}-\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}})\right\|_{W^{1,\infty}([-1,1]^{2^{m-1}})}\le 4^{m-2}\cdot 6\cdot 2^{-k},$$

then we have $\phi_k^m\in\mathcal{NN}(2\cdot 3\cdot 2^{m+k-1},4m-4+4)=\mathcal{NN}(3\cdot 2^{m+k},4m)$ and

$$\begin{aligned}
&\|x_1\cdots x_{2^m}-\phi_k^m(x_1,\cdots,x_{2^m})\|_{W^{1,\infty}([-1,1]^{2^m})}\\
=&\left\|x_1\cdots x_{2^{m-1}}\cdot x_{2^{m-1}+1}\cdots x_{2^m}-\phi_k^1\left(\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}}),\phi_k^{m-1}(x_{2^{m-1}+1},\cdots,x_{2^m})\right)\right\|_{W^{1,\infty}}\\
\le&\|x_1\cdots x_{2^{m-1}}\cdot(x_{2^{m-1}+1}\cdots x_{2^m})-\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}})\cdot(x_{2^{m-1}+1}\cdots x_{2^m})\|_{W^{1,\infty}}\\
&+\|\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}})\|_{W^{1,\infty}}\cdot\|(x_{2^{m-1}+1}\cdots x_{2^m})-\phi_k^{m-1}(x_{2^{m-1}+1},\cdots,x_{2^m})\|_{W^{1,\infty}}\\
&+\|\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}})\cdot\phi_k^{m-1}(x_{2^{m-1}+1},\cdots,x_{2^m})\\
&\qquad-\phi_k^1\left(\phi_k^{m-1}(x_1,\cdots,x_{2^{m-1}}),\phi_k^{m-1}(x_{2^{m-1}+1},\cdots,x_{2^m})\right)\|_{W^{1,\infty}}\\
\le&4^{m-2}\cdot 6\cdot 2^{-k}+(1+4^{m-2}\cdot 6\cdot 2^{-k})4^{m-2}\cdot 6\cdot 2^{-k}+6\cdot 2^{-k}(1+2\cdot 4^{m-1}\cdot 6\cdot 2^{-k})^2\\
\le&4\cdot 4^{m-2}\cdot 6\cdot 2^{-k}=4^{m-1}\cdot 6\cdot 2^{-k},
\end{aligned}$$

where the first inequality is due to the triangle inequality. By the induction hypothesis and Lemma A.3, we get the second inequality. Since $k\ge 2\lceil\log_2 d\rceil+2=2m+2$, we have $4^{m-1}\cdot 6\cdot 2^{-k}<1$, which can derive the last inequality. Hence, the assertion is true for m, the proof is complete. □

In Lemma A.4 we construct neural networks to approximate monomials. We can then approximate any $f\in W^{s,\infty}$ by approximating its local Taylor expansion.

$$p(\boldsymbol{x})=\sum_{\boldsymbol{n}\in\{0,1,\cdots,N\}^d}\psi_{\boldsymbol{n}}(\boldsymbol{x})\sum_{\|\boldsymbol{\alpha}\|_1\le s}\frac{\partial^{\boldsymbol{\alpha}}f\left(\frac{\boldsymbol{n}}{N}\right)}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}},\tag{A.1}$$

where we use the usual conventions $\boldsymbol{\alpha}!=\prod_{i=1}^d\alpha_i!$ and $\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}=\prod_{i=1}^d\left(x_i-\frac{n_i}{N}\right)^{\alpha_i}$. The functions $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}}$ form a partition of unity of $[0,1]^d$ and each $\psi_{\boldsymbol{n}}$ is supported on a sufficiently small neighborhood of $\boldsymbol{n}/N$.

Theorem A.1. *For any $N,k\in\mathbb{N}$, $k\geq 2\lceil\log_2(d+s-1)\rceil+2$ and $f\in\mathcal{F}_{s,d,\infty}$ where $s\in\mathbb{N}_0$ and $s\geq 2$, there exists $\phi\in\mathcal{NN}(W,L)$ where*

$$\begin{aligned}&W=6s(d+s-1)d^{s-1}(N+1)^d2^k=\mathfrak{C}_1(d,s)(N+1)^d2^k,\\&L=4\lceil\log_2(d+s-1)\rceil+2,\end{aligned}$$

such that

$$\begin{aligned}\|f-\phi\|_{W^{1,\infty}([0,1]^d)}&\leq 2^dd^sB\left(N^{-s+1}+3\cdot\frac{s}{d}\cdot 2^{2\lceil\log_2(d+s-1)\rceil-1-k}\right)\\&=\mathfrak{C}_2(d,s)(N^{-s+1}+2^{-k})B,\end{aligned}$$

where $\mathfrak{C}_1(d,s)=6s(d+s-1)d^{s-1}$ and $\mathfrak{C}_2(d,s)=2^dd^s\cdot 3\cdot\frac{s}{d}\cdot 2^{2\lceil\log_2(d+s-1)\rceil-1}$.

Proof. Let

$$\psi(t)=\sigma(1-|t|)=\sigma(1-\sigma(t)-\sigma(-t))\in[0,1],\quad t\in\mathbb{R},$$

then $\psi\in\mathcal{NN}(2,2)$ and the support of ψ is $[-1,1]$. For any $\boldsymbol{n}=(n_1,\cdots,n_d)\in\{0,1,\cdots,N\}^d$, define

$$\psi_{\boldsymbol{n}}(\boldsymbol{x}):=\prod_{i=1}^d\psi(Nx_i-n_i),\quad \boldsymbol{x}=(x_1,\cdots,x_d)^\top\in\mathbb{R}^d,$$

then $\psi_{\boldsymbol{n}}$ is supported on $\left\{\boldsymbol{x}\in\mathbb{R}^d:\left\|\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right\|_{L^\infty}\leq\frac{1}{N}\right\}$. The functions $\{\psi_{\boldsymbol{n}}\}_{\boldsymbol{n}}$ form a partition of unity of the domain $[0,1]^d$:

$$\sum_{\boldsymbol{n}\in\{0,1,\cdots,N\}^d}\psi_{\boldsymbol{n}}(\boldsymbol{x})=\prod_{i=1}^d\sum_{n_i=0}^N\psi(Nx_i-n_i)\equiv 1,\quad \boldsymbol{x}\in[0,1]^d.$$

Let $p(\boldsymbol{x})$ be the local Taylor expansion (A.1).

$$p(\boldsymbol{x})=\sum_{\boldsymbol{n}\in\{0,\cdots,N\}^d}\psi_{\boldsymbol{n}}(\boldsymbol{x})\sum_{\|\boldsymbol{\alpha}\|_1\leq s-1}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}.\tag{A.2}$$

For a fixed $\boldsymbol{n}\in\{0,1,\cdots,N\}^d$ and any $\boldsymbol{x}\in\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\|_{L^\infty}\leq\frac{1}{N}\}$, by the Taylor expansion there exists a $\xi_{\boldsymbol{x}}\in(0,1)$ such that

$$f(\boldsymbol{x})=\sum_{\|\boldsymbol{\alpha}\|_1\leq s-1}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}+\sum_{\|\boldsymbol{\alpha}\|_1=s}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N}+\xi_{\boldsymbol{x}}(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}))}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}.$$

We denote the first order derivatives of f by $\partial^\gamma f$ with $\|\gamma\|_1=1$. For $\boldsymbol{x}\in\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\|_{L^\infty}\leq\frac{1}{N}\}$, by Taylor's expansion there exists a $\xi_{\boldsymbol{x}}^\gamma\in(0,1)$ such that

$$\begin{aligned}\partial^\gamma f(\boldsymbol{x})&=\sum_{\|\boldsymbol{\alpha}\|_1\leq s-2}\frac{\partial^{\boldsymbol{\alpha}}\partial^\gamma f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}+\sum_{\|\boldsymbol{\alpha}\|_1=s-1}\frac{\partial^{\boldsymbol{\alpha}}\partial^\gamma f(\frac{\boldsymbol{n}}{N}+\xi_{\boldsymbol{x}}^\gamma(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}))}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}\\&=\partial^\gamma\left(\sum_{\|\boldsymbol{\alpha}\|_1\leq s-1}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}\right)+\sum_{\|\boldsymbol{\alpha}\|_1=s-1}\frac{\partial^{\boldsymbol{\alpha}}\partial^\gamma f(\frac{\boldsymbol{n}}{N}+\xi_{\boldsymbol{x}}^\gamma(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}))}{\boldsymbol{\alpha}!}\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}.\end{aligned}$$

Then for any $\boldsymbol{x}\in[0,1]^d$, $f(\boldsymbol{x})=\sum_{\boldsymbol{n}\in\{0,1,\cdots,N\}^d}\psi_{\boldsymbol{n}}(\boldsymbol{x})f(\boldsymbol{x})$.

$$\begin{aligned}\|f(\boldsymbol{x})-p(\boldsymbol{x})\|_{L^\infty([0,1]^d)}=&\Bigg\|\sum_{\boldsymbol{n}\in\{0,1,\cdots,N\}^d}\psi_{\boldsymbol{n}}(\boldsymbol{x})\Bigg(\sum_{\|\boldsymbol{\alpha}\|_1\le s-1}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}\\&+\sum_{\|\boldsymbol{\alpha}\|_1=s}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N}+\xi_{\boldsymbol{x}}(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}))}{\boldsymbol{\alpha}!}\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}\Bigg)-p(\boldsymbol{x})\Bigg\|_{L^\infty([0,1]^d)}\\=&\Bigg\|\sum_{\boldsymbol{n}\in\{0,1,\cdots,N\}^d}\psi_{\boldsymbol{n}}(\boldsymbol{x})\sum_{\|\boldsymbol{\alpha}\|_1=s}\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N}+\xi_{\boldsymbol{x}}(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}))}{\boldsymbol{\alpha}!}\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}\Bigg\|_{L^\infty([0,1]^d)}\\\le&\sum_{\boldsymbol{n}:\|\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\|_{L^\infty}\le\frac{1}{N}}\sum_{\|\boldsymbol{\alpha}\|_1=s}\Bigg\|\frac{B}{\boldsymbol{\alpha}!}\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}\Bigg\|_{L^\infty([0,1]^d)}\\\le&2^dd^sBN^{-s},\end{aligned}$$

and

$$\begin{aligned}\|Df(\boldsymbol{x})-Dp(\boldsymbol{x})\|_{L^\infty([0,1]^d)}\le&2^d\max_{\|\gamma\|_1=1}\Bigg\|\sum_{\|\boldsymbol{\alpha}\|_1=s-1}\frac{\partial^{\boldsymbol{\alpha}}\partial^\gamma f(\frac{\boldsymbol{n}}{N}+\xi_{\boldsymbol{x}}^\gamma(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}))}{\boldsymbol{\alpha}!}\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}\Bigg\|_{L^\infty([0,1]^d)}\\\le&2^dd^{s-1}BN^{-(s-1)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\|f(\boldsymbol{x})-p(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}=&\max\Big\{\|f(\boldsymbol{x})-p(\boldsymbol{x})\|_{L^\infty([0,1]^d)},\|Df(\boldsymbol{x})-Dp(\boldsymbol{x})\|_{L^\infty([0,1]^d)}\Big\}\\\le&2^dd^sBN^{-(s-1)}.\end{aligned}$$

Next, we need to construct the neural network ϕ and bound $\|p(\boldsymbol{x})-\phi(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}$. For convenience, we denote $c_{\boldsymbol{n},\boldsymbol{\alpha}}:=\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}$, then by (A.2)

$$\begin{aligned}p(\boldsymbol{x})=&\sum_{\boldsymbol{n}\in\{0,\cdots,N\}^d}\sum_{\|\boldsymbol{\alpha}\|_1\le s-1}c_{\boldsymbol{n},\boldsymbol{\alpha}}\cdot\psi_{\boldsymbol{n}}(\boldsymbol{x})\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}\\=&\sum_{\boldsymbol{n}\in\{0,\cdots,N\}^d}\sum_{\|\boldsymbol{\alpha}\|_1\le s-1}c_{\boldsymbol{n},\boldsymbol{\alpha}}\prod_{i=1}^d\psi(Nx_i-n_i)\Big(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\Big)^{\boldsymbol{\alpha}}.\end{aligned}$$

We can approximate $p(\boldsymbol{x})$ by

$$\phi(\boldsymbol{x})=\sum_{\boldsymbol{n}\in\{0,\cdots,N\}^d}\sum_{\|\boldsymbol{\alpha}\|_1\le s-1}c_{\boldsymbol{n},\boldsymbol{\alpha}}\phi_{\boldsymbol{n},\boldsymbol{\alpha}}(\boldsymbol{x}).$$

When we set $m=\lceil\log_2(d+\|\boldsymbol{\alpha}\|_1)\rceil$,

$$\phi_{\boldsymbol{n},\boldsymbol{\alpha}}(\boldsymbol{x}):=\phi_k^m\Big(\psi(Nx_1-n_1),\cdots,\psi(Nx_d-n_d),\cdots,x_i-\frac{n_i}{N},\cdots\Big),$$

where the term $x_i-\frac{n_i}{N}$ appears in the input only when $\alpha_i\neq 0$ and it repeats α_i times. Since $x_i-n_i/N=\sigma(x_i-n_i/N)-\sigma(-x_i+n_i/N)$, by Lemma A.4, $\phi_{\boldsymbol{n},\boldsymbol{\alpha}}\in\mathcal{NN}(6(d+s-1)\cdot 2^k,4\lceil\log_2(d+s-1)\rceil)$ and the approximation error is

$$\left\|\prod_{i=1}^{d}\psi(Nx_i-n_i)\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}-\phi_{\boldsymbol{n},\boldsymbol{\alpha}}(\boldsymbol{x})\right\|_{W^{1,\infty}([0,1]^d)}\leq 3\cdot 2^{2\lceil\log_2(d+\|\boldsymbol{\alpha}\|_1)\rceil-1-k}.$$

Observe that $|c_{\boldsymbol{n},\boldsymbol{\alpha}}|=|\frac{\partial^{\boldsymbol{\alpha}}f(\frac{\boldsymbol{n}}{N})}{\boldsymbol{\alpha}!}|\leq B$ and the number of terms in the inner summation is

$$\sum_{\|\boldsymbol{\alpha}\|_1\leq s-1}1=\sum_{j=0}^{s-1}\sum_{\|\boldsymbol{\alpha}\|_1=j}1\leq\sum_{j=0}^{s-1}d^j\leq sd^{s-1}.$$

The approximation error is, for any $\boldsymbol{x}\in[0,1]^d$,

$$\begin{aligned}\|p(\boldsymbol{x})-\phi(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}&\leq\sum_{\boldsymbol{n}}\sum_{\|\boldsymbol{\alpha}\|_1\leq s-1}B\left\|\prod_{i=1}^{d}\psi(Nx_i-n_i)\left(\boldsymbol{x}-\frac{\boldsymbol{n}}{N}\right)^{\boldsymbol{\alpha}}-\phi_{\boldsymbol{n},\boldsymbol{\alpha}}(\boldsymbol{x})\right\|_{W^{1,\infty}([0,1]^d)}\\&\leq 2^d sd^{s-1}B\cdot 3\cdot 2^{2\lceil\log_2(d+s-1)\rceil-1-k}.\end{aligned}$$

Hence, the total approximation error is

$$\begin{aligned}\|f(\boldsymbol{x})-\phi(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}&\leq\|f(\boldsymbol{x})-p(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}+\|p(\boldsymbol{x})-\phi(\boldsymbol{x})\|_{W^{1,\infty}([0,1]^d)}\\&\leq 2^d d^s BN^{-(s-1)}+2^d sd^{s-1}B\cdot 3\cdot 2^{2\lceil\log_2(d+s-1)\rceil-1-k}\\&=2^d d^s B\left(N^{-s+1}+3\cdot\frac{s}{d}\cdot 2^{2\lceil\log_2(d+s-1)\rceil-1-k}\right)\\&=\mathfrak{C}_2(d,s)(N^{-s+1}+2^{-k})B,\end{aligned}$$

where $\mathfrak{C}_2(d,s)=2^d d^s\cdot 3\cdot\frac{s}{d}\cdot 2^{2\lceil\log_2(d+s-1)\rceil-1}$. The width W of the network ϕ is $6(d+s-1)\cdot 2^k\cdot sd^{s-1}(N+1)^d$ and the depth L is $4\lceil\log_2(d+s-1)\rceil+2$. □

Theorem A.2. *For $k\in\mathbb{N}$, $k\geq 2\lceil\log_2(d+s-1)\rceil+2$ and $f\in\mathcal{H}^\beta([0,1]^d,B)$ where $\beta\geq 2$, $s=\lfloor\beta\rfloor\in\mathbb{N}_0$, there exists $\phi\in\mathcal{NN}(W,L)$ where*

$$\begin{aligned}&W=\mathfrak{C}_1(d,s)2^{k+\frac{kd}{s-1}},\\&L=4\lceil\log_2(d+s-1)\rceil+2,\end{aligned}$$

such that

$$\|f-\phi\|_{W^{1,\infty}([0,1]^d)}\leq 2\mathfrak{C}_2(d,s)2^{-k}B,$$

where $\mathfrak{C}_1(d,s)$ and $\mathfrak{C}_2(d,s)$ are defined as Theorem A.1.

Proof. Choose $N=\lceil 2^{\frac{k}{s-1}}\rceil$, by the result in Theorem A.1, we have $W=\mathfrak{C}_1(d,s)2^{k+\frac{kd}{s-1}}$, $L=4\lceil\log_2(d+s-1)\rceil+2$ and $\|f-\phi\|_{W^{1,\infty}([0,1]^d)}\leq 2\mathfrak{C}_2(d,s)2^{-k}B$.

Combining Theorem A.2 with $k\geq\log_2\mathfrak{C}_2(d,s)+1$, we have that $\|f-\phi\|_{W^{1,\infty}([0,1]^d)}\leq B$, and the proof of Theorem 4.1 is complete. □

A.1.3 Proof of Lemma 4.1

We take $\widetilde{Z}=\{\widetilde{Z}_i\}_{i=1}^n$ as an independent copy of $\{Z_i\}_{i=1}^n$, then

$$\begin{aligned}\mathcal{L}(f)-\widehat{\mathcal{L}}(f)&=\mathop{\mathbb{E}}_{Z}(\ell(f,Z))-\frac{1}{n}\sum_{i=1}^n\ell(f,Z_i)\\&=\mathop{\mathbb{E}}_{\{\widetilde{Z}_i\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n\ell(f,\widetilde{Z}_i)-\frac{1}{n}\sum_{i=1}^n\ell(f,Z_i)\right]\\&=\mathop{\mathbb{E}}_{\{\widetilde{Z}_i\}_{i=1}^n}\frac{1}{n}\sum_{i=1}^n\left[\ell(f,\widetilde{Z}_i)-\ell(f,Z_i)\right].\end{aligned}$$

Since ℓ is the least squares loss function, combined with Talagrand's lemma [40], we have

$$\begin{aligned}\mathop{\mathbb{E}}_{\{Z_i\}_{i=1}^n}\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|&\leq\mathop{\mathbb{E}}_{\{Z_i,\widetilde{Z}_i\}_{i=1}^n}\sup_{f\in\mathcal{P}}\left|\frac{1}{n}\sum_{i=1}^n\left[\ell(f,\widetilde{Z}_i)-\ell(f,Z_i)\right]\right|\\&=\mathop{\mathbb{E}}_{\{Z_i,\widetilde{Z}_i,\sigma_i\}_{i=1}^n}\sup_{f\in\mathcal{P}}\frac{1}{n}\sum_{i=1}^n\sigma_i\left[\ell(f,\widetilde{Z}_i)-\ell(f,Z_i)\right]\\&\leq\mathop{\mathbb{E}}_{\{\widetilde{Z}_i,\sigma_i\}_{i=1}^n}\sup_{f\in\mathcal{P}}\frac{1}{n}\sum_{i=1}^n\sigma_i\ell(f,\widetilde{Z}_i)+\mathop{\mathbb{E}}_{\{Z_i,\sigma_i\}_{i=1}^n}\sup_{f\in\mathcal{P}}\frac{1}{n}\sum_{i=1}^n(-\sigma_i)\ell(f,Z_i)\\&=2\mathop{\mathbb{E}}_{\{Z_i,\sigma_i\}_{i=1}^n}\sup_{f\in\mathcal{P}}\frac{1}{n}\sum_{i=1}^n\sigma_i\ell(f,Z_i)\\&=2\mathcal{R}(\ell\circ\mathcal{P})\leq2\mathcal{R}(\mathcal{P}),\end{aligned}$$

where the second step is due to the fact that the insertion of Rademacher variables doesn't change the distribution.

A.1.4 Proof of Theorem 4.3

Lemma A.5. *Let $\mathcal{F}$ be a class of functions from Ω to $\mathbb{R}$ such that $0\in\mathcal{F}$ and the diameter of $\mathcal{F}$ is less than M, i.e., $||u||_{L^\infty(\Omega)}\leq M$, $\forall u\in\mathcal{F}$. Then*

$$\mathcal{R}(\mathcal{F})\leq\inf_{0<\delta<M}\left(4\delta+\frac{12}{\sqrt{N}}\int_\delta^M\sqrt{\log(\mathcal{C}(\epsilon,\mathcal{F},d_\infty))}d\epsilon\right).\tag{A.3}$$

Proof. The proof is based on the chaining method. Set $\epsilon_k=2^{-k+1}M$. We denote by $\mathcal{F}_k$ such that $\mathcal{F}_k$ is an ϵ_k-cover of $\mathcal{F}$ and $|\mathcal{F}_k|=\mathcal{C}(\epsilon_k,\mathcal{F},d_\infty)$. Hence for any $u\in\mathcal{F}$, there exists $u_k\in\mathcal{F}_k$ such that

$$d_\infty(u,u_k)\leq\epsilon_k.$$

Let K be a positive integer determined later. We have

$$
\begin{aligned}
\mathcal{R}(\mathcal{F}) &= \mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\left[\sup_{u\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N\sigma_i u(X_i)\right]\\
&= \mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\frac{1}{N}\sup_{u\in\mathcal{F}}\left[\sum_{i=1}^N\sigma_i(u(X_i)-u_K(X_i))+\sum_{j=1}^{K-1}\sum_{i=1}^N\sigma_i\left(u_{j+1}(X_i)-u_j(X_i)\right)+\sum_{i=1}^N\sigma_i u_1(X_i)\right]\\
&\le \mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\left[\sup_{u\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N\sigma_i(u(X_i)-u_K(X_i))\right]+\sum_{j=1}^{K-1}\mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\left[\sup_{u\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N\sigma_i\left(u_{j+1}(X_i)-u_j(X_i)\right)\right]\\
&\quad+\mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\left[\sup_{u\in\mathcal{F}_1}\frac{1}{N}\sum_{i=1}^N\sigma_i u(X_i)\right].
\end{aligned}
$$

We can choose $\mathcal{F}_1=\{0\}$ to eliminate the third term. For the first term,

$$
\mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\sup_{u\in\mathcal{F}}\frac{1}{N}\left[\sum_{i=1}^N\sigma_i\left(u(X_i)-u_K(X_i)\right)\right]\le\mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\sup_{u\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N|\sigma_i|\,\|u-u_K\|_{L^\infty}\le\epsilon_K.
$$

For the second term, for any fixed samples $\{X_i\}_{i=1}^N$, we define

$$
V_j:=\Big\{\left(u_{j+1}(X_1)-u_j(X_1),\cdots,u_{j+1}(X_N)-u_j(X_N)\right)\in\mathbb{R}^N: u\in\mathcal{F}\Big\}.
$$

Then, for any $v^j\in V_j$,

$$
\begin{aligned}
\left\|v^j\right\|_2&=\left(\sum_{i=1}^n\left|u_{j+1}(X_i)-u_j(X_i)\right|^2\right)^{1/2}\le\sqrt{n}\left\|u_{j+1}-u_j\right\|_{L^\infty}\\
&\le\sqrt{n}\left\|u_{j+1}-u\right\|_{L^\infty}+\sqrt{n}\left\|u_j-u\right\|_{L^\infty}=\sqrt{n}\epsilon_{j+1}+\sqrt{n}\epsilon_j=3\sqrt{n}\epsilon_{j+1}.
\end{aligned}
$$

Applying Massart's lemma [40], we have

$$
\begin{aligned}
&\sum_{j=1}^{K-1}\mathop{\mathbb{E}}_{\{\sigma_i\}_{i=1}^N}\left[\sup_{u\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N\sigma_i\left(u_{j+1}(X_i)-u_j(X_i)\right)\right]\\
=&\sum_{j=1}^{K-1}\mathop{\mathbb{E}}_{\{\sigma_i\}_{i=1}^N}\left[\sup_{v^j\in V_j}\frac{1}{N}\sum_{i=1}^N\sigma_i v_i^j\right]\le\sum_{j=1}^{K-1}\frac{3\epsilon_{j+1}}{\sqrt{N}}\sqrt{2\log|V_j|}.
\end{aligned}
$$

By the definition of V_j, we know that $|V_j|\le|\mathcal{F}_j||\mathcal{F}_{j+1}|\le|\mathcal{F}_{j+1}|^2$. Hence

$$
\sum_{j=1}^{K-1}\mathop{\mathbb{E}}_{\{\sigma_i,X_i\}_{i=1}^N}\left[\sup_{u\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N\sigma_i\left(u_{j+1}(X_i)-u_j(X_i)\right)\right]\le\sum_{j=1}^{K-1}\frac{6\epsilon_{j+1}}{\sqrt{N}}\sqrt{\log|\mathcal{F}_{j+1}|}.
$$

Now we obtain

$$\begin{aligned}\mathcal{R}(\mathcal{F}) &\leq \epsilon_K + \sum_{j=1}^{K-1} \frac{6\epsilon_{j+1}}{\sqrt{N}} \sqrt{\log|\mathcal{F}_{j+1}|}\\ &= \epsilon_K + \frac{12}{\sqrt{N}} \sum_{j=1}^{K} (\epsilon_j - \epsilon_{j+1}) \sqrt{\log \mathcal{C}(\epsilon_j, \mathcal{F}, d_\infty)}\\ &\leq \epsilon_K + \frac{12}{\sqrt{N}} \int_{\epsilon_{K+1}}^{M} \sqrt{\log \mathcal{C}(\epsilon, \mathcal{F}, d_\infty)} d\epsilon\\ &\leq \inf_{0<\delta<M} \left(4\delta + \frac{12}{\sqrt{N}} \int_{\delta}^{M} \sqrt{\log(\mathcal{C}(\epsilon, \mathcal{F}, d_\infty))} d\epsilon \right),\end{aligned}$$

where last inequality holds since for $0 \leq \delta \leq M$, we can choose K to be the largest integer such that $\epsilon_{K+1} > \delta$, at this time $\epsilon_K \leq 4\epsilon_{K+2} \leq 4\delta$. □

Proof of Theorem 4.3. From Lemma 4.1 we have

$$\mathcal{E}_{sta} = \mathop{\mathbb{E}}_{Z} \left[\sup_{f \in \mathcal{P}} |\mathcal{L}(f) - \widehat{\mathcal{L}}(f)| \right] \leq 2\mathcal{R}(\mathcal{P}).$$

Combining with Lemma A.5 and Theorem 4.2, when $d > 2$,

$$\begin{aligned}\mathcal{R}(\mathcal{P}) &\leq \inf_{0<\delta<2B} \left(4\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{2B} \sqrt{\log(\mathcal{C}(\epsilon, \mathcal{P}, d_\infty))} d\epsilon \right)\\ &\leq \inf_{0<\delta<2B} \left(4\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{2B} \sqrt{\mathfrak{C}(d,B)\epsilon^{-d}} d\epsilon \right)\\ &\leq \inf_{0<\delta<2B} \left(4\delta + \frac{12}{\sqrt{n}} \mathfrak{C}(d,B) \frac{2}{d-2} \left(\delta^{-\frac{d}{2}+1} - (2B)^{-\frac{d}{2}+1} \right) \right)\\ &\leq \inf_{0<\delta<2B} \left(4\delta + \frac{\mathfrak{C}(d,B)}{\sqrt{n}} \delta^{-\frac{d}{2}+1} \right)\\ &\leq \mathfrak{C}(d,B) n^{-\frac{1}{d}}.\end{aligned}$$

Therefore, $\mathcal{E}_{sta} \leq 2\mathfrak{C}(d,B) n^{-\frac{1}{d}} = \mathfrak{C}(d,B) n^{-\frac{1}{d}}$. □

A.1.5 Proof of Theorem 4.4

Firstly, since $f_0 \in \mathcal{H}^\beta([0,1]^d, B)$, by Corollary 4.1 we know that for any $k \geq \log_2 \mathfrak{C}_2(d,s) + 1$, there exists a neural network $f_{\boldsymbol{\theta}} \in \mathcal{P}$ with $W = \mathfrak{C}_1(d,s) 2^{k + \frac{kd}{s-1}}$ and $L = 4\lceil \log_2(d+s-1) \rceil + 2$ such that

$$\|f_0 - f_{\boldsymbol{\theta}}\|_{W^{1,\infty}([0,1]^d)} \leq 2\mathfrak{C}_2(d,s) B \cdot 2^{-k}.$$

Secondly, by Theorem 4.2, we have that the statistic error

$$\mathop{\mathbb{E}}_{S_n} \left[\sup_{f \in \mathcal{P}} |\mathcal{L}(f) - \widehat{\mathcal{L}}(f)| \right] \leq \mathfrak{C}(d,B) n^{-\frac{1}{d}}.$$

Finally, take $(2\mathfrak{C}_2(d,s)B\cdot 2^{-k})^2 \le \mathfrak{C}(d,B)n^{-\frac{1}{d}}$, we have that $k \ge -\frac{1}{2}\log_2(\mathfrak{C}(d,B,s)n^{-\frac{1}{d}}) = \mathfrak{C}(d,B,s)\log_2 n$. Hence take $k \ge \max\{\log_2 \mathfrak{C}_2(d,s)+1, \mathfrak{C}(d,B,s)\log_2 n\}$, and by Proposition 4.1 the total error is

$$\mathop{\mathbb{E}}_{S_n}\left[\|\widehat{f}_{\boldsymbol{\theta}}-f_0\|^2_{L^2(\mu)}\right] \le \mathfrak{C}(s,d,B)n^{-\frac{1}{d}}.$$

A.2 Proof of Deep Ritz Method

In this section, $\mathcal{P}$ consists of the feed-forward neural networks with the *Sigmoidal* activation function.

A.2.1 Proof of Proposition 4.2

For any $u\in\mathcal{P}$, set $v=u-u^*$, then

$$\begin{aligned}
\mathcal{L}(u) =& \mathcal{L}(u^*+v)\\
=&\frac{1}{2}(\nabla(u^*+v),\nabla(u^*+v))_{L^2(\Omega)}+\frac{1}{2}(u^*+v,u^*+v)_{L^2(\Omega;w)}-\langle u^*+v,f\rangle_{L^2(\Omega)}\\
=&\frac{1}{2}(\nabla u^*,\nabla u^*)_{L^2(\Omega)}+\frac{1}{2}(u^*,u^*)_{L^2(\Omega;w)}-\langle u^*,f\rangle_{L^2(\Omega)}\\
&+\frac{1}{2}(\nabla v,\nabla v)_{L^2(\Omega)}+\frac{1}{2}(v,v)_{L^2(\Omega;w)}+\left[(\nabla u^*,\nabla v)_{L^2(\Omega)}+(u^*,v)_{L^2(\Omega;w)}-\langle v,f\rangle_{L^2(\Omega)}\right]\\
=&\mathcal{L}(u^*)+\frac{1}{2}(\nabla v,\nabla v)_{L^2(\Omega)}+\frac{1}{2}(v,v)_{L^2(\Omega;w)},
\end{aligned}$$

where the last equality is due to the fact that u^* is the weak solution of Eq. equation 3.4. Hence

$$\begin{aligned}
\frac{c_1\wedge 1}{2}\|v\|^2_{H^1(\Omega)} \le \mathcal{L}(u)-\mathcal{L}(u^*) = \frac{1}{2}(\nabla v,\nabla v)_{L^2(\Omega)}+\frac{1}{2}(v,v)_{L^2(\Omega;w)}\\
\le \frac{\|w\|_{L^\infty(\Omega)}\vee 1}{2}\|v\|^2_{H^1(\Omega)},
\end{aligned}$$

that is,

$$\frac{c_1\wedge 1}{2}\|u-u^*\|^2_{H^1(\Omega)} \le \mathcal{L}(u)-\mathcal{L}(u^*) \le \frac{\|w\|_{L^\infty(\Omega)}\vee 1}{2}\|u-u^*\|^2_{H^1(\Omega)}.$$

A.2.2 Proof of Theorem 4.7

Since $\Omega=[0,1]^d$, by Lemma 4.4, we have

$$\mathcal{E}_{sta} \le \mathfrak{C}(d,M,|\partial\Omega|,|\Omega|)\sum_{j=1}^4 \mathcal{R}(\mathcal{F}_j).$$

Since $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^2},2B)$ and the activation function is infinitely differentiable, for all j, $1\le j\le 4$, $\mathcal{F}_j\in\mathcal{M}_0 := \{f\in C^1 : \|f\|_{C^1}\le 2B\}$. Then by Theorem 4.6, $\log\mathcal{C}(\epsilon,\mathcal{M}_0,d_\infty)\le$

$\mathfrak{C}(d,B)\epsilon^{-d}$. By Lemma A.5 and Theorem 4.6, when $d>2$,

$$\begin{aligned}\mathcal{R}(\mathcal{F}_j) &\leq \inf_{0<\delta<2B}\left(4\delta+\frac{12}{\sqrt{n}}\int_{\delta}^{2B}\sqrt{\log(\mathcal{C}(\epsilon,\mathcal{M}_0,d_\infty))}d\epsilon\right)\\ &\leq \inf_{0<\delta<2B}\left(4\delta+\frac{12}{\sqrt{n}}\int_{\delta}^{2B}\sqrt{\mathfrak{C}(d,B)\epsilon^{-d}}d\epsilon\right)\\ &\leq \inf_{0<\delta<2B}\left(4\delta+\frac{12}{\sqrt{n}}\mathfrak{C}(d,B)\frac{2}{d-2}\left(\delta^{-\frac{d}{2}+1}-(2B)^{-\frac{d}{2}+1}\right)\right)\\ &\leq \inf_{0<\delta<2B}\left(4\delta+\frac{\mathfrak{C}(d,B)}{\sqrt{n}}\delta^{-\frac{d}{2}+1}\right)\\ &\leq \mathfrak{C}(d,B)n^{-\frac{1}{d}}.\end{aligned}$$

Therefore,

$$\mathcal{E}_{sta}=2\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]\leq\mathfrak{C}(d,B,M)n^{-\frac{1}{d}}.$$

A.2.3 Proof of Theorem 4.8

Firstly, by Corollary 4.3 we know that the assumption $\mathcal{P}=\mathcal{N}\mathcal{N}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^2},2B)$ is reasonable and for any $0<\epsilon<B$, there exists a neural network $u_{\boldsymbol{\theta}}\in\mathcal{P}$ with depth L and total number of nonzero weights $\mathfrak{n}_{\boldsymbol{\theta}}\geq\mathfrak{C}(d,\mu,p,B)\epsilon^{-\frac{d}{1-3\mu}}$ such that

$$\|u^*-u_{\boldsymbol{\theta}}\|_{W^{2,\infty}(\Omega)}\leq\epsilon<B,$$

where μ is an arbitrarily small positive number and L depends on d,μ,p,B.

Secondly, by Theorem 4.7,

$$\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]\leq\mathfrak{C}(d,B,M)n^{-\frac{1}{d}}.$$

Finally, take $\epsilon^2=\mathfrak{C}(d,B,M)n^{-\frac{1}{d}}$, we have that $\mathfrak{n}_{\boldsymbol{\theta}}\geq\mathfrak{C}(d,\mu,B,M,p)n^{\frac{1}{2(1-3\mu)}}$. On the other hand, since $\epsilon<B$, $\mathfrak{n}_{\boldsymbol{\theta}}\geq\mathfrak{C}(d,\mu,p,B)\epsilon^{-\frac{d}{1-3\mu}}\geq\mathfrak{C}(d,\mu,p,B)B^{-\frac{d}{1-3\mu}}$.

Hence take $\mathfrak{n}_{\boldsymbol{\theta}}\geq\max\{\mathfrak{C}(d,\mu,B,M,p)n^{\frac{1}{2(1-3\mu)}},\mathfrak{C}(d,\mu,B,p)B^{-\frac{d}{1-3\mu}}\}$, by Proposition 4.3, the total error is

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[||\widehat{u}_{\boldsymbol{\theta}}-u^*||^2_{H^1(\Omega)}\right]\\
\leq&\frac{2}{c_1\wedge 1}\Bigg\{\underbrace{\frac{M\vee 1}{2}\inf_{u\in\mathcal{P}}||u-u^*||^2_{H^1(\Omega)}}_{\mathcal{E}_{app}}+2\underbrace{\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]}_{\mathcal{E}_{sta}}\Bigg\}\\
\leq&\frac{2}{c_1\wedge 1}\Bigg\{\underbrace{\frac{M\vee 1}{2}\inf_{u\in\mathcal{P}}||u-u^*||^2_{W^{2,\infty}(\Omega)}}_{\mathcal{E}_{app}}+2\underbrace{\mathop{\mathbb{E}}_{\{X_k\}_{k=1}^{N_{in}},\{Y_k\}_{k=1}^{N_b}}\left[\sup_{u\in\mathcal{P}}|\mathcal{L}(u)-\widehat{\mathcal{L}}(u)|\right]}_{\mathcal{E}_{sta}}\Bigg\}\\
\leq&\mathfrak{C}(d,\mu,B,M)n^{-\frac{1}{d}}.
\end{aligned}
$$

A.3 Proof of smoothness

A.3.1 Proof of Theorem 4.9

For the regression model, by Lemma 4.1, we have $\mathcal{E}_{sta}\leq 2\mathcal{R}(\mathcal{P})$. Since $\mathcal{P}=\mathcal{NN}(L,\mathfrak{n}_{\boldsymbol{\theta}},\|\cdot\|_{C^t},2B)$, by Theorem 4.6, $\log\mathcal{C}(\epsilon,\mathcal{P},d_\infty)\leq\mathfrak{C}(d,B,t)\epsilon^{-\frac{d}{t}}$. By Lemma A.5 when $t<d/2$,

$$
\begin{aligned}
\mathcal{R}(\mathcal{P})&\leq\inf_{0<\delta<2B}\left(4\delta+\frac{12}{\sqrt{n}}\int_\delta^{2B}\sqrt{\log(\mathcal{C}(\epsilon,\mathcal{P},d_\infty))d\epsilon}\right)\\
&\leq\inf_{0<\delta<2B}\left(4\delta+\frac{12}{\sqrt{n}}\int_\delta^{2B}\sqrt{\mathfrak{C}(d,B,t)\epsilon^{-\frac{d}{t}}d\epsilon}\right)\\
&\leq\inf_{0<\delta<2B}\left(4\delta+\frac{12}{\sqrt{n}}\mathfrak{C}(d,B,t)\frac{2t}{d-2t}\left(\delta^{-\frac{d}{2t}+1}-(2B)^{-\frac{d}{2t}+1}\right)\right)\\
&\leq\inf_{0<\delta<2B}\left(4\delta+\frac{\mathfrak{C}(d,B,t)}{\sqrt{n}}\delta^{-\frac{d}{2t}+1}\right)\\
&\leq\mathfrak{C}(d,B,t)n^{-\frac{t}{d}}.
\end{aligned}
$$

Therefore,

$$
\mathcal{E}_{sta}=\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]\leq\mathfrak{C}(d,B,t)n^{-\frac{t}{d}}.
$$

A.3.2 Proof of Theorem 4.10

Firstly, since $f_0\in\mathcal{H}^\beta([0,1]^d,B)$, $\beta\geq 2$ and $1\leq t\leq\eta\leq s-1$, by Corollary 4.4 we know that for any $0<\epsilon<B$, there exists a neural network $f_{\boldsymbol{\theta}}\in\mathcal{P}$ with depth L and total number of nonzero weights $\mathfrak{n}_{\boldsymbol{\theta}}\geq\mathfrak{C}(d,\mu,B,s,\eta)\epsilon^{-\frac{d}{s-\eta-\mu\eta}}$ such that

$$
\|f_0-f_{\boldsymbol{\theta}}\|_{W^{t,\infty}([0,1]^d)}\leq\|f_0-f_{\boldsymbol{\theta}}\|_{W^{\eta,\infty}([0,1]^d)}\leq\epsilon<B,
$$

where μ is an arbitrarily small positive number and L depends on d,μ,B,s.

Secondly, by Theorem 4.9, we have that the statistic error satisfies

$$\mathop{\mathbb{E}}_{S_n}\left[\sup_{f\in\mathcal{P}}|\mathcal{L}(f)-\widehat{\mathcal{L}}(f)|\right]\le\mathfrak{C}(d,B,t)n^{-\frac{t}{d}}.$$

Finally, take $\epsilon^2=\mathfrak{C}(d,B,t)n^{-\frac{t}{d}}$, we have that $\mathfrak{n}_{\boldsymbol{\theta}}\ge\mathfrak{C}(d,\mu,B,s,\eta,t)n^{\frac{t}{2(s-\eta-\mu\eta)}}$.

On the other hand, since $\epsilon<B$, $\mathfrak{n}_{\boldsymbol{\theta}}\ge\mathfrak{C}(d,\mu,B,s,\eta)\epsilon^{-\frac{d}{s-\eta-\mu\eta}}\ge\mathfrak{C}(d,\mu,B,s,\eta)B^{-\frac{d}{s-\eta-\mu\eta}}$. Hence take

$$\mathfrak{n}_{\boldsymbol{\theta}}\ge\max\{\mathfrak{C}(d,\mu,B,s,t,\eta)n^{\frac{t}{2(s-\eta-\mu\eta)}},\mathfrak{C}(d,\mu,B,s,\eta)B^{-\frac{d}{s-\eta-\mu\eta}}\},$$

and by Proposition 4.1, the total error is

$$\mathop{\mathbb{E}}_{S_n}\left[\mathcal{L}(\widehat{f}_{\boldsymbol{\theta}})-\mathcal{L}(f_0)\right]\le\mathfrak{C}(s,d,B,t,\eta,\mu)n^{-\frac{t}{d}}.$$

References

[1] Shmuel Agmon, Avron Douglis, and Louis Nirenberg. Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions. I. *Communications on Pure and Applied Mathematics*, 12(4):623–727, 1959.

[2] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[3] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[4] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

[5] Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

[6] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[8] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

[9] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, 2019.

[10] Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 2021.

[11] Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4):1203–1253, 2022.

[12] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pages 1675–1685. PMLR, 2019.

[14] Chenguang Duan, Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, and Jerry Zhijian Yang. Convergence rate analysis for deep Ritz method. *Communications in Computational Physics*, 31(4):1020–1048, 2022.

[15] Weinan E, Chao Ma, and Lei Wu. A priori estimates of the population risk for two-layer neural networks. *arXiv preprint arXiv:1810.06397*, 2018.

[16] Jianqing Fan, Yihong Gu, and Wen-Xin Zhou. How do noise tails impact on deep ReLU networks? *arXiv preprint arXiv:2203.10418*, 2022.

[17] Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

[18] David Gilbarg, Neil S Trudinger, David Gilbarg, and NS Trudinger. *Elliptic Partial Differential Equations of Second Order*, volume 224. Springer, 1977.

[19] Evarist Gin and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, USA, 1st edition, 2015.

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

[21] Sean Hon and Haizhao Yang. Simultaneous neural network approximation for smooth functions, 2021.

[22] Sean Hon and Haizhao Yang. Simultaneous neural network approximations in Sobolev spaces. *arXiv preprint arXiv:2109.00161*, 2021.

[23] Qingguo Hong, Jonathan W Siegel, and Jinchao Xu. Rademacher complexity and numerical quadrature analysis of stable neural networks with applications to numerical PDEs. *arXiv preprint arXiv:2104.02903*, 2021.

[24] Martin Hutzenthaler, Arnulf Jentzen, Thomas Kruse, Tuan Anh Nguyen, and Philippe von Wurstemberger. Overcoming the curse of dimensionality in the numerical approximation of semilinear parabolic partial differential equations. *Proceedings of the Royal Society A*, 476(2244):20190630, 2020.

[25] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[26] Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, Fengru Wang, Jerry Zhijian Yang, et al. A rate of convergence of physics informed neural networks for the linear second order elliptic PDEs. *Communications in Computational Physics*, 31(4):1272–1295, 2022.

[27] Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv preprint arXiv:2104.06708*, 2021.

[28] Yuling Jiao, Yang Wang, and Yunfei Yang. Approximation bounds for norm constrained neural networks with applications to regression and gans. *arXiv preprint arXiv:2201.09418*, 2022.

[29] Michael Kohler and Adam Krzyzak. Over-parametrized deep neural networks minimizing the empirical risk do not generalize well. *Bernoulli*, 27(4):2564–2597, 2021.

[30] Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

[31] Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A theoretical

analysis of deep neural networks and parametric PDEs. *arXiv preprint arXiv:1904.00377*, 2019.

[32] Samuel Lanthaler, Siddhartha Mishra, and George Em Karniadakis. Error estimates for DeepONets: A deep learning framework in infinite dimensions. *arXiv preprint arXiv:2102.09618*, 2021.

[33] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[34] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 2022.

[35] Jianfeng Lu, Yulong Lu, and Min Wang. A priori generalization analysis of the deep Ritz method for solving high dimensional elliptic equations. *arXiv preprint arXiv:2101.01708*, 2021.

[36] Yiping Lu, Haoxuan Chen, Jianfeng Lu, Lexing Ying, and Jose Blanchet. Machine learning for elliptic PDEs: Fast rate generalization bound, neural scaling law and minimax optimality. *ICLR*, 2021.

[37] Tao Luo and Haizhao Yang. Two-layer neural networks for partial differential equations: Optimization and generalization theory. *arXiv preprint arXiv:2006.15733*, 2020.

[38] Siddhartha Mishra and Roberto Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs. *IMA Journal of Numerical Analysis*, 42(2):981–1022, 2022.

[39] Siddhartha Mishra and T Konstantin Rusch. Enhancing accuracy of deep learning algorithms by training with low-discrepancy sequences. *arXiv preprint arXiv:2005.12564*, 2020.

[40] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2018.

[41] Johannes Müller and Marius Zeinhofer. Error estimates for the variational training of neural networks with boundary penalty. *arXiv preprint arXiv:2103.01007*, 2021.

[42] Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21:174–1, 2020.

[43] Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

[44] Johannes Schmidt-Hieber et al. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

[45] Guohao Shen, Yuling Jiao, Yuanyuan Lin, and Jian Huang. Robust nonparametric regression with deep neural networks. *arXiv preprint arXiv:2107.10343*, 2021.

[46] Yeonjong Shin, Jerome Darbon, and George Em Karniadakis. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. *arXiv preprint arXiv:2004.01806*, 2020.

[47] Hwijae Son, Jin Woo Jang, Woo Jin Han, and Hyung Ju Hwang. Sobolev training for the neural network solutions of PDEs. *arXiv preprint arXiv:2101.08932*, 2021.

[48] Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

[49] Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *Advances in Neural Information Processing Systems*, 34, 2021.

[50] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint*

arXiv:2009.14286, 2020.

[51] Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *arXiv preprint arXiv:2007.14527*, 2020.

[52] E Weinan, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, pages 1–24, 2020.

[53] E Weinan and Bing Yu. The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

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

[55] Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.