Deeper or Wider: A Perspective from Optimal Generalization Error with Sobolev Loss

Yahong Yang [1] **Juncai He** [2]

Abstract

Constructing the architecture of a neural network is a challenging pursuit for the machine learning community, and the dilemma of whether to go deeper or wider remains a persistent question. This paper explores a comparison between deeper neural networks (DeNNs) with a flexible number of layers and wider neural networks (WeNNs) with limited hidden layers, focusing on their optimal generalization error in Sobolev losses. Analytical investigations reveal that the architecture of a neural network can be significantly influenced by various factors, including the number of sample points, parameters within the neural networks, and the regularity of the loss function. Specifically, a higher number of parameters tends to favor WeNNs, while an increased number of sample points and greater regularity in the loss function lean towards the adoption of DeNNs. We ultimately apply this theory to address partial differential equations using deep Ritz and physics-informed neural network (PINN) methods, guiding the design of neural networks.

1. Introduction

Recently, Sobolev training (Czarnecki et al., 2017; Son et al., 2021; Vlassis & Sun, 2021) for neural networks (NNs) with the rectified linear unit (ReLU) activation function has improved model prediction and generalization capabilities by incorporating derivatives of the target output with respect to the model input. This training approach has proven impactful in various areas, including the solution of partial differential equations (Lagaris et al., 1998; E et al., 2017; Raissi et al., 2019; De Ryck & Mishra, 2022), operator learning (Lu et al., 2021b), network compression (Sau & Balasubramanian, 2016), distillation (Hinton et al., 2015; Rusu et al., 2015), regularization (Czarnecki et al., 2017), and dynamic programming (Finlay et al., 2018; Werbos, 1992), among others.

In Sobolev training, two architectural parameters of fully connected neural networks have been considered: L, the number of hidden layers, and N, the number of neurons in each layer. In this paper, we consider the following two different NNs with a fixed number of parameters $W = \mathcal{O}(N^2L)$:

- Wider neural networks (WeNNs) with a limited number of hidden layers but flexible width, namely, $L = \mathcal{O}(1)$ or $\mathcal{O}(\log(N))$;
- Deeper neural networks (DeNNs) with a flexible number of layers but limited width: $N = \mathcal{O}(1)$ or $\mathcal{O}(\log(L))$.

In the literature, WeNNs, as evidenced by recent works such as (Schmidt-Hieber, 2020; Suzuki, 2018; Ma et al., 2022a; Barron, 1993; Mhaskar, 1996; Ma et al., 2022b; Siegel & Xu, 2022; Montanelli & Du, 2019), demonstrate efficacy in tasks that require shallower architectures. In contrast, DeNNs, supported by recent research (Yarotsky, 2017; Yarotsky & Zhevnerchuk, 2020; Siegel, 2022; Yang et al., 2023b;a; He, 2023; Yang & Lu, 2023), exhibit extensive capacity for intricate representations and excel in handling complex computations.

These two categories of neural networks have distinct advantages and disadvantages. WeNNs boast a simple structure that facilitates easy training and reduces the risk of overfitting. However, this simplicity comes at a cost, as the limited richness within this space constrains their approximation capability. Consequently, achieving a satisfactory approximation necessitates an increased number of parameters in the network structure. On the other hand, DeNNs offer a high-complexity space, enabling accurate approximations with fewer parameters. Works such as (Yarotsky, 2017; Siegel, 2022; Yang et al., 2023b;a; He, 2023; Yang & Lu, 2023) demonstrate that DeNNs can achieve significantly

[1]Department of Mathematics, The Pennsylvania State University, University Park, State College, PA 16802, USA [2]Computer, Electrical and Mathematical Science and Engineering Division, The King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia. Correspondence to: Yahong Yang <yxy5498@psu.edu>, Juncai He <juncai.he@kaust.edu.sa>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

better approximations than shallow neural networks and traditional methods. This improved approximation rate is referred to as super-convergence. However, these studies do not explicitly compare the generalization error with shallow neural networks. Nevertheless, the heightened complexity of DeNNs introduces challenges in the learning process. Training such networks is not a straightforward task and demands a more extensive set of sample points to effectively capture the intricate relationships within the data.

In this paper, our objective is to conduct a comprehensive analysis of the optimal generalization errors between these two types of NNs in Sobolev training setup. While existing literature, such as (Schmidt-Hieber, 2020; Suzuki, 2018), has extensively explored WeNNs, our focus is on the generalization error of DeNNs, breaking it down into two integral components: the approximation error and the sampling error. Despite previous studies, including (Duan et al., 2021; Jakubovitz et al., 2019; Advani et al., 2020; Berner et al., 2020), analyzing the generalization error of DeNNs, these works impose a constraint requiring the parameters in DeNNs to be uniformly bounded. This constraint limits the complexity of DeNNs and compromises the approximation advantages inherent in DeNNs. Notably, without such a bounded restriction, DeNNs can achieve a superior approximation rate compared to traditional NNs (Yarotsky, 2017; Siegel, 2022; Yang et al., 2023b;a; He, 2023; Yang & Lu, 2023). Furthermore, our result is different from (Yang et al., 2023b;a), which is not nearly optimal in those papers. Moreover, in this paper, we will extend the generalization error analysis to Sobolev training to further understand the capabilities of DeNNs in a more general setting.

We investigate the approximation of the target function $f(\boldsymbol{x})$ defined on $[0,1]^d$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$ using a finite set of sample data $\{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^M$. We employ DeNNs with a width of N and a depth of L, leading to a total number of parameters denoted as $W := \mathcal{O}(N^2L)$. The main results of this paper are summarized in Table 1 and Figure 1. Table 1 lists the approximation and sampling errors for both WeNNs and DeNNs with different loss functions. Fig. 1 summarizes the genearlization error with different soloblev loss functions with respect to the number of parameters W and the number of sample points M. For instance, the blue curve represents the generalization error with L^2 loss function. When a pair (W, M) is located to the left of this curve, we demonstrate that the DeNNs performs better. Conversely, when the pair (W, M) is situated to the right of this curve, WeNNs are more effective. Similar curves are drawn for H^k loss functions, revealing that WeNNs perform better as the loss function requires higher regularity. Here, the H^1 loss is equivalent to the deep Ritz method, and the H^2 loss represents the PINN method.

Based on Fig. 1, valuable insights for selecting neural networks can be gained. If the number of parameters in our neural networks is fixed, the choice between WeNNs and DeNNs depends on the availability of sample points. Opting for WeNNs is advisable when the number of sample points is limited, whereas choosing DeNNs becomes preferable when a substantial number of sample points are available. Conversely, when the number of sample points is fixed, the neural network architecture will depend on the number of parameters. DeNNs are preferable if one seeks fewer parameters, while WeNNs are more suitable when a larger number of parameters is acceptable.

Moreover, it is evident that with an increase in the order of derivatives in the loss functions, the effective region for DeNNs expands. In cases where the pair (W, M) falls within transitional areas between the two curves, enhancing the depth of the neural network is advisable to improve generalization accuracy.

We highlight the contributions of the present paper as follows:

• We establish the optimal generalization error by dissecting it into two components—approximation error and sampling error—for DeNNs. Furthermore, we extend this result to Sobolev training, specifically involving loss functions defined by H^1 and H^2 norms. Our result differs from existing works (Schmidt-Hieber, 2020; Suzuki, 2018) which only consider the generalization error of WNNs. (Duan et al., 2021; Yang et al., 2023b;a) obtain the generalization error using Rademacher complexity, which is suboptimal. (Shi et al., 2024; Zhang et al., 2023; Feng et al., 2021) do not analyze the generalization error for Sobolev training.

• We compare the optimal generalization error of neural networks (Schmidt-Hieber, 2020; Suzuki, 2018) with our findings and conclude that DeNNs perform better than WeNNs when the sample points are abundant but the number of parameters is limited. Conversely, WeNNs are superior to DeNNs if the number of parameters is large but the number of sample points is limited. Furthermore, as the required order of the derivative in the loss function increases, WeNNs may transition to DeNNs, influencing the performance of NNs in generalization error.

• We utilize our findings to analyze the generalization error of deep Ritz and PINN methods when employing DeNNs with a flexible number of layers. Our approach surpasses existing work, as prior studies have solely focused on the approximation and sample errors for neural networks with a limited number of hidden layers.

1.1. Organization of the paper

The remainder of the paper unfolds as follows: Section 2 provides a compilation of useful notations and definitions related to Sobolev spaces. Subsequently, in Section 3, we

	WeNNs		DeNNs	
	Approximation	Sampling (Optimal)	Approximation (Optimal)	Sampling (Optimal)
L^2-loss	$\mathcal{O}(W^{-\frac{n}{d}})$	$\mathcal{O}(W/M)$	$\mathcal{O}(W^{-\frac{2n}{d}})$	$\mathcal{O}(W^2/M)$
H^1-loss	$\mathcal{O}(W^{-\frac{n-1}{d}})$	$\mathcal{O}(W/M)$	$\mathcal{O}(W^{-\frac{2n-2}{d}})$	$\mathcal{O}(W^2/M)$
H^2-loss	$\mathcal{O}(W^{-\frac{n-2}{d}})$	$\mathcal{O}(W/M)$	$\mathcal{O}(W^{-\frac{2n-4}{d}})$	$\mathcal{O}(W^2/M)$

Table 1. The approximation and sampling errors for both WeNNs and DeNNs (W represents the number of parameters, defined as $W := \mathcal{O}(N^2L)$ with a width of N and a depth of L, while M denotes the number of sample points. All values in the table are up to a logarithmic factor. The results for both WeNNs and DeNNs are near optimality in terms of the sample number (Schmidt-Hieber, 2020; Suzuki, 2018). Additionally, the results for DeNNs approach near optimality in terms of the parameter count W (Lu et al., 2021a; Siegel, 2022; Yang et al., 2023b;a).

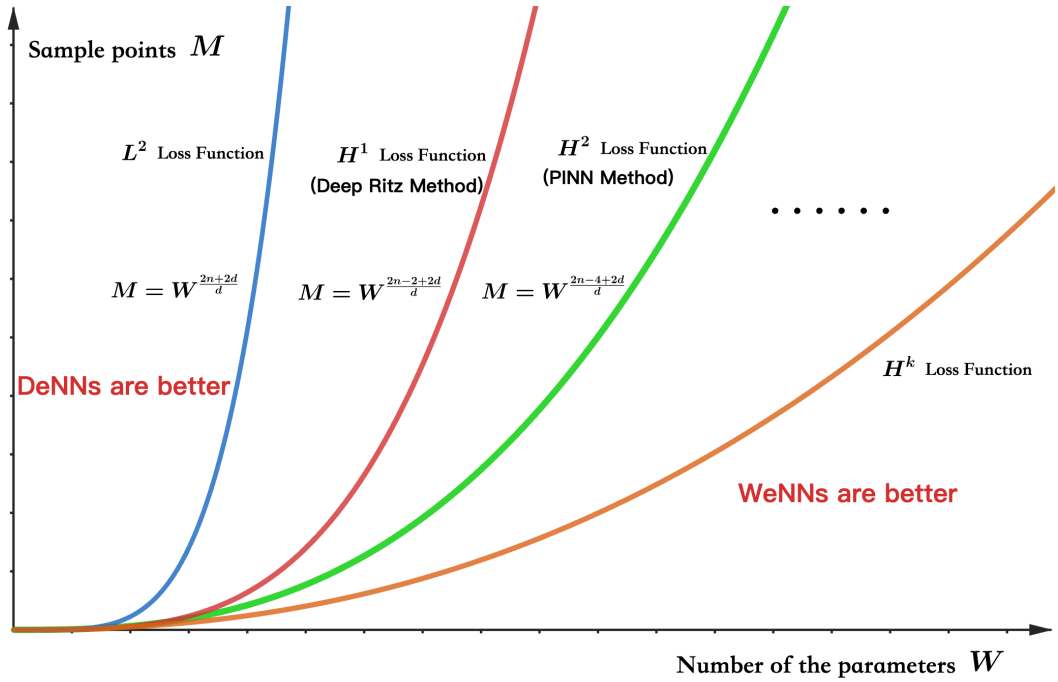

Figure 1. Generalization error with respect to the number of parameters, W, and sample points, M, for L^2 loss functions, H^1 loss functions, H^2 loss functions, and H^k loss functions with $k \geq 3$ ($W = \mathcal{O}(N^2L)$, where N is the width of NNs and L is the depth of NNs).

delve into establishing the optimal generalization error of DeNNs with respect to the H^k norms for $k = 0, 1, 2$. In Section 4, we apply our results to the deep Ritz method and PINN methods.

2. Preliminaries

2.1. Neural networks

Let us summarize all basic notations used in the DeNNs as follows:

1. Matrices are denoted by bold uppercase letters. For example, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a real matrix of size $m \times n$ and $\boldsymbol{A}^\intercal$ denotes the transpose of $\boldsymbol{A}$. Vectors are denoted by bold lowercase letters. For example, $\boldsymbol{v} \in \mathbb{R}^n$ is a column vector of size n.

2. For a d-dimensional multi-index $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots \alpha_d] \in \mathbb{N}^d$, we denote several related notations as follows: (a) $|\boldsymbol{\alpha}| = |\alpha_1| + |\alpha_2| + \cdots + |\alpha_d|$; (b) $\boldsymbol{x}^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$, $\boldsymbol{x} = [x_1, x_2, \cdots, x_d]^\intercal$; (c) $\boldsymbol{\alpha}! = \alpha_1!\alpha_2! \cdots \alpha_d!$.

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

4. Define $\sigma_1(x) := \sigma(x) = \max\{0, x\}$ and $\sigma_2 := \sigma^2(x)$. We call the neural networks with activation function σ_t with $t \leq i$ as σ_i neural networks (σ_i-NNs), $i = 1, 2$. With the abuse of notations, we define $\sigma_i : \mathbb{R}^d \to \mathbb{R}^d$ as $\sigma(\boldsymbol{x}) =$

$[\sigma(x_1), \cdots, \sigma(x_d)]^\intercal \in \mathbb{R}^d$. for any $\boldsymbol{x} = [x_1, \cdots, x_d]^\intercal$.

5. Define $L, N \in \mathbb{N}_+$, $N_0 = d$ and $N_{L+1} = 1$, $N_i \in \mathbb{N}_+$ for $i = 1, 2, \ldots, L$, then a σ_i-NN ϕ with the width N and depth L can be described as follows:

$$\boldsymbol{x} = \tilde{\boldsymbol{h}}_0 \xrightarrow{W_1, b_1} \boldsymbol{h}_1 \xrightarrow{\sigma_i} \tilde{\boldsymbol{h}}_1 \ldots \xrightarrow{\sigma_i} \tilde{\boldsymbol{h}}_L \xrightarrow{W_{L+1}, b_{L+1}} \phi(\boldsymbol{x}),$$

where $\boldsymbol{W}_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the i-th linear transform in ϕ, respectively, i.e., $\boldsymbol{h}_i := \boldsymbol{W}_i \tilde{\boldsymbol{h}}_{i-1} + \boldsymbol{b}_i$, for $i = 1, \ldots, L+1$ and $\tilde{\boldsymbol{h}}_i = \sigma_i(\boldsymbol{h}_i)$, for $i = 1, \ldots, L$. In this paper, an DeNN with the width N and depth L, means (a) The maximum width of this DeNN for all hidden layers is less than or equal to N. (b) The number of hidden layers of this DeNN is less than or equal to L.

2.2. Sobolev spaces and covering number

Denote Ω as $[0,1]^d$, D as the weak derivative of a single variable function, $D^{\boldsymbol{\alpha}} = D_1^{\alpha_1} D_2^{\alpha_2} \ldots D_d^{\alpha_d}$ as the partial derivative of a multivariable function, where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_d]^\intercal$ and D_i is the derivative in the i-th variable.

Definition 2.1 (Sobolev Spaces (Evans, 2022))**.** Let $n \in \mathbb{N}$ and $1 \le p \le \infty$. Then we define Sobolev spaces

$$W^{n,p}(\Omega) := \{f \in L^p(\Omega) : D^{\boldsymbol{\alpha}} f \in L^p(\Omega) \text{ for all } |\boldsymbol{\alpha}| \le n\}$$

with $\|f\|_{W^{n,p}(\Omega)} := \left(\sum_{0 \le |\alpha| \le n} \|D^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)}\right)^{1/p}$, if $p < \infty$, and $\|f\|_{W^{n,\infty}(\Omega)} := \max_{0 \le |\alpha| \le n} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)}$. Furthermore, for $\boldsymbol{f} = (f_1, f_2, \ldots, f_d)$, $\boldsymbol{f} \in W^{1,\infty}(\Omega, \mathbb{R}^d)$ if and only if $f_i \in W^{1,\infty}(\Omega)$ for each $i = 1, 2, \ldots, d$ and $\|\boldsymbol{f}\|_{W^{1,\infty}(\Omega, \mathbb{R}^d)} := \max_{i=1,\ldots,d} \{\|f_i\|_{W^{1,\infty}(\Omega)}\}$. When $p = 2$, denote $W^{n,2}(\Omega)$ as $H^n(\Omega)$ for $n \in \mathbb{N}_+$.

Definition 2.2 (covering number (Anthony et al., 1999))**.** Let $(V, \|\cdot\|)$ be a normed space, and $\Theta \in V$. $\{V_1, V_2, \ldots, V_n\}$ is an ε-covering of Θ if $\Theta \subset \cup_{i=1}^n B_{\varepsilon, \|\cdot\|}(V_i)$. The covering number $\mathcal{N}(\varepsilon, \Theta, \|\cdot\|)$ is defined as $\mathcal{N}(\varepsilon, \Theta, \|\cdot\|) := \min\{n : \exists \varepsilon\text{-covering over } \Theta \text{ of size } n\}$.

Definition 2.3 (Uniform covering number (Anthony et al., 1999))**.** Suppose the $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $\mathbb{R}$. Given n samples $\boldsymbol{Z}_n = (z_1, \ldots, z_n) \in \mathcal{X}^n$, define

$$\mathcal{F}|_{\boldsymbol{Z}_n} = \{(u(z_1), \ldots, u(z_n)) : u \in \mathcal{F}\}.$$

The uniform covering number $\mathcal{N}(\varepsilon, \mathcal{F}, n)$ is defined as

$$\mathcal{N}(\varepsilon, \mathcal{F}, n) = \max_{\boldsymbol{Z}_n \in \mathcal{X}^n} \mathcal{N}\left(\varepsilon, \mathcal{F}|_{\boldsymbol{Z}_n}, \|\cdot\|_\infty\right),$$

where $\mathcal{N}\left(\varepsilon, \mathcal{F}|_{\boldsymbol{Z}_n}, \|\cdot\|_\infty\right)$ denotes the ε-covering number of $\mathcal{F}|_{\boldsymbol{Z}_n}$ w.r.t the L_∞-norm.

3. Generalization Error in H^k Loss Functions for Supervised Learning

In this section, our primary emphasis is on the generalization error of NNs utilizing Sobolev losses with orders H^k where k takes values of 0, 1, and 2. Noting that when k equals 0, the Sobolev loss reduces to L^2 loss. As highlighted in the introduction, Sobolev training has found widespread application across numerous tasks such as (Adler & Lunz, 2018; Gu & Rigazio, 2014; Finlay et al., 2018; Mroueh et al., 2018; Finlay et al., 2018; Werbos, 1992). These specialized loss functions empower models to learn DeNNs capable of approximating the target function with minimal discrepancies in both magnitudes and derivatives.

All the proofs for the content discussed in this section are provided in the appendix. Here, we delve into the generalization error of NNs with H^k loss functions for values of k ranging from 0 to 2. Although the proofs in the appendix address each case individually due to differences in their respective formulations, in this section, we consolidate our discussion to encompass all the cases jointly.

3.1. Notations in Sobolev training

In a typical supervised learning algorithm, the objective is to learn a high-dimensional target function $f(\boldsymbol{x})$ defined on $[0,1]^d$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$ from a finite set of data samples $\{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^M$. When training a DeNN, we aim to identify a DeNN ϕ that approximates $f(\boldsymbol{x})$ based on random data samples $\{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^M$. We assume that $\{\boldsymbol{x}_i\}_{i=1}^M$ is an i.i.d. sequence of random variables uniformly distributed on $[0,1]^d$ in this section. For H^k loss functions for $k = 0, 1$, we investigate the hypothesis spaces defined as follows:

$$\begin{aligned}\mathcal{F}_{B,0} &= \{\phi : \sigma_1\text{-NNs with width} \le C_1 \log L, \\ &\qquad \text{depth} \le C_2 L \log L, \|\phi\|_{L^\infty([0,1]^d)} \le B\} \\ \mathcal{F}_{B,1} &= \{\phi : \phi \in \mathcal{F}_{B,0},\ \|\phi\|_{W^{1,\infty}([0,1]^d)} \le B\},\end{aligned} \tag{1}$$

where $L \in \mathbb{N}_+$ and C_i, B are constant will be defined later in Proposition A.1. However, the limitation arises when relying solely on ReLU neural networks, as they lack higher-order derivatives. Consequently, we explore an alternative neural network architecture that extends beyond traditional ReLU networks. This concept, elucidated in (Yang et al., 2023b;a), is termed Deep Super ReLU Networks (DSRNs). Define a subset of σ_2-NNs characterized by $L \gg 1$ and $C = \mathcal{O}(1)$ concerning L as follows:

$$\begin{aligned}\mathcal{N}_{C,L} := \{&\psi(\boldsymbol{x}) = \psi_2 \circ \boldsymbol{\psi}_1(\boldsymbol{x}) : \psi_2 \text{ is a } \sigma_2\text{-NN with depth} \\ &L_2, \text{each component of } \boldsymbol{\psi}_1 \text{ is a } \sigma_1\text{-NN with depth } L_1, \\ &L_1 + L_2 \le L,\ L_2 \le C \log L.\}\end{aligned} \tag{2}$$

Let us denote the set as

$$\mathcal{F}_{B,2} = \{\phi : \phi \in \mathcal{N}_{6, 10(L+1)\log_2 4L}, \|\phi\|_{W^{2,\infty}([0,1]^d)} \le B\},$$

where the constants in $\mathcal{N}_{6,10(L+1)\log_2 4L}$ are chosen to match the approximation results in (Yang et al., 2023a), and we cite the results in the appendix.

Next, let's proceed to define the loss functions in Sobolev training. We denote

$$\mathcal{R}_{D,k}(\boldsymbol{\theta}) := \int_{[0,1]^d} h_k(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x},\ \mathcal{R}_{S,k}(\boldsymbol{\theta}) := \sum_{i=1}^{M} \frac{h_k(\boldsymbol{x}_i;\boldsymbol{\theta})}{M}. \tag{3}$$

Here, $h_k(\boldsymbol{x};\boldsymbol{\theta})$ is defined as follows:

$$\begin{aligned} h_0(\boldsymbol{x};\boldsymbol{\theta}) &:= |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta})|^2, \\ h_1(\boldsymbol{x};\boldsymbol{\theta}) &:= h_0 + |\nabla(f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}))|^2, \\ h_2(\boldsymbol{x};\boldsymbol{\theta}) &:= h_0 + |\Delta(f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}))|^2 \end{aligned}$$

For $k = 0, 1, 2$, denote

$$\boldsymbol{\theta}_{D,k} := \arg\inf_{\boldsymbol{\theta}\in\Theta_k} \mathcal{R}_{D,k}(\boldsymbol{\theta}),\ \boldsymbol{\theta}_{S,k} := \arg\inf_{\boldsymbol{\theta}\in\Theta_k} \mathcal{R}_{S,k}(\boldsymbol{\theta})$$

with $\Theta_k := \{\boldsymbol{\theta} \in \mathbb{R}^W \mid \phi(\boldsymbol{x};\boldsymbol{\theta}) \in \mathcal{F}_{B,k}\}$.

The overall inference error (generalization error) is $\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k})$ for $k = 0, 1, 2$, which can be divided into two parts:

$$\begin{aligned} &\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \\ =&\mathcal{R}_{D,k}(\boldsymbol{\theta}_{D,k}) + \mathbf{E}\mathcal{R}_{S,k}(\boldsymbol{\theta}_{D,k}) - \mathcal{R}_{D,k}(\boldsymbol{\theta}_{D,k}) \\ +&\mathbf{E}(\mathcal{R}_{S,k}(\boldsymbol{\theta}_{S,k}) - \mathcal{R}_{S,k}(\boldsymbol{\theta}_{D,k}) + \mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) - \mathcal{R}_{S,k}(\boldsymbol{\theta}_{S,k})) \\ \le& \underbrace{\mathcal{R}_{D,k}(\boldsymbol{\theta}_{D,k})}_{\text{approximation error}} + \underbrace{\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) - \mathbf{E}\mathcal{R}_{S,k}(\boldsymbol{\theta}_{S,k})}_{\text{sample error}}, \end{aligned} \tag{4}$$

where the last inequality is due to $\mathbf{E}\mathcal{R}_{S,k}(\boldsymbol{\theta}_{S,k}) \le \mathbf{E}\mathcal{R}_{S,k}(\boldsymbol{\theta}_{D,k})$ by the definition of $\boldsymbol{\theta}_{S,k}$, and $\mathbf{E}\mathcal{R}_{S,k}(\boldsymbol{\theta}_{D,k}) - \mathcal{R}_{D,k}(\boldsymbol{\theta}_{D,k}) = 0$ due to the definition of integration.

Remark 3.1. Our method in the paper can easily generalize to a more general case. If we do not require uniformly distributed input, instead we ask for $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \sim \mu$ i.i.d. For the sample error, we consider the gap between continuous loss function and discrete loss functions, we just need to change the continuous loss function measured by $\mu(\boldsymbol{x})$. For the sample error part, every analysis is totally the same. For the approximation error, we need to bound this with continuous loss functions. Our result of approximation error is measured in $W^{k,\infty}$, therefore, the new loss function can also be bounded. Therefore, we can still obtain the optimal generalization error for the new input distribution. For the assumption about $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$, we can easily generalize it to $\|f\|_{W^{n,\infty}([0,1]^d)} \le D$ for any positive number D. Then the approximation error will be multiplied by D, and the sample error will not be affected since it only depends on the bound of neural networks B. For large D, we just need to choose $B \ge D$ in order to contain the neural networks that can approximate the target functions well.

3.2. Main results

In this paper, we explore the generalization error of DeNNs, considering the flexibility in determining the depth of the neural network.

Theorem 3.2. *Let $d, L, M \in \mathbb{N}_+$, B, C_1, C_2[1] $\in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$ for $n > k$ and $k = 0, 1, 2$, we have*

$$\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \le C\left[\left(\frac{W}{(\log W)^2}\right)^{-\frac{4(n-k)}{d}} + \frac{W^2}{M}\log M\right]$$

where $W = \mathcal{O}(L(\log L)^3)$ is the number of parameters in DeNNs, $\mathbf{E}$ is expected responding to X, $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$, and C is independent with M, L.

Remark 3.3. The expression $W = \mathcal{O}(L(\log L)^3)$ arises from the fact that the number of parameters is equal to the depth times the square of the width. Here, based on the definition of $\mathcal{F}_{B,k}$, the depth is $\mathcal{O}(L\log L)$ and the width is $\mathcal{O}(\log L)$. Therefore, we have $W = \mathcal{O}(L(\log L)^3)$.

Remark 3.4. The generalization error is segmented into two components: the approximation error, denoted by $\left(\frac{W}{(\log W)^2}\right)^{-\frac{4(n-k)}{d}}$, and the sample error, represented as $\frac{W^2}{M}\log M$, where W is the number of parameters in the neural network. Note that while the approximation error is contingent on k, the sample error remains unaffected by k. The sample error's nature is rooted in the Rademacher complexity of several sets, specifically involving derivatives or higher-order derivatives of neural networks. In the appendix, our analysis demonstrates that these complexities maintain a consistent order relative to the depth and width of NNs. Consequently, the constant C in the expression $\frac{W^2}{M}\log M$ is dependent on k, while the other terms remain independent of it.

The outcome is nearly optimal with respect to the number of sample points M (Schmidt-Hieber, 2020) based on following corollary.

Corollary 3.5. *Let $d, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$ for $n > k$ and $k = 0, 1, 2$, we have*

$$\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \le CM^{-\frac{2(n-k)}{2(n-k)+d}}$$

where the result is up to the logarithmic term, $\mathbf{E}$ is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent set of random variables uniformly distributed on $[0,1]^d$. C is a constant independent of M.

[1] Here, the occurrences of C_1 and C_2 in $\mathcal{F}_{B,k}$ continue to hold throughout the rest of the propositions and theorems in this paper.

Proof. Set $W = M^{\frac{d}{2d+4n}}$ in Theorem 3.2, and the result can be obtained directly. □

Remark 3.6. The result in the corollary is nearly optimal, as indicated by the findings in (Schmidt-Hieber, 2020). In other words, we achieve nearly optimal generalization error for DeNNs with any number of hidden layers, and this differs from the results presented in (Schmidt-Hieber, 2020; Suzuki, 2018). Furthermore, the consideration of Sobolev training within Sobolev spaces indeed introduces challenges due to the curse of dimensionality inherent in such spaces. This issue has been extensively discussed and proven in the paper (Yang et al., 2023b). It's important to emphasize that addressing the curse of dimensionality within Sobolev spaces is not the focus of our method. Instead, in this paper, we propose a method to determine the optimal generalization error of Sobolev training.

If one aims to achieve generalization results in Sobolev training without encountering the curse of dimensionality, exploring smaller spaces such as Korobov spaces and Barron spaces, which are subspaces of Sobolev spaces, becomes crucial. Our method is fully applicable in these spaces and can accurately determine the generalization error of Sobolev training without being affected by the curse of dimensionality.

Based on our analysis, it's important to recognize that DeNNs ($N = \mathcal{O}(1)$ or $\mathcal{O}(\log(L))$) with super convergence approximation rates may not necessarily outperform shallow or less DeNNs. This is because, in very deep NNs lacking proper parameter control boundaries, the hypothesis space can become excessively large, resulting in a significant increase in sample error. Nonetheless, it's crucial to acknowledge that DeNNs still offer advantages. Consider a large number $W \gg 1$. Let's assume that both deep and shallow (or not very deep) neural networks have W parameters. The generalization error in DeNNs can be characterized as $\mathcal{O}\left(W^{-\frac{4(n-k)}{d}} + \frac{W^2}{M}\right)$, While the generalization error in shallow (or not very deep) neural networks can be obtained by the following proposition shown in (Schmidt-Hieber, 2020; Suzuki, 2018):

Proposition 3.7 ((Schmidt-Hieber, 2020; Suzuki, 2018))**.** *Let $\mathcal{F}$ be a set of functions defined as follows:*

$\mathcal{F} := \{\phi$ *is a σ_2-NN with number of parameters $\mathcal{O}(W)$ and depth $\log(W)$ and parameters bounded by F*$\}$,

where F is a universal constant larger than 1. Assume that $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$ for $n > k$ and $k = 0, 1, 2$. If $\varepsilon > 0$ satisfies $\mathcal{N}(\varepsilon, \mathcal{F}, n) \geq 3$, then it holds that

$$\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \leq C\Big[W^{-2(n-k)/d} + \varepsilon(1+\sigma) + \left(1+\sigma^2\right) \cdot \frac{\sum_{i=1}^{d} \log \mathcal{N}(\varepsilon, D_i^k\mathcal{F}, \|\cdot\|_\infty) + \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)}{M}\Big]$$

for $k = 0, 1, 2$, where C, σ are universal constants, $\mathcal{N}(\varepsilon, D_i\mathcal{F}, \|\cdot\|_\infty)$ is covering number, and

$$D_i^k\mathcal{F} := \{\bar{\phi} : \bar{\phi} = D_i^k\phi, \phi \in \mathcal{F}\},\ i = 1, \ldots, d.$$

($D_i^0\mathcal{F}$ is an empty set.) Here, the definition of $\boldsymbol{\theta}_{S,k}$ is similar to that in Eq. (3), with the only difference being the substitution of $\mathcal{F}_{B,k}$ with $\mathcal{F}$ here.

Proof. For $k = 0$, the proof can be found directly in (Schmidt-Hieber, 2020). For $k = 1$, the proof can be obtained by combining the approximation results from (Gühring et al., 2020) and (Schmidt-Hieber, 2020). Similarly, for $k = 2$, the proof can be obtained by combining the approximation results from (Gühring & Raslan, 2021) and (Schmidt-Hieber, 2020). □

The outcome is nearly optimal with respect to the number of sample points M, yet it may not be applicable in our work. For DeNNs, the approximation can be much better than for shallow neural networks or neural networks with a small number of hidden layers, as seen in Proposition A.1. This convergence is called super-convergence (Siegel, 2022). This is due to the composition of functions that can make the overall function very complex and allow for excellent approximation. However, we cannot assume that the parameters can be bounded by universal constants in the hypothesis space since the neural network needs to be very flexible in some points, as seen in (Shen et al., 2019). Due to this result, we cannot bound the covering number $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)$ well in DeNNs with super-convergence, while it remains $\mathcal{O}(W)$ for shallow neural networks or those with a small number of hidden layers (Schmidt-Hieber, 2020; Suzuki, 2018). Based on the result in Proposition 3.7, we know that the generalization error up to a logarithmic factor of WeNNs ($L = \mathcal{O}(1)$ or $\mathcal{O}(\log(N))$) is $\mathcal{O}\left(W^{-\frac{2(n-k)}{d}} + \frac{W}{M}\right)$. When $M \geq W^{\frac{2n+2d-2k}{d}}$, the order of the generalization error in DeNNs surpasses that of WeNNs. This relationship is illustrated in Fig. 1. This tells us that for a fixed number of sample points, if you want your networks to have a few parameters, DeNNs are a better choice; otherwise, opt for a shallow neural network. If the number of parameters is nearly fixed from the start and the sample size is very large, consider using DeNNs; otherwise, a shallow network is better to control the generalization error.

Remark 3.8. Our analysis in the paper focuses on $k = 0, 1, 2$, and it is straightforward to generalize to $k \geq 3$. Regarding the approximation result, reference can be made to (Yang et al., 2023a; Gühring & Raslan, 2021). For the generalization error, the estimation method of H^2 loss functions in this paper can be directly extended to estimate the sample error in H^k loss functions.

Another finding based on our analysis is that, as the value of k increases, the curve will shift to the left in Fig. 1, expanding the region of better performance for DeNNs. This discovery suggests that if the loss function requires higher regularity, DeNNs may be a preferable choice over wider neural networks. In this paper, we consider the underparameterized case, while the overparameterized cases are regarded as future work.

Above all, we consider the generalization error of WeNNs and DeNNs. Our method can easily obtain the optimal generalization error for Sobolev training for any fully connected neural network, as stated in the following corollary:

Corollary 3.9. *Let $d, L, N, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$ for $n > k$ and $k = 0, 1, 2$, there exists a σ_2-neural network with $C_1 N \log N$ width and $C_2 L \log L$ depth such that*

$$\begin{aligned}&\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k})\\ \leq & C\left[(NL)^{-\frac{4(n-k)}{d}} + \frac{N^2L^2\log N\log L}{M}\log M\right], \quad (5)\end{aligned}$$

where $\mathbf{E}$ is the expected value with respect to X, $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent set of random variables uniformly distributed on $[0,1]^d$, and C is independent of M, L, N.

3.3. Proof Sketche of Theorem 3.2

In this subsection, we present the proof sketches for estimating the generalization error of DeNNs in Sobolev training and compare them with the differences observed in shallow neural networks with L^2-loss.

• First and foremost, we apply the inequality established in (Györfi et al., 2002), as presented in Lemma B.1, which utilizes uniform covering numbers (refer to Definition 2.3) to bound the generalization errors. This method allows us to obtain the optimal generalization error with respect to the number of samples. Additionally, in Sect. 5, we introduce another approach using the Rademacher complexity to bound the generalization error. While this method may yield suboptimal results in terms of the number of sample points, it proves to be more friendly to the number of parameters, making it suitable for overparameter analysis. This is regarded as future work.

• Secondly, we employ pseudo-dimension (see Definition B.2) to bound the uniform covering number.

The outcomes of these first two steps differ from those in shallow or not very deep neural networks, as presented in (Gühring et al., 2020; Schmidt-Hieber, 2020; Suzuki, 2018). In those studies, a universal bound on parameters was established for all functions. In such instances, the complexity of neural networks could be effectively regulated using covering numbers instead of uniform covering numbers. For covering numbers with a uniform bound parameter in neural networks, estimation can be performed directly, even if they contain derivatives. We present the proof in the appendix.

However, in the approximation of DeNNs, some parameters may become large to approximate specific functions due to the use of techniques called bit extraction in the approximation. These techniques enable better approximation results, as demonstrated in (Lu et al., 2021a; Siegel, 2022). Therefore, we cannot assume a uniform bound like in shallow neural networks. This makes the estimation more challenging, and this is the result we will use uniform covering numbers to bound the generalization errors as Lemma B.1 and then use the pseudo-dimension to bound the uniform covering number.

• Last of all, we estimate the pseudo-dimension of DeNNs and functions represented as higher-order differential operator actions on DeNNs. The structure of derivative and higher-order derivative DeNNs is very complex, involving tensor products and the addition of many relative parts due to the chain rule, which makes estimating the richness or complexity of such spaces difficult.

Our theoretical results can also be validated through experimentation Table 3.2. In the appendix (see Appendix G), we provide a detailed account of a simple experiment conducted to corroborate our findings.

4. Applications for Solving Partial Differential Equations

One of the most crucial applications of neural networks is in solving partial differential equations (PDEs), exemplified in methods such as deep Ritz (E et al., 2017), PINN (Raissi et al., 2019), and deep Galerkin method (DGM) (Sirignano & Spiliopoulos, 2018). These methods construct solutions using neural networks, designing specific loss functions and learning parameters to approximate the solutions. The choice of loss function varies among methods; for instance, deep Ritz methods use the energy variation formula, while PINN and DGM employ residual error.

In these approaches, loss functions integrate orders of, or higher-order derivatives of, neural networks. This design allows models to learn neural networks capable of approximating the target function with minimal discrepancies in both magnitude and derivative. We specifically focuses on solving partial differential equations using neural networks by directly learning the solution. Another approach, known as operator learning, involves learning the mapping between input functions and solutions, but this method is not discussed in this paper. In this section, we delve into the Poisson equation solved using both the deep Ritz method

Neural Network	Large Data Regime	Small Data Regime
Shallow (Depth 1, Width 20)	Mean: 6.18e-4, Std: 9.86e-05	Mean: 1.51e-3, Std: 2.52e-4
Deep (Depth 4, Width 10)	Mean: 3.69e-4, Std: 2.57e-05	Mean: 4.96e-3, Std: 4.46e-3

Table 2. This table compares the performance of shallow and deep neural networks in terms of mean test performance and standard deviation across different data regimes. It illustrates how network depth and data availability impact learning outcomes, with shallow networks performing better in small data scenarios, while deep networks excel with larger datasets.

(E et al., 2017) and the PINN method (Raissi et al., 2019):

$$\begin{cases} -\Delta u = f & \text{in } \widehat{\Omega}, \\ u = 0 & \text{on } \partial\widehat{\Omega}, \end{cases} \tag{6}$$

where $f \in L^2(\widehat{\Omega})$. The domain $\widehat{\Omega}$ is defined in $[0,1]^d$ with a smooth boundary. The assumption of a smooth and bounded $\widehat{\Omega}$ is made to establish a connection between the regularity of the boundary and the interior, facilitated by trace inequalities. The proofs of theorem in this section are presents in the appendix.

4.1. Generalization Error in deep Ritz Method

We initially focus on solving the Poisson equation, as represented by Eq. (6), using the deep Ritz method. This approach utilizes the energy as the loss function during training to guide the neural networks in approximating the solutions. In the context of the Poisson equation, the energy functional encompasses the first derivative of the neural networks. Consequently, the corresponding loss function in deep Ritz methods can be expressed as:

$$\begin{aligned} \mathcal{E}_{D,r}(\phi) :=& \int_{\widehat{\Omega}} g(\boldsymbol{x};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x} + \lambda \int_{\partial\widehat{\Omega}} \phi(\boldsymbol{x};\boldsymbol{\theta})^2\,\mathrm{d}s, \\ g(\boldsymbol{x};\boldsymbol{\theta}) :=& \frac{1}{2}|\nabla\phi(\boldsymbol{x};\boldsymbol{\theta})|^2 - f\phi(\boldsymbol{x};\boldsymbol{\theta}), \end{aligned}$$

where $\boldsymbol{\theta}$ represents all the parameters in the neural network, and λ is a constant to balance the interior and boundary terms. To discretize the loss function, we randomly choose M_1 points in the interior and M_2 points on the boundary:

$$\mathcal{E}_{S,r}(\phi) := \frac{1}{2M_1}\sum_{i=1}^{M_1} g(\boldsymbol{x}_i;\boldsymbol{\theta}) + \frac{\lambda}{M_2}\sum_{i=1}^{M_2} \phi^2(\boldsymbol{y}_i;\boldsymbol{\theta}).$$

Denote $\boldsymbol{\theta}_{D,r} := \arg\inf_{\boldsymbol{\theta}\in\Theta_1} \mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}))$, $\boldsymbol{\theta}_{S,r} := \arg\inf_{\boldsymbol{\theta}\in\Theta_1} \mathcal{E}_{S,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}))$.

In the deep Ritz method, a well-learned solution is characterized by the smallness of the error term $\mathbf{E}\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,r})) - \mathcal{E}_{D,r}(u^*)$, where $u^*(\boldsymbol{x})$ represents the exact solution of Eq. (6). This error can be decomposed into two parts:

$$\begin{aligned} &\mathbf{E}\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,r})) - \mathcal{E}_{D,r}(u^*) \\ \le& \underbrace{\mathbf{E}\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,r})) - \mathbf{E}\mathcal{E}_{S,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,r}))}_{\text{sample error}} \\ &+ \underbrace{\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,r})) - \mathcal{E}_{D,r}(u^*)}_{\text{approximation error}}. \end{aligned} \tag{7}$$

For the approximation error, it can be controlled by the H^1 gap between $\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,r})$ and u^* based on the trace inequality.

Lemma 4.1. *Suppose that $\phi \in \mathcal{F}_{B,1}$ and $u^*(\boldsymbol{x})$ is the exact solution of Eq. (6), then there is a constant C such that $\mathcal{E}_{D,r}(\phi) - \mathcal{E}_{D,r}(u^*) \le C\|\phi - u^*\|_{H^1(\widehat{\Omega})}$.*

The proof is presented in the appendix. Regarding the sample error, it can be estimated using the results in Theorem 3.2 for $k = 1$. Therefore, by combining the estimations of the sample error and the approximation error, we obtain the following theorem:

Theorem 4.2. *Let $d, L, M_1, M_2 \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$, we have an independent constant C such that*

$$\begin{aligned} &\mathbf{E}\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,r})) - \mathcal{E}_{D,r}(u^*) \\ \le& C\left[\left(\frac{W}{(\log W)^2}\right)^{-\frac{4(n-1)}{d}} + W^2\left(\frac{\log M_1}{M_1} + \frac{\log M_2}{M_2}\right)\right] \end{aligned}$$

where $W = \mathcal{O}(L(\log L)^3)$ is the number of parameters in DeNNs, u^* is the exact solution of Eq. (6), $\mathbf{E}$ is expected responding to X and Y, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{M_1}\}$ and $Y := \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{M_2}\}$ is an independent random variables set uniformly distributed on $\widehat{\Omega}$ and $\partial\widehat{\Omega}$.

The error is decomposed into three components. The term $\left(\frac{W}{(\log W)^2}\right)^{-\frac{4(n-1)}{d}}$ represents the optimal approximation error, while $\frac{\log M_1}{M_1}$ corresponds to the sample error in the interior, and $\frac{\log M_2}{M_2}$ denotes the sample error in the boundary.

4.2. Generalization error of Physics-Informed Neural Network Methods

In this subsection, we apply PINN methods to solve the Poisson equation Eq. (6). PINN differs from the deep Ritz

method in that it uses the residual error as the loss function, which is an H^2 loss function, similar to the DGM. Therefore, we will leverage the result from Theorem 3.2 for the case where $k = 2$. The corresponding loss function in Physics-Informed Neural Network methods can be expressed as:

$$\mathcal{E}_{D,p}(\phi) := \int_{\widehat{\Omega}} |\Delta\phi(\boldsymbol{x};\boldsymbol{\theta}) + f(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} + \lambda \int_{\partial\widehat{\Omega}} \phi(\boldsymbol{x};\boldsymbol{\theta})^2 \,\mathrm{d}s.$$

$$\mathcal{E}_{S,p}(\phi) := \sum_{i=1}^{M_1} \frac{|\Delta\phi(\boldsymbol{x}_i;\boldsymbol{\theta}) + f(\boldsymbol{x})|^2}{M_1} + \lambda \sum_{i=1}^{M_2} \frac{\phi^2(\boldsymbol{y}_i;\boldsymbol{\theta})}{M_2}.$$

Denote $\boldsymbol{\theta}_{D,p} := \arg\inf_{\boldsymbol{\theta}\in\Theta_2} \mathcal{E}_{D,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}))$, $\boldsymbol{\theta}_{S,p} := \arg\inf_{\boldsymbol{\theta}\in\Theta_2} \mathcal{E}_{S,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}))$. In the PINN method, if the solution is learned well, meaning that $\mathbf{E}\mathcal{E}_{D,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,p}))$ should be small. It still can be divided into two parts, approximation error and sample error. The approximation error can be read as $\mathcal{E}_{D,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p}))$ and sample error can be read as $\mathbf{E}\mathcal{E}_{D,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,p})) - \mathbf{E}\mathcal{E}_{S,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,p}))$. The approximation error can be bounded as the following lemma:

Lemma 4.3. *Suppose that $\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p}) \in \mathcal{F}_{B,2}$ and u^* is the exact solution of Eq. (6), then there is a constant C such that $\mathcal{E}_{D,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p})) \le C\|\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p}) - u^*\|_{H^2(\widehat{\Omega})}$.*

By applying Theorem 3.2 and above lemma, we can estimate overall inference error for DeNNs .

Theorem 4.4. *Let $d, L, M_1, M_2 \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$, we have an independent constant C_9 such that*

$$\mathbf{E}\mathcal{E}_{D,p}(\boldsymbol{\theta}_{S,p}) \le C\big[\,\big(W/(\log W)^2\big)^{-\frac{4(n-2)}{d}} + W^2\left(\log M_1/M_1 + \log M_2/M_2\right)\big]$$

where $W = \mathcal{O}(L(\log L)^3)$ is the number of parameters in DeNNs, C is a constant independent with M, L.

The performance comparison between shallow NNs and DeNNs in supervised learning (Section 3) can be directly applied here. There is no need to reiterate the discussion in this context.

Remark 4.5. In this section, we only consider Laplace equations. However, our method can work in a more general case. The difference between Laplace equations and general PDEs lies in two aspects. Firstly, we need to utilize a PDE method to bound the approximation error in methods like deep Ritz or PINN by the Sobolev norm of the gap between the neural network and the exact solution. This step only requires assuming that the coefficients in the PDEs are smooth and bounded. Secondly, in the sample error, we need to consider the VC-dimension of the neural network acting as an operator in the PDEs. The approach to this is similar to what we have presented in our paper. Hence, our method works for general PDEs with some assumptions on the coefficients of the PDEs.

5. Improving Generalization Error Across Parameter Counts

Another approach to bounding the generalization error is based on the Rademacher complexity. The result differs from Theorem 3.2, and it provides a better outcome with respect to the number of parameters. However, it is not nearly optimal in terms of the number of sample points. The detailed proof is provided in the appendix. In the subsequent comparison between WeNNs and DeNNs, we will utilize Theorem 3.2 instead of this result, as this paragraph focuses on the underparameterized cases. For the overparameterized cases, estimation of Rademacher complexity may prove useful.

Theorem 5.1. *Let $d, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$ for $n > k$ and $k = 0, 1, 2$, we have*

$$\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \le C\left[\left(\frac{W}{(\log W)^2}\right)^{-\frac{4(n-k)}{d}} + \frac{W}{\sqrt{M}}\log M\right]$$

where $W = \mathcal{O}(L(\log L)^3)$ is the number of parameters in DeNNs, $\mathbf{E}$ is expected responding to X, $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$, and C is independent with M, L.

6. Conclusions and Future Works

In this paper, we examine the optimal generalization error in strategies involving DeNNs versus WeNNs. Our finding suggests that when the parameter count in neural networks remains constant, the decision between WeNNs and DeNNs hinges on the volume of available sample points. On the other hand, when the sample points are predetermined, the choice between WeNNs and DeNNs is contingent upon specific objective functions. Furthermore, our analysis underscores the pivotal role of the loss function's regularity in determining neural network stability. Specifically, in the context of Sobolev training, when the derivative order within the loss function is elevated, favoring DeNNs over their shallow counterparts emerges as a prudent strategy.

We specifically compare DeNNs, characterized by an arbitrary number of hidden layers, in the under-parameterized case. The overparameterized scenario is considered a topic for future research. Additionally, the extension to more complex spaces, such as Korobov spaces (Montanelli & Du, 2019; Yang & Lu, 2023) and functional spaces (Yang & Xiang, 2022), is envisaged as part of our future research endeavors. Moreover, in light of the recently established approximation theory for convolutional neural networks (CNNs) (Zhou, 2020; He et al., 2022), we plan to investigate the generalization error estimates for CNNs.

7. Acknowledgments and Statements

We extend our heartfelt gratitude to Prof. Wenrui Hao and Dr. Xinliang Liu for their invaluable contributions to discussions, experiments, and feedback, which greatly enhanced the quality of this work. The work of the second author was supported by the KAUST Baseline Research Fund.

This paper represents an effort to push the boundaries of the field of Machine Learning. While our work may have various potential societal implications, we do not find it necessary to highlight any specific ones at this time.

References

Abu-Mostafa, Y. The Vapnik-Chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, 1(3):312–317, 1989.

Adler, J. and Lunz, S. Banach wasserstein Gan. *Advances in neural information processing systems*, 31, 2018.

Advani, M., Saxe, A., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Anthony, M., Bartlett, P., et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Barron, A. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Bartlett, P., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Berner, J., Grohs, P., and Jentzen, A. Analysis of the generalization error: Empirical risk minimization over deep artificial neural networks overcomes the curse of dimensionality in the numerical approximation of black–scholes partial differential equations. *SIAM Journal on Mathematics of Data Science*, 2(3):631–657, 2020.

Czarnecki, W., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. Sobolev training for neural networks. *Advances in neural information processing systems*, 30, 2017.

De Ryck, T. and Mishra, S. Error analysis for physics-informed neural networks (PINNs) approximating kolmogorov PDEs. *Advances in Computational Mathematics*, 48(6):1–40, 2022.

Duan, C., Jiao, Y., Lai, Y., Lu, X., and Yang, Z. Convergence rate analysis for Deep Ritz method. *arXiv preprint arXiv:2103.13330*, 2021.

E, W., Han, J., and Jentzen, A. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

Evans, L. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Feng, H., Huang, S., and Zhou, D. Generalization analysis of cnns for classification on spheres. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Finlay, C., Calder, J., Abbasi, B., and Oberman, A. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint arXiv:1808.09540*, 2018.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Gühring, I. and Raslan, M. Approximation rates for neural networks with encodable weights in smoothness spaces. *Neural Networks*, 134:107–130, 2021.

Gühring, I., Kutyniok, G., and Petersen, P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05):803–859, 2020.

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

He, J. On the optimal expressive power of relu dnns and its application in approximation with kolmogorov superposition theorem. *arXiv preprint arXiv:2308.05509*, 2023.

He, J., Li, L., and Xu, J. Approximation properties of deep relu cnns. *Research in the mathematical sciences*, 9(3): 38, 2022.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jakubovitz, D., Giryes, R., and Rodrigues, M. Generalization error in deep learning. In *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*, pp. 153–193. Springer, 2019.

Jiao, Y., Lai, Y., Lo, Y., Wang, Y., and Yang, Y. Error analysis of Deep Ritz methods for elliptic equations. *arXiv preprint arXiv:2107.14478*, 2021.

Lagaris, I., Likas, A., and Fotiadis, D. Artificial neural networks for solving ordinary and partial differential equations. *IEEE Transactions on Neural Networks*, 9 (5):987–1000, 1998.

Lu, J., Shen, Z., Yang, H., and Zhang, S. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021a.

Lu, L., Jin, P., Pang, G., Zhang, Z., and Karniadakis, G. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021b.

Ma, C., Wu, L., and E, W. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022a.

Ma, L., Siegel, J., and Xu, J. Uniform approximation rates and metric entropy of shallow neural networks. *Research in the Mathematical Sciences*, 9(3):46, 2022b.

Mhaskar, H. Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8(1): 164–177, 1996.

Montanelli, H. and Du, Q. New error bounds for deep relu networks using sparse grids. *SIAM Journal on Mathematics of Data Science*, 1(1):78–92, 2019.

Mroueh, Y., Li, C., Sercu, T., Raj, A., and Cheng, Y. Sobolev Gan. In *International Conference on Learning Representations*. International Conference on Learning Representations, ICLR, 2018.

Pollard, D. Empirical processes: theory and applications. Ims, 1990.

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

Rusu, A. A., Colmenarejo, S., Gulcehre, C., Desjardins, G., Kirkpatrick, J., Pascanu, R., Mnih, V., Kavukcuoglu, K., and Hadsell, R. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Sau, B. and Balasubramanian, V. Deep model compression: Distilling knowledge from noisy teachers. *arXiv preprint arXiv:1610.09650*, 2016.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. 2020.

Shen, Z., Yang, H., and Zhang, S. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

Shi, Z., Fan, J., Song, L., Zhou, D., and Suykens, J. Nonlinear functional regression by functional deep neural network with kernel embedding. *arXiv preprint arXiv:2401.02890*, 2024.

Siegel, J. Optimal approximation rates for deep ReLU neural networks on Sobolev spaces. *arXiv preprint arXiv:2211.14400*, 2022.

Siegel, J. and Xu, J. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, pp. 1–57, 2022.

Sirignano, J. and Spiliopoulos, K. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 375:1339–1364, 2018.

Son, H., Jang, J., Han, W., and Hwang, H. Sobolev training for the neural network solutions of pdes. *arXiv preprint arXiv:2101.08932*, 2021.

Suzuki, T. 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.

Vlassis, N. and Sun, W. Sobolev training of thermodynamic-informed neural networks for interpretable elasto-plasticity models with level set hardening. *Computer Methods in Applied Mechanics and Engineering*, 377: 113695, 2021.

Wainwright, M. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Werbos, P. Approximate dynamic programming for real-time control and neural modeling. *Handbook of intelligent control*, 1992.

Yang, Y. and Lu, Y. Optimal deep neural network approximation for Korobov functions with respect to Sobolev norms. *arXiv preprint arXiv:2311.04779*, 2023.

Yang, Y. and Xiang, Y. Approximation of functionals by neural network without curse of dimensionality. *Journal of Machine Learning*, 2022.

Yang, Y., Wu, Y., Yang, H., and Xiang, Y. Nearly optimal approximation rates for deep super ReLU networks on Sobolev spaces. *arXiv preprint arXiv:2310.10766*, 2023a.

Yang, Y., Yang, H., and Xiang, Y. Nearly optimal VC-dimension and pseudo-dimension bounds for deep neural network derivatives. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

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

Yarotsky, D. and Zhevnerchuk, A. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Zhang, Z., Shi, L., and Zhou, D. Classification with deep neural networks and logistic loss. *arXiv preprint arXiv:2307.16792*, 2023.

Zhou, D.-X. Universality of deep convolutional neural networks. *Applied and computational harmonic analysis*, 48 (2):787–794, 2020.

A. Approximation Results for DeNNs with Super Convergence Rates

To bound the approximation error in Eq. (4) for $k = 0$, we employ the following nearly optimal approximation result, which can be found in (Lu et al., 2021a; Siegel, 2022):

Proposition A.1 ((Lu et al., 2021a; Siegel, 2022))**.** *Given a function $f(\boldsymbol{x})$ defined on $[0,1]^d$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, for any $N, L \in \mathbb{N}^+$, there exists a constant $B \geq 1$ and a function implemented by a σ_1-NN with $C_1 N$ width and $C_2 L$ depth such that*

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

where C_1, C_2 and $C_{3,0}$ are constants independent with N and L.

To bound the approximation error in Eq. (4) for $k = 1$, we employ the following nearly optimal approximation result, which can be found in (Yang et al., 2023b)

Proposition A.2 ((Yang et al., 2023b))**.** *Given a function $f(\boldsymbol{x})$ defined on $[0,1]^d$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, for any $N, L \in \mathbb{N}^+$, there exists a constant $B \geq 1$ and a function implemented by a σ_1-NN with $C_1 N$ width and $C_2 L$ depth such that*

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

where C_1, C_2 and $C_{3,1}$ are constants independent with N and L.

For the approximation error n Eq. (4) for $k = 2$, we can directly leverage the findings presented in (Yang et al., 2023a):

Proposition A.3. *[(Yang et al., 2023a)] For any $f \in W^{n,p}((0,1)^d)$ with $\|f\|_{W^{n,p}((0,1)^d)} \leq 1$ for $p \in [1,\infty]$, any $d, n \geq 2$ and $N, L \in \mathbb{N}_+$ with $N \log_2 L + 2^{\lfloor \log_2 N \rfloor} \geq \max\{d, n\}$ and $L \geq N$, there is a DSRN $\phi(\boldsymbol{x})$ in $\mathcal{N}_{6,10(L+1)\log_2 4L}$ with the width $2^{d+6} n^{d+1}(N+d)\log_2(8N)$ such that*

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

where $C_{3,2}$ is the constant independent with N, L.

B. Lemmas Related to Covering Numbers

The following lemma will be used to bounded generalization error by covering numbers:

Lemma B.1 ((Györfi et al., 2002), Theorem 11.4)**.** *Let $M \in \mathbb{N}$, and assume $\|f\|_{L^\infty(\Omega)} \leq B$ for some $B \geq 1$. Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $[-B, B]$. Then for any $0 < \epsilon \leq 1/2$ and $\alpha, \beta > 0$,*

$$\begin{aligned}
&\mathbf{P}\left\{\exists \phi \in \mathcal{F} : \|f - \phi\|_{L^2}^2 - (\mathcal{R}_{S,0}(\phi) - \mathcal{R}_{S,0}(f)) \geq \epsilon\left(\alpha + \beta + \|f - \phi\|_{L^2}^2\right)\right\} \\
&\leq 14\mathcal{N}\left(\frac{\beta\epsilon}{20B}, \mathcal{F}, M\right) \exp\left(-\frac{\epsilon^2(1-\epsilon)\alpha M}{214(1+\epsilon)B^4}\right),
\end{aligned}$$

where

$$\mathcal{R}_{S,0}(g) = \frac{1}{M}\sum_{i=1}^{M} |g(\boldsymbol{x}_i) - y_i|^2$$

for $\{y_i\}_{i=1}^M \subset [-B, B]$ and $\{\boldsymbol{x}_i\}_{i=1}^M \subset \Omega$.

We estimate the uniform covering number by the pseudo-dimension based on the following lemma.

Definition B.2 (pseudo-dimension (Pollard, 1990))**.** Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The pseudo-dimension of $\mathcal{F}$, denoted by $\mathrm{Pdim}(\mathcal{F})$, is the largest integer m for which there exists $(x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_m) \in \mathcal{X}^m \times \mathbb{R}^m$ such that for any $(b_1, \ldots, b_m) \in \{0,1\}^m$ there is $f \in \mathcal{F}$ such that $\forall i : f(x_i) > y_i \iff b_i = 1$.

Lemma B.3 ((Anthony et al., 1999))**.** *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $[-B, B]$. For any $\varepsilon > 0$, we have*

$$\mathcal{N}(\varepsilon, \mathcal{F}, n) \leq \left(\frac{2enB}{\varepsilon Pdim(\mathcal{F})}\right)^{Pdim(\mathcal{F})}$$

for $n \geq Pdim(\mathcal{F})$.

C. Proof of Theorem 3.2

There are three cases in Theorem 3.2. In order to prove Theorem 3.2, we consider these three cases individually.

Remark C.1. For simplicity in notation, we denote the width as $N = \mathcal{O}(\log L)$ in $\mathcal{F}_{B,k}$ in the following proofs. In order to obtain the result for DeNNs, we just substitute $N = \mathcal{O}(\log L)$ in the final result.

C.1. Proof of Theorem 3.2 for $k = 0$.

We first show the proof of the Theorem 3.2 in $k = 0$ case:

Proposition C.2. *Let $d, N, L, M \in \mathbb{N}_+$, B, C_1, C_2[2] $\in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, we have*

$$\mathbf{E}\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) \leq \frac{5136B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,0}, M\right)\right) + 1\right] + 2\inf_{\phi\in\mathcal{F}_{B,0}} \|\phi - f\|_{L^\infty([0,1]^d)},$$

where $\mathcal{F}^2_{B,0} := \{\bar{\phi} : \bar{\phi} = \phi^2, \phi \in \mathcal{F}_{B,0}\}$, $\mathbf{E}$ is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$.

Proof. Due to the definition, we know that

$$\mathcal{R}_{D,0}(\boldsymbol{\theta}_{D,0}) = \inf_{\phi\in\mathcal{F}_{B,0}} \|\phi - f\|_{L^2([0,1]^d)}.$$

For the sample error, due to $\boldsymbol{\theta}_{S,0}$ and $\boldsymbol{\theta}_{D,0}$ belong to Θ almost surely, we have

$$\begin{aligned}
&\mathbf{E}\left[\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) - \mathcal{R}_{S,0}(\boldsymbol{\theta}_{S,0})\right] \\
=&\mathbf{E}\left[\int_{[0,1]^d} |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})|^2 \,\mathrm{d}\boldsymbol{x} - \frac{1}{M}\sum_{i=1}^M |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2\right] \\
=&\mathbf{E}\left[\int_{[0,1]^d} |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})|^2 \,\mathrm{d}\boldsymbol{x} - \frac{2}{M}\sum_{i=1}^M |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2\right] + \mathbf{E}\frac{1}{M}\sum_{i=1}^M |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2 .
\end{aligned}$$

Based on the definition of $\boldsymbol{\theta}_{S,0}$, we know that

$$\mathbf{E}\frac{1}{M}\sum_{i=1}^M |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2 \leq \inf_{\phi\in\mathcal{F}_{B,0}} \|\phi - f\|_{L^\infty([0,1]^d)} \tag{8}$$

Next we denote

$$\mathbf{A}_0 := \int_{[0,1]^d} |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})|^2 \,\mathrm{d}\boldsymbol{x} - \frac{2}{M}\sum_{i=1}^M |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2 . \tag{9}$$

Note that $\mathcal{R}_{S,0}(f) = 0$ and

$$\begin{aligned}
\mathbf{P}(\mathbf{A}_0 \geq \epsilon) &= \mathbf{P}\left(2\int_{[0,1]^d} |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})|^2 \,\mathrm{d}\boldsymbol{x} - \frac{2}{M}\sum_{i=1}^M |f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2 \geq \epsilon + \int_{[0,1]^d} |f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})|^2 \,\mathrm{d}\boldsymbol{x}\right) \\
&= \mathbf{P}\left(2\left(\|f-\phi\|_{L^2([0,1]^d)} - (\mathcal{R}_{S,0}(\phi) - \mathcal{R}_{S,0}(f))\right) \geq \epsilon + \|f-\phi\|_{L^2([0,1]^d)}\right) \\
&= \mathbf{P}\left(\|f-\phi\|_{L^2([0,1]^d)} - (\mathcal{R}_{S,0}(\phi) - \mathcal{R}_{S,0}(f)) \geq \frac{1}{2}\left(\frac{1}{2}\epsilon + \frac{1}{2}\epsilon + \|f-\phi\|_{L^2([0,1]^d)}\right)\right) \\
&\leq 14\mathcal{N}\left(\frac{\epsilon}{80B}, \mathcal{F}_{B,0}, M\right)\exp\left(-\frac{\epsilon M}{5136B^4}\right),
\end{aligned} \tag{10}$$

where the last inequality is due to Lemma B.1.

[2]Here, the occurrences of C_1 and C_2 in $\mathcal{B}_{B,0}$ continue to hold throughout the rest of the propositions and theorems in this paper.

Therefore, we have that

$$\begin{aligned}\mathbf{E}\mathbf{A}_0 &\leq \int_0^\infty \mathbf{P}(\mathbf{A}_0 \geq y)\,\mathrm{d}y \leq \epsilon + \int_\epsilon^\infty \mathbf{P}(\mathbf{A}_0 \geq y)\,\mathrm{d}y \\ &\leq \epsilon + \int_\epsilon^\infty 14\mathcal{N}\left(\frac{\epsilon}{80B}, \mathcal{F}_{B,0}, M\right)\exp\left(-\frac{yM}{5136B^4}\right)\,\mathrm{d}y \end{aligned} \tag{11}$$

By the direct calculation, we have

$$\int_\epsilon^\infty 14\mathcal{N}\left(\frac{\epsilon}{80B}, \mathcal{F}_{B,0}, M\right)\exp\left(-\frac{yM}{5136B^4}\right)\,\mathrm{d}y \leq 14\mathcal{N}\left(\frac{\epsilon}{80B}, \mathcal{F}_{B,0}, M\right)\frac{5136B^4}{M}\exp\left(-\frac{\epsilon M}{5136B^4}\right). \tag{12}$$

Set

$$\epsilon = \frac{5136B^4}{M}\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,0}, M\right)\right) \geq \frac{1}{M}$$

and we have

$$\mathbf{E}\mathbf{A}_0 \leq \frac{5136B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,0}, M\right)\right) + 1\right]$$

Hence we have

$$\mathbf{E}\left[\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) - \mathcal{R}_{S,0}(\boldsymbol{\theta}_{S,0})\right] \leq \frac{5136B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,0}, M\right)\right) + 1\right] + \inf_{\phi\in\mathcal{F}_{B,0}}\|\phi - f\|_{L^\infty([0,1]^d)}. \tag{13}$$

□

Next we need to bound the covering number $\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,0}, M\right)$, which can be estimate by the $\text{Pdim}(\mathcal{F}_{B,0})$ based on Lemma B.3. Based on (Bartlett et al., 2019), $\text{Pdim}(\mathcal{F}_{B,0}) = \mathcal{O}(L^2N^2\log_2 L\log_2 N)$:

Proposition C.3 ((Bartlett et al., 2019)). *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\widehat{C}$ independent with N, L such that*

$$\textit{Pdim}(\mathcal{F}_{B,0}) \leq \widehat{C}N^2L^2\log_2 L\log_2 N. \tag{14}$$

Now we can show the proof of Theorem 3.2 for $k = 0$:

Proof of Theorem 3.2 for $k = 0$. Let $J = \text{Pdim}(\mathcal{F}_{B,0})$. Due to Lemma B.3, for any $M \geq J$, we have

$$\begin{aligned}\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,0}, M\right)\right) \leq& \log\left(14\left(\frac{160eM^2B^2}{\text{Pdim}(\mathcal{F}_{B,0})}\right)^{\text{Pdim}(\mathcal{F}_{B,0})}\right) \\ \leq& C\text{Pdim}(\mathcal{F}_{B,0})\log M \\ \leq& C\widehat{C}N^2L^2\log_2 L\log_2 N\log M. \end{aligned} \tag{15}$$

Hence, we have that there is a constant $C_5 = C_5(B, d, C_1, C_2) := 10272B^4C\widehat{C}$ such that

$$\mathbf{E}\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) \leq C_{3,0}N^{-2n/d}L^{-2n/d} + C_5\frac{N^2L^2\log_2 L\log_2 N}{M}\log M. \tag{16}$$

Finally, we set $N = \mathcal{O}(\log L)$ according to the definition of DeNNs, and the number of parameters $W = \mathcal{O}(N^2L\log L) = \mathcal{O}(L(\log L)^3)$ to complete the proof. □

C.2. Proof of Theorem 3.2 for $k = 1$

The Theorem 3.2 in the $k = 1$ case can be read as follows:

Proposition C.4. *Let $d, N, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, we have*

$$\begin{aligned}\mathbf{E}\mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1}) \leq & \frac{5136B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,1}, M\right)\right)+1\right] \\ & + \frac{5136d^4B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80dBM}, \nabla\mathcal{F}_{B,1}, M\right)\right)+1\right] + 3\inf_{\phi\in\mathcal{F}_{B,1}}\|\phi - f\|_{W^{1,\infty}([0,1]^d)},\end{aligned} \tag{17}$$

where

$$\nabla\mathcal{F}_{B,1} := \{\bar{\phi} : \bar{\phi} = \nabla\phi, \phi \in \mathcal{F}_{B,1}\}, \tag{18}$$

$\mathbf{E}$ *is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$.*

Proof. Due to the definition, we know that

$$\mathcal{R}_{D,1}(\boldsymbol{\theta}_{D,1}) = \inf_{\phi\in\mathcal{F}_{B,1}}\|\phi - f\|_{H^1([0,1]^d)}.$$

For the sample error, due to $\boldsymbol{\theta}_{S,1}$ and $\boldsymbol{\theta}_{D,1}$ belong to Θ almost surely, we have

$$\begin{aligned}&\mathbf{E}\left[\mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1}) - \mathcal{R}_{S,1}(\boldsymbol{\theta}_{S,1})\right] \\ \leq &\mathbf{E}\left[\int_{[0,1]^d}|\nabla f(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})|^2\,\mathrm{d}\boldsymbol{x} - \frac{1}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2\right] + \mathbf{E}\left[\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) - \mathcal{R}_{S,0}(\boldsymbol{\theta}_{S,0})\right].\end{aligned} \tag{19}$$

For the upper bound of $\mathbf{E}\left[\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) - \mathcal{R}_{S,0}(\boldsymbol{\theta}_{S,0})\right]$ can be obtained by the Proposition C.2.

For the rest of sample error, we have

$$\begin{aligned}&\mathbf{E}\left[\int_{[0,1]^d}|\nabla f(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})|^2\,\mathrm{d}\boldsymbol{x} - \frac{1}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2\right] \\ =&\mathbf{E}\left[\int_{[0,1]^d}|\nabla f(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})|^2\,\mathrm{d}\boldsymbol{x} - \frac{2}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2\right] + \mathbf{E}\frac{1}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2 \\ \leq&\mathbf{E}\left[\int_{[0,1]^d}|\nabla f(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})|^2\,\mathrm{d}\boldsymbol{x} - \frac{2}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2\right] + \inf_{\phi\in\mathcal{F}_{B,1}}\|\phi - f\|_{W^{1,\infty}([0,1]^d)}\end{aligned} \tag{20}$$

where the last inequality is due to the definition of $\boldsymbol{\theta}_{S,1}$.

Set

$$\mathbf{A}_1 = \int_{[0,1]^d}|\nabla f(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})|^2\,\mathrm{d}\boldsymbol{x} - \frac{2}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2. \tag{21}$$

Note that $\mathcal{R}_{S,0}(\nabla f) = 0$ and

$$\begin{aligned}\mathbf{P}(\mathbf{A}_1 \geq \epsilon) &= \mathbf{P}\left(2\|\nabla f - \nabla\phi\|_{L^2([0,1]^d)} - \frac{2}{M}\sum_{i=1}^{M}|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2 \geq \epsilon + \|\nabla f - \nabla\phi\|_{L^2([0,1]^d)}\right) \\ &= \mathbf{P}\left(2\left(\|\nabla f - \nabla\phi\|_{L^2([0,1]^d)} - (\mathcal{R}_{S,0}(\nabla\phi) - \mathcal{R}_{S,0}(\nabla f))\right) \geq \epsilon + \|\nabla f - \nabla\phi\|_{L^2([0,1]^d)}\right) \\ &= \mathbf{P}\left(\|\nabla f - \nabla\phi\|_{L^2([0,1]^d)} - (\mathcal{R}_{S,0}(\nabla\phi) - \mathcal{R}_{S,0}(\nabla f)) \geq \frac{1}{2}\left(\frac{1}{2}\epsilon + \frac{1}{2}\epsilon + \|\nabla f - \nabla\phi\|_{L^2([0,1]^d)}\right)\right) \\ &\leq 14\mathcal{N}\left(\frac{\epsilon}{80dB}, \nabla\mathcal{F}_{B,1}, M\right)\exp\left(-\frac{\epsilon M}{5136d^4B^4}\right),\end{aligned} \tag{22}$$

where the last inequality is due to Lemma B.1.

Therefore, we have that

$$\begin{aligned}\mathbf{E}\mathbf{A}_1 &\leq \int_0^\infty \mathbf{P}(\mathbf{A}_1 \geq y)\,\mathrm{d}y \leq \epsilon + \int_\epsilon^\infty \mathbf{P}(\mathbf{A}_1 \geq y)\,\mathrm{d}y \\ &\leq \epsilon + \int_\epsilon^\infty 14\mathcal{N}\left(\frac{\epsilon}{80dB}, \nabla\mathcal{F}_{B,1}, M\right)\exp\left(-\frac{yM}{5136d^4B^4}\right)\mathrm{d}y \end{aligned} \tag{23}$$

By the direct calculation, we have

$$\int_\epsilon^\infty 14\mathcal{N}\left(\frac{\epsilon}{80dB}, \nabla\mathcal{F}_{B,1}, M\right)\exp\left(-\frac{yM}{5136d^4B^4}\right)\mathrm{d}y \leq 14\mathcal{N}\left(\frac{\epsilon}{80dB}, \nabla\mathcal{F}_{B,1}, M\right)\frac{5136d^4B^4}{M}\exp\left(-\frac{\epsilon M}{5136d^4B^4}\right). \tag{24}$$

Set

$$\epsilon = \frac{5136d^4B^4}{M}\log\left(14\mathcal{N}\left(\frac{1}{80dBM}, \nabla\mathcal{F}_{B,1}, M\right)\right) \geq \frac{1}{M}$$

and we have

$$\mathbf{E}\mathbf{A}_1 \leq \frac{5136B^4d^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \nabla\mathcal{F}_{B,1}, M\right)\right) + 1\right]$$

Hence we have

$$\begin{aligned}&\mathbf{E}\left[\mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1}) - \mathcal{R}_{S,1}(\boldsymbol{\theta}_{S,1})\right] \leq \frac{5136B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,1}, M\right)\right) + 1\right] \\ &+ \frac{5136d^4B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80dBM}, \nabla\mathcal{F}_{B,1}, M\right)\right) + 1\right] + 2\inf_{\phi\in\mathcal{F}_{B,1}}\|\phi - f\|_{W^{1,\infty}([0,1]^d)}. \end{aligned} \tag{25}$$

Finally, we set $N = \mathcal{O}(\log L)$ according to the definition of DeNNs, and the number of parameters $W = \mathcal{O}(N^2L\log L) = \mathcal{O}(L(\log L)^3)$ to complete the proof. □

The remainder of the proof focuses on estimating $\mathcal{N}\left(\frac{1}{80dBM}, \nabla\mathcal{F}_{B,1}, M\right)$ with the aim of limiting the generalization error. The pseudo-dimension serves as a valuable tool for establishing such constraints. Notably, (Yang et al., 2023b) provides nearly optimal bounds for the pseudo-dimension in the context of DeNN derivatives. Drawing inspiration from the proof presented in (Yang et al., 2023b), the upcoming propositions will illustrate the pseudo-dimension of $\nabla\mathcal{F}_{B,1}$.

Before we estimate pseudo-dimension of $\nabla\mathcal{F}_{B,1}$, we first introduce Vapnik–Chervonenkis dimension (VC-dimension)

Definition C.5 (VC-dimension (Abu-Mostafa, 1989))**.** Let H denote a class of functions from $\mathcal{X}$ to $\{0,1\}$. For any non-negative integer m, define the growth function of H as

$$\Pi_H(m) := \max_{x_1,x_2,\ldots,x_m\in\mathcal{X}}\left|\{(h(x_1), h(x_2), \ldots, h(x_m)) : h \in H\}\right|.$$

The Vapnik–Chervonenkis dimension (VC-dimension) of H, denoted by $\mathrm{VCdim}(H)$, is the largest m such that $\Pi_H(m) = 2^m$. For a class $\mathcal{G}$ of real-valued functions, define $\mathrm{VCdim}(\mathcal{G}) := \mathrm{VCdim}(\mathrm{sgn}(\mathcal{G}))$, where $\mathrm{sgn}(\mathcal{G}) := \{\mathrm{sgn}(f) : f \in \mathcal{G}\}$ and $\mathrm{sgn}(x) = 1[x > 0]$.

In the proof of Proposition C.8, we use the following lemmas:

Lemma C.6 ((Bartlett et al., 2019; Anthony et al., 1999))**.** *Suppose $W \leq M$ and let $P_1, \ldots, P_M$ be polynomials of degree at most D in W variables. Define*

$$K := \left|\{(\mathrm{sgn}(P_1(a)), \ldots, \mathrm{sgn}(P_M(a))) : a \in \mathbb{R}^W\}\right|,$$

then we have $K \leq 2(2eMD/W)^W$.

Lemma C.7 ((Bartlett et al., 2019))**.** *Suppose that $2^m \leq 2^t(mr/w)^w$ for some $r \geq 16$ and $m \geq w \geq t \geq 0$. Then, $m \leq t + w\log_2(2r\log_2 r)$.*

Proposition C.8. *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\bar{C}$ independent with N, L such that*

$$Pdim(\nabla \mathcal{F}_{B,1}) \leq \bar{C} N^2 L^2 \log_2 L \log_2 N. \tag{26}$$

Proof. Denote

$$\mathcal{F}_{B,1,\mathcal{N}} := \{\eta(\boldsymbol{x}, y) : \eta(\boldsymbol{x}, y) = \psi(\boldsymbol{x}) - y, \psi \in \nabla \mathcal{F}_{B,1}, (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}\}.$$

Based on the definition of VC-dimension and pseudo-dimension, we have that

$$\text{Pdim}(\nabla \mathcal{F}_{B,1}) \leq \text{VCdim}(\mathcal{F}_{B,1,\mathcal{N}}). \tag{27}$$

For the $\text{VCdim}(\mathcal{F}_{B,1,\mathcal{N}})$, it can be bounded by following way. The proof is similar to that in (Yang et al., 2023b;a).

For a DeNN with N width and L depth, it can be represented as

$$\phi = \boldsymbol{W}_{L+1}\sigma_1(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L) + b_{L+1}.$$

Therefore,

$$\begin{aligned}\psi(\boldsymbol{x}) = \nabla\phi(\boldsymbol{x}) = &\sum_{i=1}^{d} \boldsymbol{W}_{L+1}\sigma_0(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L) \\ &\cdot \boldsymbol{W}_L\sigma_0(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots)\ldots\boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)(\boldsymbol{W}_1)_i,\end{aligned} \tag{28}$$

where $\boldsymbol{W}_i \in \mathbb{R}^{N_i \times N_{i-1}}$ ($(\boldsymbol{W})_i$ is i-th column of $\boldsymbol{W}$) and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the i-th linear transform in ϕ, and $\sigma_0(x) = \text{sgn}(x) = 1[x > 0]$, which is the derivative of the ReLU function and $\sigma_0(\boldsymbol{x}) = \text{diag}(\sigma_0(x_i))$.

Let $\bar{\boldsymbol{x}} = (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}$ be an input and $\boldsymbol{\theta} \in \mathbb{R}^W$ be a parameter vector in $\eta := \psi^2 - y$. We denote the output of ψ with input $\boldsymbol{x}$ and parameter vector $\boldsymbol{\theta}$ as $f(\boldsymbol{x}, \boldsymbol{\theta})$. For fixed $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ in $\mathbb{R}^d$, we aim to bound

$$K := \left|\{(\text{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \text{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in \mathbb{R}^W\}\right|. \tag{29}$$

The proof is inspired by (Yang et al., 2023b, Theorem 1). For any partition $\mathcal{S} = \{P_1, P_2, \ldots, P_T\}$ of the parameter domain $\mathbb{R}^W$, we have $K \leq \sum_{i=1}^{T} |\{(\text{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \text{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in P_i\}|$. We choose the partition such that within each region P_i, the functions $f(\boldsymbol{x}_j, \cdot)$ are all fixed polynomials of bounded degree. This allows us to bound each term in the sum using Lemma C.6.

We define a sequence of sets of functions $\{\mathbb{F}_j\}_{j=0}^{L}$ with respect to parameters $\boldsymbol{\theta} \in \mathbb{R}^W$:

$$\begin{aligned}
\mathbb{F}_0 &:= \{\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1\} \cup \{(\boldsymbol{W}_1)_i\}_{i=1}^{d} \\
\mathbb{F}_1 &:= \{\boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2\} \cup \{(\boldsymbol{W}_1)_i\}_{i=1}^{d} \\
\mathbb{F}_2 &:= \{\boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \boldsymbol{W}_3\sigma_0(\boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2), \boldsymbol{W}_3\sigma_1(\boldsymbol{W}_2\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3\} \cup \{(\boldsymbol{W}_1)_i\}_{i=1}^{d} \\
&\vdots \\
\mathbb{F}_L &:= \{\boldsymbol{W}_2\sigma_0(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \ldots, \boldsymbol{W}_{L+1}\sigma_0(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L)\} \cup \{(\boldsymbol{W}_1)_i\}_{i=1}^{d}.
\end{aligned} \tag{30}$$

The partition of $\mathbb{R}^W$ is constructed layer by layer through successive refinements denoted by $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_L$. These refinements possess the following properties:

1. We have $|\mathcal{S}_0| = 1$, and for each $n = 1, \ldots, L$, we have $\frac{|\mathcal{S}_n|}{|\mathcal{S}_{n-1}|} \leq 2\left(\frac{2emnN_k}{\sum_{i=1}^{n} W_i}\right)^{\sum_{i=1}^{n} W_i}$.

2. For each $n = 0, \ldots, L-1$, each element S of $\mathcal{S}_n$, when $\boldsymbol{\theta}$ varies in S, the output of each term in $\mathbb{F}_n$ is a fixed polynomial function in $\sum_{i=1}^{n} W_i$ variables of $\boldsymbol{\theta}$, with a total degree no more than $n+1$.

3. For each element S of $\mathcal{S}_L$, when $\boldsymbol{\theta}$ varies in S, the h-th term in $\mathbb{F}_L$ for $h \in \{1, 2, \ldots, L+1\}$ is a fixed polynomial function in W_h variables of $\boldsymbol{\theta}$, with a total degree no more than 1.

We define $\mathcal{S}_0 = \{\mathbb{R}^W\}$, which satisfies properties 1,2 above, since $\boldsymbol{W}_1\boldsymbol{x}_j + \boldsymbol{b}_1$ and $(\boldsymbol{W}_1)_i$ are affine functions of $\boldsymbol{W}_1, \boldsymbol{b}_1$.

To define $\mathcal{S}_n$, we use the last term of $\mathbb{F}_{n-1}$ as inputs for the last two terms in $\mathbb{F}_n$. Assuming that $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{n-1}$ have already been defined, we observe that the last two terms are new additions to $\mathbb{F}_n$ when comparing it to $\mathbb{F}_{n-1}$. Therefore, all elements in $\mathbb{F}_n$ except the last two are fixed polynomial functions in W_n variables of $\boldsymbol{\theta}$, with a total degree no greater than n when $\boldsymbol{\theta}$ varies in $S \in \mathcal{S}_n$. This is because $\mathcal{S}_n$ is a finer partition than $\mathcal{S}_{n-1}$.

We denote $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$ as the output of the k-th node in the last term of $\mathbb{F}_{n-1}$ in response to $\boldsymbol{x}_j$ when $\boldsymbol{\theta} \in S$. The collection of polynomials

$$\{p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta}) : j = 1, \ldots, m,\ k = 1, \ldots, N_n\}$$

can attain at most $2\left(\frac{2emnN_n}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}$ distinct sign patterns when $\boldsymbol{\theta} \in S$ due to Lemma C.6 for sufficiently large m. Therefore, we can divide S into $2\left(\frac{2emnN_n}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}$ parts, each having the property that $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$ does not change sign within the subregion. By performing this for all $S \in \mathcal{S}_{n-1}$, we obtain the desired partition $\mathcal{S}_n$. This division ensures that the required property 1 is satisfied.

Additionally, since the input to the last two terms in $\mathbb{F}_n$ is $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$, and we have shown that the sign of this input will not change in each region of $\mathcal{S}_n$, it follows that the output of the last two terms in $\mathbb{F}_n$ is also a polynomial without breakpoints in each element of $\mathcal{S}_n$. Therefore, the required property 2 is satisfied.

In the context of DeNNs, the last layer is characterized by all terms containing the activation function σ_0. Consequently, for any element S of the partition $\mathcal{S}_L$, when the vector of parameters $\boldsymbol{\theta}$ varies within S, the h-th term in $\mathbb{F}_L$ for $h \in \{1, 2, \ldots, L+1\}$ can be expressed as a polynomial function of at most degree 1, which depends on at most W_h variables of $\boldsymbol{\theta}$. Hence, the required property 3 is satisfied.

Due to property 3, note that there is a partition $\mathcal{S}$ where

$$|\mathcal{S}| \leq \prod_{n=1}^{L} 2\left(\frac{2emnN_n}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i}.$$

And the output of ψ is a polynomial function in $\sum_{i=1}^{L+1} W_i$ variables of $\boldsymbol{\theta} \in S \in \mathcal{S}_L$, of total degree no more than $L+1$. Therefore, for each $S \in \mathcal{S}_L$ we have

$$|\{(\operatorname{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \operatorname{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in S\}| \leq 2\left(2em(L+1)/\sum_{i=1}^{L+1} W_i\right)^{\sum_{i=1}^{L+1} W_i}.$$

Then

$$K \leq 2^{L+1}\left(\frac{em(L+1)(L+2)N}{U}\right)^U \tag{31}$$

where $U := \sum_{n=1}^L \sum_{i=1}^n W_i = \mathcal{O}(N^2L^2)$, N is the width of the network, and the last inequality is due to weighted AM-GM. For the definition of the VC-dimension, we have

$$2^{\text{VCdim}(D\mathcal{F}_{B,1,\mathcal{N}})} \leq 2^L\left(\frac{e\text{VCdim}(D\mathcal{F}_{B,1,\mathcal{N}})(L+1)LN}{U}\right)^U. \tag{32}$$

Due to Lemma C.7, we obtain that

$$\text{VCdim}(\mathcal{F}_{B,1,\mathcal{N}}) \leq L + 1 + U\log_2[2(L+1)(L+2)\log_2(L+1)(L+2)] = \mathcal{O}(N^2L^2\log_2 L\log_2 N) \tag{33}$$

since $U = \mathcal{O}(N^2L^2)$. □

Proof of Theorem 3.2 for $k = 1$. The proofs of Theorem 3.2 for $k = 1$ are analogous to those of Theorem 3.2 for $k = 0$ by combining the results of Propositions C.4 and C.8. □

C.3. Proof of Theorem 3.2 for $k = 2$

The Theorem 3.2 in the $k = 2$ case can be read as follows:

Proposition C.9. *Let $d, N, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, we have*

$$\begin{aligned}\mathbf{E}\mathcal{R}_{D,2}(\boldsymbol{\theta}_{S,2}) \leq & \frac{5136B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80BM}, \mathcal{F}_{B,2}, M\right)\right)+1\right] \\ & + \frac{5136d^4B^4}{M}\left[\log\left(14\mathcal{N}\left(\frac{1}{80dBM}, \Delta\mathcal{F}_{B,2}, M\right)\right)+1\right] + 3\inf_{\phi\in\mathcal{F}_{B,2}}\|\phi - f\|_{W^{2,\infty}([0,1]^d)},\end{aligned} \tag{34}$$

where

$$\Delta\mathcal{F}_{B,2} := \{\bar{\phi} : \bar{\phi} = \Delta\phi, \phi \in \mathcal{F}_{B,2}\}, \tag{35}$$

$\mathbf{E}$ *is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$.*

Proof. The proof closely resembles that of Proposition C.4. □

The subsequent part of the proof centers on estimating $\mathcal{N}\left(\frac{1}{80dBM}, \Delta\mathcal{F}_{B,2}, M\right)$ to constrain the generalization error. Utilizing the pseudo-dimension proves instrumental in establishing these bounds.

Proposition C.10. *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\bar{C}$ independent with N, L such that*

$$\textit{Pdim}(\Delta\mathcal{F}_{B,2}) \leq \bar{C}N^2L^2\log_2 L\log_2 N. \tag{36}$$

Proof. Denote

$$\mathcal{F}_{B,2,\mathcal{N}} := \{\eta(\boldsymbol{x}, y) : \eta(\boldsymbol{x}, y) = \psi(\boldsymbol{x}) - y, \psi \in \Delta\mathcal{F}_{B,2}, (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}\}.$$

Based on the definition of VC-dimension and pseudo-dimension, we have that

$$\text{Pdim}(\Delta\mathcal{F}_{B,2}) \leq \text{VCdim}(\mathcal{F}_{B,2,\mathcal{N}}). \tag{37}$$

For the $\text{VCdim}(\mathcal{F}_{B,2,\mathcal{N}})$, it can be bounded by following way. The proof is similar to that in (Yang et al., 2023a).

For a DeNN with N width and L depth, it can be represented as

$$\phi = \boldsymbol{W}_{L+1}\sigma(\boldsymbol{W}_L\sigma(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L) + b_{L+1},$$

where σ can be either the ReLU or the ReLU square. Then the first order derivative can be read as

$$\begin{aligned}\psi(\boldsymbol{x}) = D_i\phi(\boldsymbol{x}) = & \boldsymbol{W}_{L+1}\sigma'(\boldsymbol{W}_L\sigma(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L) \\ & \cdot \boldsymbol{W}_L\sigma'(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots)\ldots\boldsymbol{W}_2\sigma'(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)(\boldsymbol{W}_1)_i,\end{aligned} \tag{38}$$

where $\boldsymbol{W}_i \in \mathbb{R}^{N_i\times N_{i-1}}$ ($(\boldsymbol{W})_i$ is i-th column of $\boldsymbol{W}$) and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the i-th linear transform in ϕ, and $\sigma'(\boldsymbol{x}) = \text{diag}(\sigma'(x_i))$. Then we have

$$\lambda(\boldsymbol{x}) = \Delta\phi(\boldsymbol{x}) = \sum_{a=1}^{L+1}\lambda_a(\boldsymbol{x}). \tag{39}$$

For

$$\begin{aligned}\lambda_a(\boldsymbol{x}) = & \sum_{i=1}^{d}\boldsymbol{W}_{L+1}\sigma'(\boldsymbol{W}_L\sigma(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L)\cdot\boldsymbol{W}_L\sigma'(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots)\ldots \\ & \cdot[\boldsymbol{W}_a\sigma''(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots)\cdot\boldsymbol{W}_{a-1}\sigma'(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots)\boldsymbol{W}_2\sigma'(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)(\boldsymbol{W}_1)_i] \\ & \cdot\ldots\boldsymbol{W}_{a-1}\sigma'(\ldots\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots)\boldsymbol{W}_2\sigma'(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)(\boldsymbol{W}_1)_i,\end{aligned} \tag{40}$$

where $\sigma''(\boldsymbol{x}) = \operatorname{diag}(\sigma''(x_i))$ is a three-order tensor. Denote W_i as the number of parameters in $\boldsymbol{W}_i, \boldsymbol{b}_i$, i.e., $W_i = N_i N_{i-1} + N_i$.

Let $\bar{\boldsymbol{x}} = (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}$ be an input and $\boldsymbol{\theta} \in \mathbb{R}^W$ be a parameter vector in $\eta := \lambda^2 - y$. We denote the output of λ_a with input $\boldsymbol{x}$ and parameter vector $\boldsymbol{\theta}$ as $f(\boldsymbol{x}, \boldsymbol{\theta})$. For fixed $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ in $\mathbb{R}^d$, we aim to bound

$$K := \left|\{(\operatorname{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \operatorname{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in \mathbb{R}^W\}\right|. \tag{41}$$

The proof is inspired by (Yang et al., 2023a, Theorem 2). For any partition $\mathcal{S} = \{P_1, P_2, \ldots, P_T\}$ of the parameter domain $\mathbb{R}^W$, we have

$$K \leq \sum_{i=1}^{T} \left|\{(\operatorname{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \operatorname{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in P_i\}\right|.$$

We choose the partition such that within each region P_i, the functions $f(\boldsymbol{x}_j, \cdot)$ are all fixed polynomials of bounded degree. This allows us to bound each term in the sum using Lemma C.6.

We define a sequence of sets of functions $\{\mathbb{F}_j\}_{j=0}^{L}$ with respect to parameters $\boldsymbol{\theta} \in \mathbb{R}^W$:

$$\begin{aligned}
\mathbb{F}_0 &:= \{(\boldsymbol{W}_1)_1, (\boldsymbol{W}_1)_2, \ldots, (\boldsymbol{W}_1)_d, \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1\} \\
\mathbb{F}_1 &:= \{\boldsymbol{W}_2\sigma''(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \boldsymbol{W}_2\sigma'(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1), \boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2\} \cup \mathbb{F}_0 \\
\mathbb{F}_2 &:= \{\boldsymbol{W}_3\sigma''(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2), \boldsymbol{W}_3\sigma'(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2), \\
&\qquad \boldsymbol{W}_3\sigma(\boldsymbol{W}_2\sigma(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3\} \cup \mathbb{F}_1 \\
&\quad \vdots \\
\mathbb{F}_L &:= \{\boldsymbol{W}_{L+1}\sigma'(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L), \\
&\qquad \boldsymbol{W}_{L+1}\sigma''(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L)\} \cup \mathbb{F}_{L-1}.
\end{aligned} \tag{42}$$

The partition of $\mathbb{R}^W$ is constructed layer by layer through successive refinements denoted by $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_L$. We denote $L^* = L - C\log_2 L$. These refinements possess the following properties:

1. We have $|\mathcal{S}_0| = 1$, and for each $n = 1, \ldots, L$, we have

$$\frac{|\mathcal{S}_n|}{|\mathcal{S}_{n-1}|} \leq 2\left(\frac{2em(1 + (n-1)2^{\max\{0, n-1-L^*\}})N_n}{\sum_{i=1}^{n} W_i}\right)^{\sum_{i=1}^{n} W_i}.$$

2. For each $n = 0, \ldots, L^*$, each element S of $\mathcal{S}_n$, when $\boldsymbol{\theta}$ varies in S, the output of each term in $\mathbb{F}_n$ is a fixed polynomial function in $\sum_{i=1}^{n} W_i$ variables of $\boldsymbol{\theta}$, with a total degree no more than $1 + n2^{\max\{0, n-L^*\}}$.

We define $\mathcal{S}_0 = \{\mathbb{R}^W\}$, which satisfies properties 1,2 above, since $\boldsymbol{W}_1\boldsymbol{x}_j + \boldsymbol{b}_1$ and $(\boldsymbol{W}_1)_i$ for all $i = 1, \ldots, d$ are affine functions of $\boldsymbol{W}_1, \boldsymbol{b}_1$.

For each $n = 0, \ldots, L$, to define $\mathcal{S}_n$, we use the last term of $\mathbb{F}_{n-1}$ as inputs for the new terms in $\mathbb{F}_n$. Assuming that $\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_{n-1}$ have already been defined, we observe that the last two or three terms are new additions to $\mathbb{F}_n$ when comparing it to $\mathbb{F}_{n-1}$. Therefore, all elements in $\mathbb{F}_n$ except the $\mathbb{F}_n \backslash \mathbb{F}_{n-1}$ are fixed polynomial functions in W_n variables of $\boldsymbol{\theta}$, with a total degree no greater than $1 + (n-1)2^{\max\{0, n-1-L^*\}}$ when $\boldsymbol{\theta}$ varies in $S \in \mathcal{S}_n$. This is because $\mathcal{S}_n$ is a finer partition than $\mathcal{S}_{n-1}$.

We denote $p_{\boldsymbol{x}_j, n-1, S, k}(\boldsymbol{\theta})$ as the output of the k-th node in the last term of $\mathbb{F}_{n-1}$ in response to $\boldsymbol{x}_j$ when $\boldsymbol{\theta} \in S$. The collection of polynomials

$$\{p_{\boldsymbol{x}_j, n-1, S, k}(\boldsymbol{\theta}) : j = 1, \ldots, m,\ k = 1, \ldots, N_n\}$$

can attain at most $2\left(\frac{2em(1+(n-1)2^{\max\{0, n-1-L^*\}})N_n}{\sum_{i=1}^{n} W_i}\right)^{\sum_{i=1}^{n} W_i}$ distinct sign patterns when $\boldsymbol{\theta} \in S$ due to Lemma C.6 for sufficiently large m. Therefore, we can divide S into

$$2\left(\frac{2em(1 + (n-1)2^{\max\{0, n-1-L^*\}})N_n}{\sum_{i=1}^{n} W_i}\right)^{\sum_{i=1}^{n} W_i}$$

parts, each having the property that $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$ does not change sign within the subregion. By performing this for all $S \in \mathcal{S}_{n-1}$, we obtain the desired partition $\mathcal{S}_n$. This division ensures that the required property 1 is satisfied.

Additionally, since the input to the last two terms in $\mathbb{F}_n$ is $p_{\boldsymbol{x}_j,n-1,S,k}(\boldsymbol{\theta})$, and we have shown that the sign of this input will not change in each region of $\mathcal{S}_n$, it follows that the output of the last two terms in $\mathbb{F}_n$ is also a polynomial without breakpoints in each element of $\mathcal{S}_n$, therefore, the required property 2 is satisfied.

Due to the structure of λ, note that there is a partition $\mathcal{S}$ where

$$|\mathcal{S}| \leq 2\left(\frac{2em(1+(n-1)2^{\max\{0,n-1-L^*\}})N_n}{\sum_{i=1}^n W_i}\right)^{\sum_{i=1}^n W_i},$$

where $L^* = 10(L+1)\log_2 4L - 6\log_2(10(L+1)\log_2 4L)$. And the output of λ is a polynomial function in $\sum_{i=1}^{L+1} W_i$ variables of $\boldsymbol{\theta} \in S \in \mathcal{S}_L$, of total degree no more than

$$d_2 := 2\sum_{n=0}^{L}(1+n2^{\max\{0,n-L^*\}}) = 2L+2+L^6(L-1).$$

The rest of the proof is similar with those in Proposition C.8, we can obtain that

$$\text{VCdim}(\mathcal{F}_{B,2,\mathcal{N}}) \leq \mathcal{O}(N^2L^2\log_2 L\log_2 N). \tag{43}$$

□

Proof of Theorem 3.2 for $k=2$. The proofs of Theorem 3.2 for $k=2$ are analogous to those of Theorem 3.2 for $k=0$ by combining the results of Propositions C.9 and C.10. □

D. Proofs of Theorems 4.2 and 4.4

The sample errors in Theorems 4.2 and 4.4 can be obtained by applying Theorem3.2. The remaining problem is to establish the approximation error. For the approximations in Theorems 4.2, it can be estimated by the Lemma 4.1.

Lemma D.1. *Suppose $f \in L^2(\widehat{\Omega})$ and $\widehat{\Omega}$ denotes the domain belonging to $(0,1)^d$ with a smooth boundary, then $u^*(\boldsymbol{x}) \in H^2(\widehat{\Omega})$ where $u^*(\boldsymbol{x})$ is the exact solution of Eq. (6).*

Proof. The proof can be found in (Evans, 2022). □

Proof of Lemma 4.1. Denote $v = \phi(\boldsymbol{x};\boldsymbol{\theta}_{D,r}) - u^*$, then we have

$$\begin{aligned}&\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,r})) - \mathcal{E}_{D,r}(u^*) = \mathcal{E}_{D,r}(v+u^*) - \mathcal{E}_{D,r}(u^*)\\ =&\int_{\widehat{\Omega}} \nabla v\nabla u^*\,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\widehat{\Omega}}|\nabla v|^2\,\mathrm{d}\boldsymbol{x} - \int_{\widehat{\Omega}} fv\,\mathrm{d}\boldsymbol{x} + \lambda\int_{\partial\widehat{\Omega}} v^2\,\mathrm{d}s.\end{aligned} \tag{44}$$

Due to u^* being an exact solution and trace theorem inequality, we have

$$\int_{\widehat{\Omega}} \nabla v\nabla u^*\,\mathrm{d}\boldsymbol{x} - \int_{\widehat{\Omega}} fv\,\mathrm{d}\boldsymbol{x} = \int_{\partial\widehat{\Omega}} v\frac{\partial u^*}{\partial\nu}\,\mathrm{d}\boldsymbol{s} \leq C\|u^*\|_{H^2(\widehat{\Omega})}\|v\|_{H^1(\widehat{\Omega})}.$$

Therefore,

$$\mathcal{E}_{D,r}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,r})) - \mathcal{E}_{D,r}(u^*) = \frac{1}{2}\int_{\widehat{\Omega}}|\nabla v|^2\,\mathrm{d}\boldsymbol{x} + \lambda\int_{\partial\widehat{\Omega}} v^2\,\mathrm{d}s \tag{45}$$

Finally, due to the trace theorem inequality, we have that

$$\frac{1}{2}\int_{\widehat{\Omega}}|\nabla v|^2\,\mathrm{d}\boldsymbol{x} + \lambda\int_{\partial\widehat{\Omega}} v^2\,\mathrm{d}s \leq C\|v(\boldsymbol{x})\|_{H^1(\widehat{\Omega})} \leq C\|\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,r}) - u^*\|_{H^1(\widehat{\Omega})}.$$

□

For the approximation error in Theorem 4.4, we can rewrite it by the following lemma (Lemma 4.3):

Lemma D.2. *Suppose that $\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p}) \in \mathcal{F}_{B,2}$ and u^* is the exact solution of Eq. (6), then there is a constant C such that*

$$\mathcal{E}_{D,p}(\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p})) \leq C\|\phi(\boldsymbol{x};\boldsymbol{\theta}_{D,p}) - u^*\|_{H^2(\widehat{\Omega})}.$$

Proof. Due to u^* is the exact solution and trace theorem inequality, we know that

$$\begin{aligned}\mathcal{E}_{D,p}(\phi) &:= \int_{\widehat{\Omega}} |\Delta\phi(\boldsymbol{x};\boldsymbol{\theta}) - \Delta u^*(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} + \lambda \int_{\partial\widehat{\Omega}} |\phi(\boldsymbol{x};\boldsymbol{\theta}) - u^*(\boldsymbol{x})|^2 \,\mathrm{d}s, \\ &\leq C\|\phi(\boldsymbol{x};\boldsymbol{\theta}) - u^*(\boldsymbol{x})\|^2_{H^2(\widehat{\Omega})}\end{aligned} \tag{46}$$

□

E. Bounds of Covering Number for $\mathcal{F}$, $D_i\mathcal{F}$, $D_i^2\mathcal{F}$

Although we define $\mathcal{F}$ by σ_2-NNs, we only need ReLU NNs for $k = 0, 1$ (Schmidt-Hieber, 2020; Gühring et al., 2020). Therefore, in the following two lemms, we set the activation functions as ReLU only, as the square of ReLU can be generalized directly.

Lemma E.1. *The covering number of $\mathcal{F}$ can be bounded by*

$$\log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_\infty) \leq C\left[W(L+1)\log((F\vee 1)(W+1)) + W\log\left(\delta^{-1}(L+1)\right)\right],\ i = 1, \ldots, d.$$

where C is a constant[3] independent with W and L. Therefore, it is a $\mathcal{O}(W)$ term up to a logarithmic factor.

Proof. Given a network $f \in \mathcal{F}$ expressed as

$$f(\boldsymbol{x}) = (\boldsymbol{W}_{L+1}\sigma_1(\cdot) + \boldsymbol{b}_{L+1}) \circ \cdots \circ (\boldsymbol{W}_1 x + \boldsymbol{b}_1)$$

let

$$\mathcal{A}_k(f)(\boldsymbol{x}) = \sigma_1 \circ (\boldsymbol{W}_{k-1}\sigma_1(\cdot) + \boldsymbol{b}_{k-1}) \circ \cdots \circ (\boldsymbol{W}_1 x + \boldsymbol{b}_1)$$

and

$$\mathcal{B}_k(f)(\boldsymbol{x}) = (\boldsymbol{W}_{L+1}\sigma_1(\cdot) + \boldsymbol{b}_{L+1}) \circ \cdots \circ (\boldsymbol{W}_k\sigma_1(x) + \boldsymbol{b}_k),$$

for $k = 2, \ldots, L+1$. Corresponding to the last and first layer, we define $\mathcal{B}_{L+1}(f)(\boldsymbol{x}) = \boldsymbol{x}$ and $\mathcal{A}_1(f)(\boldsymbol{x}) = \boldsymbol{x}$. Then, it is easy to see that $f(\boldsymbol{x}) = \mathcal{B}_{k+1}(f) \circ (\boldsymbol{W}_k\sigma_1(\cdot) + \boldsymbol{b}_k) \circ \mathcal{A}_k(f)(\boldsymbol{x})$. Now, suppose that a pair of different two networks $f, g \in \mathcal{F}$ given by

$$f(\boldsymbol{x}) = (\boldsymbol{W}_{L+1}\sigma_1(\cdot) + \boldsymbol{b}_{L+1}) \circ \cdots \circ (\boldsymbol{W}_1 x + \boldsymbol{b}_1),\, g(\boldsymbol{x}) = \left(\boldsymbol{W}'_{L+1}\sigma_1(\cdot) + b'_{L+1}\right) \circ \cdots \circ (\boldsymbol{W}'_1\boldsymbol{x} + \boldsymbol{b}'_1),$$

has a parameters with distance $\delta : \|\boldsymbol{W}_\ell - \boldsymbol{W}'_\ell\|_\infty \leq \delta$ and $\|\boldsymbol{b}_\ell - \boldsymbol{b}'_\ell\|_\infty \leq \delta$. Now, note that

$$\|\mathcal{A}_k(f)\|_\infty \leq (F\vee 1)(W+1)\,\|\mathcal{A}_{k-1}(f)\|_\infty \leq (F\vee 1)^{k-1}(W+1)^{k-1},$$

and similarly the Lipshitz continuity of $\mathcal{B}_k(f)$ with respect to $\|\cdot\|_\infty$-norm is bounded as $(FW)^{L-k+1}$. Then, it holds that

$$\begin{aligned}&|f(\boldsymbol{x}) - g(\boldsymbol{x})| \\ =&\left|\sum_{k=1}^{L+1} \mathcal{B}_{k+1}(g) \circ (\boldsymbol{W}_k\sigma_1(\cdot) + \boldsymbol{b}_k) \circ \mathcal{A}_k(f)(\boldsymbol{x}) - \mathcal{B}_{k+1}(g) \circ (\boldsymbol{W}'_k\sigma_1(\cdot) + \boldsymbol{b}'_k) \circ \mathcal{A}_k(f)(\boldsymbol{x})\right| \\ \leq&\sum_{k=1}^{L+1}(FW)^{L-k}\left|(\boldsymbol{W}_k\sigma_1(\cdot) + \boldsymbol{b}_k) \circ \mathcal{A}_k(f)(\boldsymbol{x}) - (\boldsymbol{W}'_k\sigma_1(\cdot) + \boldsymbol{b}'_k) \circ \mathcal{A}_k(f)(\boldsymbol{x})\right| \\ \leq&\sum_{k=1}^{L+1}(FW)^{L-k}\delta\left[W(F\vee 1)^{k-1}(W+1)^{k-1} + 1\right] \\ \leq&\delta(L+1)(F\vee 1)^L(W+1)^{L+1}\end{aligned} \tag{47}$$

[3] In this section, we consistently employ the symbol C as a constant independent of W and L, which may vary from line to line.

Set the number of the parameters is CW where C is a constant independent with W and L. Then the covering number can be bounded by

$$(B(L+1)(F\vee 1)^L(W+1)^{L+1}\delta^{-1})^{CW}.$$

□

Lemma E.2. *The covering number of $\mathcal{F}$ can be bounded by*

$$\log \mathcal{N}\left(\delta, D_i\mathcal{F}, \|\cdot\|_\infty\right) \leq C\left[W(L+1)\log((F\vee 1)(W+1)) + W\log\left(\delta^{-1}(L+1)\right)\right],\ i=1,\dots,d.$$

where C is a constant independent with W and L. Therefore, it is a $\mathcal{O}(W)$ term up to a logarithmic factor.

Proof. First of all, we set and fix $i\in\{1,\dots,d\}$. For a DeNN with N width and L depth $f\in\mathcal{F}$, it can be represented as

$$\phi = \boldsymbol{W}_{L+1}\sigma(\boldsymbol{W}_L\sigma(\dots\sigma(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)\dots)+\boldsymbol{b}_L)+b_{L+1},$$

where σ can be either the ReLU or the ReLU square. Then the first order derivative can be read as

$$\begin{aligned}\psi(\boldsymbol{x}) = D_i\phi(\boldsymbol{x}) =& \boldsymbol{W}_{L+1}\sigma'(\boldsymbol{W}_L\sigma(\dots\sigma(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)\dots)+\boldsymbol{b}_L)\\ &\cdot \boldsymbol{W}_L\sigma'(\dots\sigma(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)\dots)\dots\boldsymbol{W}_2\sigma'(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)(\boldsymbol{W}_1)_i\\ &:= \prod_{k=1}^{L+1}\phi_k(\boldsymbol{x}),\end{aligned} \tag{48}$$

where $\boldsymbol{W}_i\in\mathbb{R}^{N_i\times N_{i-1}}$ ($(\boldsymbol{W})_i$ is i-th column of $\boldsymbol{W}$) and $\boldsymbol{b}_i\in\mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the i-th linear transform in ϕ, $\sigma'(\boldsymbol{x}) = \mathrm{diag}(\sigma'(x_i))$, and

$$\phi_k(\boldsymbol{x}) = \boldsymbol{W}_k\sigma'(\boldsymbol{W}_{k-1}\sigma(\dots\sigma(\boldsymbol{W}_1\boldsymbol{x}+\boldsymbol{b}_1)\dots)+\boldsymbol{b}_k),\ \phi_1 = (\boldsymbol{W}_1)_i.$$

$$\mathcal{B}_k(f)(\boldsymbol{x}) = (\boldsymbol{W}_{L+1}\sigma_1(\cdot)+\boldsymbol{b}_{L+1})\circ\dots\circ(\boldsymbol{W}_k\sigma_1(x)+\boldsymbol{b}_k),$$

for $k=2,\dots,L+1$. Corresponding to the proof of Lemma E.2, suppose a pair of different networks $f,g\in\mathcal{F}$ are given by

$$f(\boldsymbol{x}) = (\boldsymbol{W}_{L+1}\sigma_1(\cdot)+\boldsymbol{b}_{L+1})\circ\dots\circ(\boldsymbol{W}_1x+\boldsymbol{b}_1),\ g(\boldsymbol{x}) = \left(\boldsymbol{W}'_{L+1}\sigma_1(\cdot)+b'_{L+1}\right)\circ\dots\circ(\boldsymbol{W}'_1\boldsymbol{x}+\boldsymbol{b}'_1),$$

has a parameters with distance $\delta:\|\boldsymbol{W}_\ell-\boldsymbol{W}'_\ell\|_\infty\leq\delta$ and $\|\boldsymbol{b}_\ell-\boldsymbol{b}'_\ell\|_\infty\leq\delta$. Then we have

$$|f_k(\boldsymbol{x})-g_k(\boldsymbol{x})|\leq\delta k(F\vee 1)^{k-1}(W+1)^k, \tag{49}$$

where

$$D_if(\boldsymbol{x}) = \prod_{k=1}^{L+1}f_k(\boldsymbol{x}),\ D_ig(\boldsymbol{x}) = \prod_{k=1}^{L+1}g_k(\boldsymbol{x}).$$

As a result, we obtain

$$|f(\boldsymbol{x})-g(\boldsymbol{x})|\leq(L+1)^2\delta(F\vee 1)^L(W+1)^{L+1}F^L. \tag{50}$$

Set the number of the parameters is CW where C is a constant independent with W and L. Then the covering number can be bounded by

$$(B(L+1)^2(F\vee 1)^{2L}(W+1)^{L+1}\delta^{-1})^{CW}.$$

□

Lemma E.3. *The covering number of $D_i^2\mathcal{F}$ can be bounded by*

$$\log\mathcal{N}\left(\delta, D_i^2\mathcal{F}, \|\cdot\|_\infty\right)\leq C\left[W(L+1)\log((F\vee 1)(W+1)) + W\log\left(\delta^{-1}(L+1)\right)\right],\ i=1,\dots,d.$$

where C is a constant independent with W and L. Therefore, it is a $\mathcal{O}(W)$ term up to a logarithmic factor.

Proof. The proof is similar to that of Lemma E.2. □

F. Proof of Theorem 5.1

F.1. Rademacher complexity and related lemmas

Definition F.1 (Rademacher complexity (Anthony et al., 1999))**.** Given a sample set $S = \{z_1, z_2, \ldots, z_M\}$ on a domain $\mathcal{Z}$, and a class $\mathcal{F}$ of real-valued functions defined on $\mathcal{Z}$, the empirical Rademacher complexity of $\mathcal{F}$ in S is defined as

$$\mathbf{R}_S(\mathcal{F}) := \frac{1}{M}\mathbf{E}_{\Sigma_M}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{M}\sigma_i f(z_i)\right],$$

where $\Sigma_M := \{\sigma_1, \sigma_2, \ldots, \sigma_M\}$ are independent random variables drawn from the Rademacher distribution, i.e., $\mathbf{P}(\sigma_i = +1) = \mathbf{P}(\sigma_i = -1) = \frac{1}{2}$ for $i = 1, 2, \ldots, M$. For simplicity, if $S = \{z_1, z_2, \ldots, z_M\}$ is an independent random variable set with the uniform distribution, denote $\mathbf{R}_M(\mathcal{F}) := \mathbf{E}_S\mathbf{R}_S(\mathcal{F})$.

The following lemma will be used to bounded generalization error by Rademacher complexities:

Lemma F.2 ((Wainwright, 2019), Proposition 4.11)**.** *Let $\mathcal{F}$ be a set of functions. Then*

$$\mathbf{E}_X \sup_{u\in\mathcal{F}}\left|\frac{1}{M}\sum_{i=1}^{M}u(x_j) - \mathbf{E}_{x\sim\mathcal{P}_\Omega}u(x)\right| \le 2\mathbf{R}_M(\mathcal{F}),$$

where $X := \{x_1, \ldots, x_M\}$ is an independent random variable set with the uniform distribution.

Lemma F.3 ((Jiao et al., 2021))**.** *Assume that $w : \Omega \to \mathbb{R}$ and $|w(\boldsymbol{x})| \le B$ for all $\boldsymbol{x} \in \Omega$, then for any function class $\mathcal{F}$, there holds*

$$\mathbf{R}_M(w\cdot\mathcal{F}) \le B\mathbf{R}_M(\mathcal{F})$$

where $w\cdot\mathcal{F} := \{\bar{u} : \bar{u}(\boldsymbol{x}) = w(\boldsymbol{x})u(\boldsymbol{x}), u \in \mathcal{F}\}$.

Next, we present a lemma of uniform covering numbers. These concepts are essential for estimating the Rademacher complexity in the context of generalization error. Subsequently, we derive an estimate for the uniform covering number, leveraging the concept of pseudo-dimension.

Lemma F.4 (Dudley's theorem (Anthony et al., 1999))**.** *Let $\mathcal{F}$ be a function class such that $\sup_{f\in\mathcal{F}}\|f\|_\infty \le B$. Then the Rademacher complexity $\mathbf{R}_M(\mathcal{F})$ satisfies that*

$$\mathbf{R}_M(\mathcal{F}) \le \inf_{0\le\delta\le B}\left\{4\delta + \frac{12}{\sqrt{M}}\int_\delta^B \sqrt{\log 2\mathcal{N}(\varepsilon, \mathcal{F}, M)}\,\mathrm{d}\varepsilon\right\}$$

F.2. Proof of Theorem 5.1

There are three cases in Theorem 5.1. In order to prove Theorem 5.1, we consider these three cases individually. For simplicity in notation, we still denote the width as $N = \mathcal{O}(\log L)$ in $\mathcal{F}_{B,k}$ in the following proofs.

F.2.1. Proof of Theorem 5.1 for $k = 0$.

We first show the proof of the Proposition F.5:

Proposition F.5. *Let $d, N, L, M \in \mathbb{N}_+$, B, C_1, C_2[4] $\in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \le 1$, we have*

$$\begin{aligned}&\mathbf{E}\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0})\\ \le &\inf_{\phi\in\mathcal{F}_{B,0}}\|\phi - f\|_{L^2([0,1]^d)} + 4\mathbf{R}_M(\mathcal{F}_{B,0}) + 2\mathbf{R}_M(\mathcal{F}_{B,0}^2),\end{aligned}$$

where $\mathcal{F}_{B,0}^2 := \{\bar{\phi} : \bar{\phi} = \phi^2, \phi \in \mathcal{F}_{B,0}\}$, $\mathbf{E}$ is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$.

[4]Here, the occurrences of C_1 and C_2 in $\mathcal{B}_{B,0}$ continue to hold throughout the rest of the propositions and theorems in this paper.

Proof. Due to the definition, we know that

$$\mathcal{R}_{D,0}(\boldsymbol{\theta}_{D,0}) = \inf_{\phi \in \mathcal{F}_{B,0}} \|\phi - f\|_{L^2([0,1]^d)}.$$

For the sample error, due to $\boldsymbol{\theta}_{S,0}$ and $\boldsymbol{\theta}_{D,0}$ belong to Θ almost surely, we have

$$\begin{aligned}
&|\mathbf{E}\left[\mathcal{R}_{S,0}(\boldsymbol{\theta}_{S,0}) - \mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0})\right]| \\
=&\left|\mathbf{E}\left[\frac{1}{M}\sum_{i=1}^{M}|f(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})|^2 - \int_{[0,1]^d}|f(\boldsymbol{x}) - \phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})|^2\,\mathrm{d}\boldsymbol{x}\right]\right| \\
\leq& 2\left|\mathbf{E}\left(\frac{1}{M}\sum_{i=1}^{M} f(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0}) - \int_{[0,1]^d} f(\boldsymbol{x})\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})\,\mathrm{d}\boldsymbol{x}\right)\right| \\
&+\left|\mathbf{E}\left(\frac{1}{M}\sum_{i=1}^{M}\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0})^2 - \int_{[0,1]^d}\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})^2\,\mathrm{d}\boldsymbol{x}\right)\right| \\
\leq& 2\mathbf{R}_M(f\mathcal{F}_{B,0}) + \mathbf{R}_M(\mathcal{F}_{B,0}^2)
\end{aligned} \tag{51}$$

where the last inequality is due to Lemma F.2.

Due to Lemmas F.2 and F.3, and $|f| \leq 1$, we have that

$$\mathbf{E}\left|\frac{1}{M}\sum_{i=1}^{M} f(\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,0}) - \int_{[0,1]^d} f(\boldsymbol{x})\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,0})\,\mathrm{d}\boldsymbol{x})\right| \leq 2\mathbf{R}_M(f\mathcal{F}_{B,0}) \leq 2\mathbf{R}_M(\mathcal{F}_{B,0}). \tag{52}$$

Therefore, we have

$$\mathbf{E}\mathcal{R}_{S,0}(\boldsymbol{\theta}_{S,0}) - \mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) \leq 4\mathbf{R}_M(\mathcal{F}_{B,0}) + 2\mathbf{R}_M(\mathcal{F}_{B,0}^2). \tag{53}$$

□

Next we need to bound $\mathrm{Pdim}(\mathcal{F}_{B,0})$ and $\mathrm{Pdim}(\mathcal{F}_{B,0}^2)$. Based on (Bartlett et al., 2019), $\mathrm{Pdim}(\mathcal{F}_{B,0}) = \mathcal{O}(L^2N^2\log_2 L\log_2 N)$. For the $\mathrm{Pdim}(\mathcal{F}_{B,0}^2)$, we can estimate it by the simlaler way of $\mathrm{Pdim}(\mathcal{F}_{B,0})$.

Proposition F.6. *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\widehat{C}$ independent with N, L such that*

$$\textit{Pdim}(\mathcal{F}_{B,0}^2) \leq \widehat{C}N^2L^2\log_2 L\log_2 N. \tag{54}$$

Proof. Denote

$$\mathcal{F}_{B,0,\mathcal{N}}^2 := \{\eta(\boldsymbol{x},y) : \eta(\boldsymbol{x},y) = \psi(\boldsymbol{x}) - y, \psi \in \mathcal{F}_{B,0}^2, (\boldsymbol{x},y) \in \mathbb{R}^{d+1}\}.$$

Based on the definition of VC-dimension and pseudo-dimension, we have that

$$\mathrm{Pdim}(\mathcal{F}_{B,0}^2) \leq \mathrm{VCdim}(\mathcal{F}_{B,0,\mathcal{N}}^2). \tag{55}$$

For the $\mathrm{VCdim}(\mathcal{F}_{B,0,\mathcal{N}}^2)$, it can be bounded by following way. The proof is similar to that in (Bartlett et al., 2019).

For a DeNN with N width and L depth, it can be represented as

$$\phi = \boldsymbol{W}_{L+1}\sigma_1(\boldsymbol{W}_L\sigma_1(\ldots\sigma_1(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\ldots) + \boldsymbol{b}_L) + b_{L+1},$$

where $\boldsymbol{W}_i \in \mathbb{R}^{N_i \times N_{i-1}}$ ($(\boldsymbol{W})_i$ is i-th column of $\boldsymbol{W}$) and $\boldsymbol{b}_i \in \mathbb{R}^{N_i}$ are the weight matrix and the bias vector in the i-th linear transform in ϕ. Denote W_i as the number of parameters in $\boldsymbol{W}_i, \boldsymbol{b}_i$, i.e., $W_i = N_iN_{i-1} + N_i$.

Let $\bar{\boldsymbol{x}} = (\boldsymbol{x}, y) \in \mathbb{R}^{d+1}$ be an input and $\boldsymbol{\theta} \in \mathbb{R}^W$ be a parameter vector in $\psi := \phi^2 - y$. We denote the output of ψ with input $\boldsymbol{x}$ and parameter vector $\boldsymbol{\theta}$ as $f(\boldsymbol{x}, \boldsymbol{\theta})$. For fixed $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ in $\mathbb{R}^d$, we aim to bound

$$K := \left|\{(\mathrm{sgn}(f(\boldsymbol{x}_1,\boldsymbol{\theta})), \ldots, \mathrm{sgn}(f(\boldsymbol{x}_m,\boldsymbol{\theta}))) : \boldsymbol{\theta} \in \mathbb{R}^W\}\right|. \tag{56}$$

The proof is inspired by (Bartlett et al., 2019, Theorem 7). For any partition $\mathcal{S} = \{P_1, P_2, \ldots, P_T\}$ of the parameter domain $\mathbb{R}^W$, we have $K \leq \sum_{i=1}^{T} |\{(\mathrm{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \mathrm{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in P_i\}|$. We choose the partition such that within each region P_i, the functions $f(\boldsymbol{x}_j, \cdot)$ are all fixed polynomials of bounded degree. This allows us to bound each term in the sum using Lemma C.6.

Due to proof of (Bartlett et al., 2019, Theorem 7), note that there is a partition $\mathcal{S}$ where

$$|\mathcal{S}| \leq \prod_{n=1}^{L-1} 2 \left(\frac{2emnN_n}{\sum_{i=1}^{n} W_i} \right)^{\sum_{i=1}^{n} W_i}.$$

And the output ϕ is a polynomial function in $\sum_{i=1}^{L} W_i$ variables of $\boldsymbol{\theta} \in S \in \mathcal{S}_L$, of total degree no more than L. Hence the output ψ is a polynomial function in $\sum_{i=1}^{L} W_i$ variables of $\boldsymbol{\theta} \in S \in \mathcal{S}_L$, of total degree no more than $2L$. Therefore, for each $S \in \mathcal{S}_L$ we have

$$|\{(\mathrm{sgn}(f(\boldsymbol{x}_1, \boldsymbol{\theta})), \ldots, \mathrm{sgn}(f(\boldsymbol{x}_m, \boldsymbol{\theta}))) : \boldsymbol{\theta} \in S\}| \leq 2 \left(4emL / \sum_{i=1}^{L} W_i \right)^{\sum_{i=1}^{L} W_i}.$$

Then

$$\begin{aligned} K \leq & 2 \left(4emL / \sum_{i=1}^{L} W_i \right)^{\sum_{i=1}^{L} W_i} \cdot \prod_{n=1}^{L} 2 \left(\frac{2emnN_n}{\sum_{i=1}^{n} W_i} \right)^{\sum_{i=1}^{n} W_i} \leq \prod_{n=1}^{L} 2 \left(\frac{4emnN_n}{\sum_{i=1}^{n} W_i} \right)^{\sum_{i=1}^{n} W_i} \\ \leq & 2^L \left(\frac{2em(L+1)LN}{U} \right)^U \end{aligned} \tag{57}$$

where $U := \sum_{n=1}^{L} \sum_{i=1}^{n} W_i = \mathcal{O}(N^2 L^2)$, N is the width of the network, and the last inequality is due to weighted AM-GM. For the definition of the VC-dimension, we have

$$2^{\mathrm{VCdim}(\mathcal{F}^2_{B,0,\mathcal{N}})} \leq 2^L \left(\frac{2e\mathrm{VCdim}(\mathcal{F}^2_{B,0,\mathcal{N}})(L+1)LN}{U} \right)^U. \tag{58}$$

Due to Lemma C.7, we obtain that

$$\mathrm{VCdim}(\mathcal{F}^2_{B,0,\mathcal{N}}) \leq L + U \log_2[4(L+1)L \log_2(L+1)L] = \mathcal{O}(N^2 L^2 \log_2 L \log_2 N) \tag{59}$$

since $U = \mathcal{O}(N^2 L^2)$. □

Now we can show the proof of Theorem 3.2 for $k = 0$:

Proof of Theorem 3.2 for $k = 0$. Let $J = \max\{\mathrm{Pdim}(\mathcal{F}^2_{B,0}), \mathrm{Pdim}(\mathcal{F}_{B,0})\}$. Due to Lemma F.4, B.3 and Theorem F.6, for any $M \geq J$, we have

$$\begin{aligned} \mathbf{R}_M(\mathcal{F}^2_{B,0}) \leq & 4\delta + \frac{12}{\sqrt{M}} \int_\delta^B \sqrt{\log 2\mathcal{N}(\varepsilon, \mathcal{F}^2_{B,0}, M)}\, \mathrm{d}\varepsilon \\ \leq & 4\delta + \frac{12}{\sqrt{M}} \int_\delta^B \sqrt{\log 2 \left(\frac{2eMB}{\varepsilon \mathrm{Pdim}(\mathcal{F}^2_{B,0})} \right)^{\mathrm{Pdim}(\mathcal{F}^2_{B,0})}}\, \mathrm{d}\varepsilon \\ \leq & 4\delta + \frac{12B}{\sqrt{M}} + 12 \left(\frac{\mathrm{Pdim}(\mathcal{F}^2_{B,0})}{M} \right)^{\frac{1}{2}} \int_\delta^B \sqrt{\log \left(\frac{2eMB}{\varepsilon \mathrm{Pdim}(\mathcal{F}^2_{B,0})} \right)}\, \mathrm{d}\varepsilon. \end{aligned} \tag{60}$$

By the direct calculation for the integral, we have

$$\int_\delta^B \sqrt{\log \left(\frac{2eMB}{\varepsilon \mathrm{Pdim}(\mathcal{F}^2_{B,0})} \right)}\, \mathrm{d}\varepsilon \leq B \sqrt{\log \left(\frac{2eMB}{\delta \mathrm{Pdim}(\mathcal{F}^2_{B,0})} \right)}.$$

Then choosing $\delta = B\left(\frac{\text{Pdim}(\mathcal{F}_{B,0}^2)}{M}\right)^{\frac{1}{2}} \leq B$, we have

$$\mathbf{R}_M(\mathcal{F}_{B,0}^2) \leq 28B\left(\frac{\text{Pdim}(\mathcal{F}_{B,0}^2)}{M}\right)^{\frac{1}{2}}\sqrt{\log\left(\frac{2eM}{\text{Pdim}(\mathcal{F}_{B,0}^2)}\right)}. \tag{61}$$

Therefore, due to Theorem F.6, there is a constant C_4 independent with L, N, M such as

$$\mathbf{R}_M(\mathcal{F}_{B,0}^2) \leq C_4\frac{NL(\log_2 L\log_2 N)^{\frac{1}{2}}}{\sqrt{M}}\log M. \tag{62}$$

$\mathbf{R}_M(\mathcal{F}_{B,0})$ can be estimate in the similar way. Therefore, we have that there is a constant $C_5 = C_5(B, d, C_1, C_2)$ such that

$$\mathbf{E}\mathcal{R}_{D,0}(\boldsymbol{\theta}_{S,0}) \leq C_{3,0}N^{-2n/d}L^{-2n/d} + C_5\frac{NL(\log_2 L\log_2 N)^{\frac{1}{2}}}{\sqrt{M}}\log M. \tag{63}$$

Finally, we set $N = \mathcal{O}(\log L)$ according to the definition of DeNNs, and the number of parameters $W = \mathcal{O}(N^2 L\log L) = \mathcal{O}(L(\log L)^3)$ to complete the proof. □

F.2.2. Proof of Theorem 5.1 for $k = 1$

Proposition F.7. *Let $d, N, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, we have*

$$\mathbf{E}\mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1}) \leq \inf_{\phi\in\mathcal{F}_{B,1}}\|\phi - f\|_{H^1([0,1]^d)} + \mathcal{E}_1 + \mathcal{E}_0,$$

where $\mathcal{E}_1 = 4d\mathbf{R}_M(D\mathcal{F}_{B,1}) + 2\mathbf{R}_M(D\mathcal{F}_{B,1}^2)$ and $\mathcal{E}_0 = 4\mathbf{R}_M(\mathcal{F}_{B,1}) + 2\mathbf{R}_M(\mathcal{F}_{B,1}^2)$ and

$$\begin{aligned}
&D\mathcal{F}_{B,1}^2 := \{\bar{\phi} : \bar{\phi} = (D_i\phi)^2, \phi \in \mathcal{F}_{B,1},\ D_i \text{ is the weak derivative in the } i\text{-th variable}\},\\
&D\mathcal{F}_{B,1} := \{\bar{\phi} : \bar{\phi} = D_i\phi, \phi \in \mathcal{F}_{B,1},\ D_i \text{ is the weak derivative in the } i\text{-th variable}\},
\end{aligned} \tag{64}$$

$\mathbf{E}$ is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$.

Proof. Due to the definition, we know that

$$\mathcal{R}_{D,1}(\boldsymbol{\theta}_{D,1}) = \inf_{\phi\in\mathcal{F}_{B,1}}\|\phi - f\|_{H^1([0,1]^d)}.$$

For the sample error, due to $\boldsymbol{\theta}_{S,1}$ and $\boldsymbol{\theta}_{D,1}$ belong to Θ almost surely, we have

$$\begin{aligned}
&|\mathbf{E}\left[\mathcal{R}_{S,1}(\boldsymbol{\theta}_{S,1}) - \mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1})\right]|\\
\leq&\left|\mathbf{E}\left[\frac{1}{M}\sum_{i=1}^M|\nabla f(\boldsymbol{x}_i) - \nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})|^2 - \int_{[0,1]^d}|\nabla f(\boldsymbol{x}) - \nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})|^2\,\mathrm{d}\boldsymbol{x} + \mathcal{E}_0\right]\right|\\
\leq&2\left|\mathbf{E}\left(\frac{1}{M}\sum_{i=1}^M\nabla f(\boldsymbol{x}_i)\nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1}) - \int_{[0,1]^d}\nabla f(\boldsymbol{x})\nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})\,\mathrm{d}\boldsymbol{x}\right)\right|\\
&+\left|\mathbf{E}\left(\frac{1}{M}\sum_{i=1}^M\nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})^2 - \int_{[0,1]^d}\nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})^2\,\mathrm{d}\boldsymbol{x}\right)\right| + \mathcal{E}_0\\
\leq&2\mathbf{R}_M(\nabla f D\mathcal{F}_{B,1}) + \mathbf{R}_M(D\mathcal{F}_{B,1}^2) + \mathcal{E}_0
\end{aligned} \tag{65}$$

where the last inequality is due to Lemma F.2 and $\mathcal{E}_0 = 4\mathbf{R}_M(\mathcal{F}_{B,1}) + 2\mathbf{R}_M(\mathcal{F}_{B,1}^2)$.

Due to Lemmas F.2 and F.3, and $|\nabla f| \leq d$, we have that

$$\begin{aligned}&\mathbf{E}\left|\frac{1}{M}\sum_{i=1}^{M}\nabla f(\boldsymbol{x}_i)\nabla\phi(\boldsymbol{x}_i;\boldsymbol{\theta}_{S,1})-\int_{[0,1]^d}\nabla f(\boldsymbol{x})\nabla\phi(\boldsymbol{x};\boldsymbol{\theta}_{S,1})\,\mathrm{d}\boldsymbol{x})\right|\\ \leq& 2d\mathbf{R}_M(fD\mathcal{F}_{B,1})\leq 2d\mathbf{R}_M(D\mathcal{F}_{B,1}).\end{aligned} \tag{66}$$

Therefore, we have

$$\mathbf{E}\mathcal{R}_{S,1}(\boldsymbol{\theta}_{S,1})-\mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1})\leq\mathcal{E}_1+\mathcal{E}_0, \tag{67}$$

where $\mathcal{E}_1 = 4d\mathbf{R}_M(D\mathcal{F}_{B,1}) + 2\mathbf{R}_M(D\mathcal{F}_{B,1}^2)$. □

To further refine our understanding, we aim to estimate $\mathbf{R}_M(D\mathcal{F}_{B,1})$ and $\mathbf{R}_M(D\mathcal{F}_{B,1}^2)$ in order to bound the generalization error. The pseudo-dimension can be used to provide such bounds. Notably, for DeNN derivatives, (Yang et al., 2023b) presents nearly optimal bounds for the pseudo-dimension. In the upcoming propositions, we will demonstrate the pseudo-dimension of $D\mathcal{F}_{B,1}^2$.

Proposition F.8. *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\bar{C}$ independent with N, L such that*

$$\textit{Pdim}(D\mathcal{F}_{B,1}^2)\leq\bar{C}N^2L^2\log_2 L\log_2 N. \tag{68}$$

Proof. The proof is similar to those of Propositions C.8 and F.6. □

Proof of Theorem 5.1 for $k = 1$. The proofs of Theorem 5.1 for $k = 1$ are analogous to those of Theorem 5.1 for $k = 0$ by combining the results of Propositions F.7 and F.8. □

F.2.3. Proof of Theorem 5.1 for $k = 2$

Proposition F.9. *Let $d, N, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. For any $f \in W^{n,\infty}([0,1]^d)$ with $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$, we have*

$$\mathbf{E}\mathcal{R}_{D,1}(\boldsymbol{\theta}_{S,1})\leq\inf_{\phi\in\mathcal{F}_{B,2}}\|\phi-f\|_{H^2([0,1]^d)}+\mathcal{E}_2+\mathcal{E}_0,$$

where $\mathcal{E}_2 = 4d\mathbf{R}_M(D^2\mathcal{F}_{B,2}) + 2\mathbf{R}_M(D^2\mathcal{F}_{B,2}^2)$ and $\mathcal{E}_0 = 4\mathbf{R}_M(\mathcal{F}_{B,2}) + 2\mathbf{R}_M(\mathcal{F}_{B,2}^2)$ and

$$\begin{aligned}D^2\mathcal{F}_{B,2}^2&:=\{\bar{\phi}:\bar{\phi}=(D_i^2\phi)^2,\phi\in\mathcal{F}_{B,2},\ D_i^2\ \textit{is the second weak derivative in the } i\textit{-th variable}\},\\ D^2\mathcal{F}_{B,2}&:=\{\bar{\phi}:\bar{\phi}=D_i^2\phi,\phi\in\mathcal{F}_{B,2},\ D_i^2\ \textit{is the second weak derivative in the } i\textit{-th variable}\},\end{aligned} \tag{69}$$

$\mathbf{E}$ *is expected responding to X, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is an independent random variables set uniformly distributed on $[0,1]^d$.*

Proof. The proof is similar to those of Proposition F.7. □

To further refine our understanding, we aim to estimate $\mathbf{R}_M(D^2\mathcal{F}_{B,2})$ and $\mathbf{R}_M(D^2\mathcal{F}_{B,2}^2)$ in order to bound the generalization error. The pseudo-dimension can be used to provide such bounds. Notably, for DeNN derivatives, (Yang et al., 2023b) presents nearly optimal bounds for the pseudo-dimension. In the upcoming propositions, we will demonstrate the pseudo-dimension of $D^2\mathcal{F}_{B,2}^2$.

Proposition F.10. *For any $N, L, d \in \mathbb{N}_+$, there exists a constant $\bar{C}$ independent with N, L such that*

$$\textit{Pdim}(D^2\mathcal{F}_{B,2}^2)\leq\bar{C}N^2L^2\log_2 L\log_2 N. \tag{70}$$

Proof. The proof is similar to those of Propositions C.10 and F.6. □

Proof of Theorem 3.2 for $k = 2$. The proofs of Theorem 3.2 for $k = 2$ are analogous to those of Theorem 3.2 for $k = 0$ by combining the results of Propositions F.9 and F.10. □

G. Experiment

Target function: Here we define the target function as

$$f(\boldsymbol{x}) : [0,1]^2 \mapsto \mathbb{R} \tag{71}$$

with

$$f(\boldsymbol{x}) = \left(\sum_{i=1}^{n} a_i \sigma(w_i \cdot x + b_i)\right) \times \left(\sum_{i=1}^{n} \alpha_i \sigma(\omega_i \cdot x + \beta_i)\right) \tag{72}$$

where $\sigma = \mathrm{ReLU}$, $n = 1000$, and $a_i, w_i, b_i, \alpha_i, \omega_i, \beta_i$ are randomly chosen by the default sampling strategies in PyTorch. As a result,

$$f(\boldsymbol{x}) \in W^{1,\infty}\left([0,1]^2\right). \tag{73}$$

Large Data Regime: In this scenario, both networks are trained and tested on datasets with $10,000$ points sampled from the domain $[0,1] \times [0,1]$.

Small Data Regime: In this case, the training and testing are performed on a smaller dataset of $1,000$ samples from the same domain.

Test and results: We randomly sample 10,000 points for testing for both large and small data cases. Results shown in Table 3 are based on 3 tests.

Neural Network	Large Data Regime	Small Data Regime
Shallow (Depth 1, Width 20)	Mean: 6.18e-4, Std: 9.86e-05	Mean: 1.51e-3, Std: 2.52e-4
Deep (Depth 4, Width 10)	Mean: 3.69e-4, Std: 2.57e-05	Mean: 4.96e-3, Std: 4.46e-3

Table 3. This table compares the performance of shallow and deep neural networks in terms of mean test performance and standard deviation across different data regimes. It illustrates how network depth and data availability impact learning outcomes, with shallow networks performing better in small data scenarios, while deep networks excel with larger datasets.