MACHINE LEARNING MEETS NUMERICAL PDE: FAST RATE, NEURAL SCALING LAW AND MINIMAX OPTIMALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we study the statistical limits of deep learning techniques for solving a class of high-dimensional partial differential equations (PDEs) using the Deep Ritz Method (DRM) and Physics Informed Neural Networks (PINNs). To simplify the problem, we focus on the Schrödinger equation on a hypercube with zero Dirichlet boundary conditions. We establish upper and lower bounds for both methods and prove that PINN and a modified version of DRM can achieve minimax optimal bounds, which improve upon concurrent developed upper bounds for this problem. Empirically, following recent work which has shown that the deep model accuracy will improve with growing training sets according to a power law, we discuss a similar behavior of this power law for deep PDE solvers. We supply computational experiments which substantiate our theoretical results.

1 INTRODUCTION

Partial differential equations (PDEs) play a prominent role in many disciplines of science and engineering. The recent deep learning breakthrough and the rapid development of sensors, computational power, and data storage in the past decade draws attention to numerically solving PDEs via machine learning methods (Long et al., 2018; 2019; Raissi et al., 2019; Han et al., 2018; Sirignano & Spiliopoulos, 2018; Khoo et al., 2017), especially in high-dimension where conventional methods become impractical. The set of applications which motivate this interest is wide ranging including, computational physics (Han et al., 2018; Long et al., 2018; Raissi et al., 2019), inverse problem (Zhang et al., 2018; Gilton et al., 2019; Fan & Ying, 2020) and quantitative finance (Heaton et al., 2017; Germain et al., 2021). The numerical methods generated by the use of deep learning techniques are mesh-less methods, see the discussion in (Xu, 2020). A natural deep learning technique in these problems based on a of standard feed-forward type architecture consists in taking advantage (when available) of a variational formulation whose solution coincides with the solution of the PDE of interest. Despite the success and popularity of adopting neural networks for solving high-dimensional PDEs, the following question still remain poorly answered.

> *For a given PDE and data driven approximation architecture, how large a sample size and how complex such model is needed to reach a prescribed performance level?*

In this paper, we aim to establish the numerical analysis of such deep learning based PDE solvers. Inspired by recent works which showed that the empirical performance of a model is remarkably predictable via a power law of the data number, known as the neural scaling law (Kaplan et al., 2020; Hestness et al., 2017; Sharma & Kaplan, 2020), we aim to explore the neural scaling law for deep PDE solvers and compare its performance to Fourier approximation.

Among the various approaches of using deep learning methods for solving PDEs, in this work, we will focus on the deep Ritz method (DRM) (E & Yu, 2018; Khoo et al., 2017) and the Physics-Inspired Neural Networks (PINN) approach (Sirignano & Spiliopoulos, 2018; Raissi et al., 2019), both are based on minimizing neural network parameters according to some loss funcitonal related to the PDEs. To provide theoretically guarantees for DRM and PINN, following (Lu et al., 2021b; Duan et al., 2021; Bai et al., 2021), we decompose the error into approximation error (Yarotsky, 2017;

Suzuki, 2018; Shen et al., 2021) and generalization error (Bartlett et al., 2005; Yarotsky, 2017; Xu & Zeevi, 2020). However instead of the $O(1/\sqrt{n})$ (n is the number of data sampled) slow rate generalization bounds established in prior work (Lu et al., 2021b; Shen et al., 2021; Xu, 2020; Shin et al., 2020), we utilize the strongly convex structure of the DRM and PINN objective and provide a $O(1/n)$ fast rate generalization bound (Bartlett et al., 2005; Xu & Zeevi, 2020) which lead us to a non-parametric estimation bound. Our theory also suggests optimal selection of network size with respect to the number of sampled data. Moreover, to illustrate the optimiality of our upper bound, we also establish an information-theoretic lower bound which matches our upper bound for PINN and a modified version of DRM.

We also test our theory by numerical experiments. Recent works (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2019; Mikami et al., 2021) studying a variety of deep learning algorithms all find the same polynomial scaling relation between the testing error with the number of data. As the number of training data n is increased, the population loss $\mathcal{L}$ of well-trained and well-tuned models will scale with n as a power-law $\mathcal{L} \propto \frac{1}{n^\alpha}$ for some α. (Sharma & Kaplan, 2020) also scans over a large range of α and problem dimension d and find an approximately $\alpha \propto \frac{1}{d}$ scaling law. In Section 4, we conduct numerical experiments to show this phenomenon still holds for deep PDE solver and this neural scaling law test more idiosyncratic features of the theory.

1.1 Related Works

Neural Scaling Law The starting point of our work is the recent observation across speech, vision and text (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2019; Rosenfeld, 2021) that the empirical performance of a model satisfies a power law scales as a power-law with model size and dataset size. (Sharma & Kaplan, 2020) further finds out that the power of the scaling law is depends on the intrinsic dimension of the dataset. Theoretical works (Schmidt-Hieber et al., 2020; Suzuki, 2018; Suzuki & Nitanda, 2019; Chen et al., 2019b; Imaizumi & Fukumizu, 2020; Farrell et al., 2021; Jiao et al., 2021c) explore the optimal power law under the non-parametric curve estimation setting via a plug-in neural network. Our work extends this line of research to solving PDEs.

Deep Network Based PDE Solver. Soliving high dimensional partial differential equations (PDEs) have been a long-standing challenge due to the curse of dimensionality. At the same time, deep learning has shown superior flexibility and adaptivity in approximating high dimensional functions which leads to state-of-the-art performances in a wide range of tasks ranging from computer vision to natural language processing. Recent years, pioneer works (Han et al., 2018; Raissi et al., 2019; Long et al., 2018; Sirignano & Spiliopoulos, 2018; Khoo et al., 2017) have tried to utilize the deep neural networks to solve different types of PDEs and achieves impressive results in many tasks (Lu et al., 2021a; Li et al., 2020). Based on the natural idea to represent solutions of PDEs by (deep) neural networks, different loss functions for solving PDEs are proposed. (Han et al., 2018; 2020) utilize the Feyman-Kac formulation which turns solving PDE to a stochastic control problem and the weak adversarial network (Zang et al., 2020) solves the weak formulation via an adversarial network. In this paper, we focus on the convergence rate of the Deep Ritz Method (DRM) (E & Yu, 2018; Khoo et al., 2017) and Physic–informed neural network(PINN) (Raissi et al., 2019; Sirignano & Spiliopoulos, 2018). DRM(E & Yu, 2018; Khoo et al., 2017) utilize the variational structure of the PDE, similar to the Ritz-Galerkin method in classical numerical analysis of PDEs, and trains the neural network to minimize the variational objective. PINN(Raissi et al., 2019; Sirignano & Spiliopoulos, 2018) train the neural network directly to minimize the residual of the PDE.

Theoretical Guarantee For Machine Learning Based PDE Solver. Theoretical convergence results for deep learning based PDE solvers raises wide interest recently. Specifically, (Lu et al., 2021b; Grohs & Herrmann, 2020; Marwah et al., 2021; Wojtowytsch et al., 2020; Xu, 2020; Shin et al., 2020; Bai et al., 2021) investigated the regularity of PDEs approximated by neuarl network and (Lu et al., 2021b; Luo & Yang, 2020) further provided a generalization analysis. (Nickl et al., 2020) considered a prior and an equivalent white noise model (Brown & Low, 1996) and considered the rate of convergence of the posterior. Our paper does not need to introduce and provided a non-asymptotic result. At the same time, (Nickl et al., 2020) can only be applied to linear PDEs while our proof technique can be extend to nonlinear one. All these paper also failed to answer the question that how to determine the network size corresponding to the sampled data number to achieve a desired statistical convergence rate. (Hütter & Rigollet, 2019; Manole et al., 2021) consider the similar

Upper Bounds				Lower Bound
Objective Function	Neural Network	Previous Bound	Fourier Basis	
Deep Ritz	$n^{-\frac{2s-2}{d+2s-2}}\log n$	$n^{-\frac{2s-2}{d+4s-4}}\log n$ (Duan et al., 2021)	$n^{-\frac{2s-2}{d+2s-2}}$	$n^{-\frac{2s-2}{d+2s-4}}$
Modified Deep Ritz	$n^{-\frac{2s-2}{d+2s-2}}\log n$	/	$n^{-\frac{2s-2}{d+2s-4}}$	$n^{-\frac{2s-2}{d+2s-4}}$
PINN	$n^{-\frac{2s-4}{d+2s-4}}\log n$	$n^{-\frac{2s-4}{d+4s-8}}\log n$ (Jiao et al., 2021a)	$n^{-\frac{2s-4}{d+2s-4}}$	$n^{-\frac{2s-4}{d+2s-4}}$

Table 1: Upper bounds and lower bounds we achieved in this paper and previous work. The upper bound colored in red indicates the convergence rate matches the min-max lower bound.

problem for the optimal transport problem, *i.e.* Monge-ampere equation. However the variational problem we considered is different from (Hütter & Rigollet, 2019; Manole et al., 2021) and leads to technical difference. The most related paper to us is a concurrent paper (Duan et al., 2021; Jiao et al., 2021a;b). However our upper is faster than (Duan et al., 2021; Jiao et al., 2021a;b) and we have shown that the lower bound conjecture in (Duan et al., 2021) is wrong. In this paper, we showed that generalization analysis in (Lu et al., 2021b; Duan et al., 2021; Luo & Yang, 2020) are loose due to lack of localization technique (De Boor & De Boor, 1978; Bartlett et al., 2005; Koltchinskii, 2011; Xu, 2020). With observation of the strong convexity of the loss function, we follow the fast rate results for ERM (Schmidt-Hieber et al., 2020; Xu & Zeevi, 2020; Farrell et al., 2021) and provided a near optimal bound for both DRM and PINN.

1.2 Contribution

In short, we summarize our contribution as following

- In this paper, we first consider the statistical limit of learning a PDE solution from sampled observations. The lower bound showed a non-standard exponent different from non-parametric estimating a function, which breaks the conjecture listed in the concurrent work (Duan et al., 2021).
- Instead of the $O(1/\sqrt{n})$ slow rate generalization bounds in (Lu et al., 2021b; Duan et al., 2021; Jiao et al., 2021a;c), we utilize the strongly convex nature of the variational form and provide a fast rate generalization bound via the localization methods (Van De Geer, 1987; Bartlett et al., 2005; Koltchinskii, 2011; Srebro et al., 2010; Xu & Zeevi, 2020). We showed that PINN and a modified version of DRM can achieve nearly min-max optimal convergence rate. Our result is listed in Table 1.
- We tested the recently discovered neural scaling law (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2019; Hashimoto, 2021) for deep PDE solvers numerically and the empirical results verified our theory.

2 Set-up

We consider the static Schrödinger equation with zero Dirichlet boundary conditions on the domain Ω, which we assume to be the unit hypercube in $\mathbb{R}^d$. In order to precisely introduce the problem, we recall some standard notions. We consider our domain as $\Omega = [0,1]^d$ and use $L^2(\Omega)$ to denote the space of square integrable functions on Ω with respect to the Lebesgue measure. We let $L^\infty(\Omega)$ be the space of essentially bounded (with respect to the Lebesgue measure) functions on Ω and $C(\partial\Omega)$ denotes the space of continuous functions on $\partial\Omega$.

Let $f \in L^2(\Omega)$, $V \in L^\infty(\Omega)$, and , $g \in L^\infty(\Omega)$. Our focus is on the analysis of Deep-Learning-based numerical methods to solve the elliptic equations

$$\begin{aligned} -\Delta u + Vu &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

2.1 Loss Functions for Solving PDEs and Induced Evaluation Metric

In this paper, we mainly focus on analyzing Deep Ritz Methods (DRM) and Physics Informed Neural Network (PINN). In this subsection, we first introduce the objective function and algorithm of the two methods.

Deep Ritz Methods (E & Yu, 2018; Sirignano & Spiliopoulos, 2018) Recall that the equation 2.1 is equivalent to following variational form

$$u^* = \arg\min_{H_0^1(\Omega)} \boldsymbol{E}^{\text{DRM}}(u) := \frac{1}{2}\int_\Omega |\nabla u|^2 + V|u|^2\, dx - \int_\Omega f u dx, \tag{2.2}$$

where u is minimized over $H_0^1(\Omega)$ with boundary condition given by g on $\partial\Omega$.

This variational form provides the basis for the DRM type method for solving the static Schrödinger equation based on neural network ansatz. More specifically, the energy functional given in equation 2.2 is viewed as the population risk function to train an optimal estimator approximation of the solution to the PDE within a parameterized hypothesis function class $\boldsymbol{F} \subset H^1(\Omega)$. In this paper, we also rely on the strong convexity of the DRM objective respect to the H^1 norm.

Proposition 2.1. *We further assume $0 < V_{\min} \le V(x) \le V_{\max}$, then we have*

$$\frac{2}{\max\{1, V_{\max}\}}\left(\boldsymbol{E}^{DRM}(u) - \boldsymbol{E}^{DRM}(u^*)\right) \le \|u - u^*\|_{H^1}^2 \le \frac{2}{\max\{1, V_{\min}\}}\left(\boldsymbol{E}^{DRM}(u) - \boldsymbol{E}^{DRM}(u^*)\right)$$

holds for all $u \in H_0^1(\Omega)$

Physics Informed Neural Network (Raissi et al., 2019; Sirignano & Spiliopoulos, 2018). PINN solves 2.1 via minizing the following objective function

$$u^* = \arg\min_{H_0^1(\Omega)} \boldsymbol{E}^{\text{PINN}}(u) := \int_\Omega |\Delta u(x) - V(x)u(x) + f(x)|^2 dx.$$

The objective function $\boldsymbol{E}^{\text{PINN}}$ can also be viewed as the population risk function and we can train an optimal estimator approximation of the solution to the PDE within a parameterized hypothesis function class $\boldsymbol{F} \subset H_0^1(\Omega)$. In this paper, we also rely on the strong convexity of the PINN objective respect to the H^2 norm.

Proposition 2.2. *For PINN, we further assume $V \in L^\infty(\Omega)$ with $0 < C_{\min} < V^2 - \Delta V, 0 < C_{\min} < V(x) \le V_{\max}$ and $-\Delta V(x) \le V_{\max}$, then we have*

$$\frac{1}{2\left(1 + V_{\max} + V_{\max}^2\right)}\left(\boldsymbol{E}^{PINN}(u) - \boldsymbol{E}^{PINN}(u^*)\right) \le \|u - u^*\|_{H^2}^2 \le \frac{2}{\max\{1, C_{\min}\}}\left(\boldsymbol{E}^{PINN}(u) - \boldsymbol{E}^{PINN}(u^*)\right)$$

holds for all $u \in H_0^1(\Omega)$

2.2 Estimator Setting

Empirical Loss Minimization In order to access the d-dimensional integrals, DRM (E & Yu, 2018; Khoo et al., 2017) and PINN(Raissi et al., 2019; Sirignano & Spiliopoulos, 2018) employ a Monte-Carlo method for computing the high dimensional integrals, which leads to the so-called *empirical risk minimization* training for neural networks. To define the empirical loss, let $\{X_j\}_{j=1}^n$ be an i.i.d. sequence of random variables distributed according to the uniform distribution in domain $\boldsymbol{P}_\Omega$. We also have access to $f_j = f(X_j) + \xi_i, j = 1, \cdots, n$ is the noisy observation of the right hand side of the PDE (2.1) and ξ is a bounded random variable with mean zero and independent with X_j. Define the empirical losses $\boldsymbol{E}_{n,S}$ by setting

$$\boldsymbol{E}_n^{\text{DRM}}(u) = \frac{1}{n}\sum_{j=1}^n \left[|\Omega| \cdot \left(\frac{1}{2}|\nabla u(X_j)|^2 + \frac{1}{2}V(X_j)|u(X_j)|^2 - f_j u(X_j)\right)\right], \tag{2.3}$$

$$\boldsymbol{E}_n^{\text{PINN}}(u) = \frac{1}{n}\sum_{j=1}^n \left[|\Omega| \cdot \left(\Delta u(X_j) - V(X_j)u(X_j) + f_j\right)^2\right], \tag{2.4}$$

where $|\Omega|$ represent the Lebesgue measure of the sets.

Once given an empirical loss $\boldsymbol{E}'_n$, we apply the empirical loss minimization to seek the estimation u_n, i.e. $u_n = \arg\min_{u \in \boldsymbol{F}} \boldsymbol{E}_n(u)$ where $\boldsymbol{F}$ is the parametrized hypothesis function space we consider. For example, reproducing kernel Hilbert space(Chen et al., 2021) and tensor training format(Richter et al., 2021). In this paper, we consider sparse neural network and truncated fourier basis, which can achieves min-max optimal estimation rate for the non-parametric function estimation(Tsybakov, 2008; Schmidt-Hieber et al., 2020; Farrell et al., 2021; Suzuki, 2018; Chen et al., 2019b; Jiao et al., 2021c; Nitanda & Suzuki, 2020).

Neural Network Function Space In this paper, the hypothesis function space $\mathcal{F}$ is expressed by the neural network following (Suzuki, 2018). Let us denote the ReLU^3 activation by $\eta_3(x) = \max\{x^3, 0\}$ $(x \in \mathbb{R})$ which is used in (E & Yu, 2018), and for a vector x, $\eta(x)$ is operated in an element-wise manner. Define the neural network with height L, width W, sparsity constraint S and norm constraint B as

$$\begin{aligned}
&\Phi(L, W, S, B) := \{(\mathcal{W}^{(L)}\eta_3(\cdot) + b^{(L)}) \circ \cdots (\mathcal{W}^{(2)}\eta_3(\cdot) + b^{(2)}) \circ (\mathcal{W}^{(1)}x + b^{(1)}) \mid \\
&\mathcal{W}^{(L)} \in \mathbb{R}^{1\times W}, b^{(L)} \in \mathbb{R}, \mathcal{W}^{(1)} \in \mathbb{R}^{W\times d}, b^{(1)} \in \mathbb{R}^{W}, \mathcal{W}^{(l)} \in \mathbb{R}^{W\times W}, b^{(l)} \in \mathbb{R}^{W} (1 < l < L), \\
&\sum_{l=1}^{L} (\|\mathcal{W}^{(l)}\|_0 + \|b^{(l)}\|_0) \leq S, \max_l \|\mathcal{W}^{(l)}\|_{\infty,\infty} \vee \|b^{(l)}\|_\infty \leq B\},
\end{aligned} \tag{2.5}$$

where $\circ$ denotes the function composition, $\|\cdot\|_0$ is the ℓ_0-norm of the matrix (the number of non-zero elements of the matrix) and $\|\cdot\|_{\infty,\infty}$ is the ℓ_∞-norm of the matrix (maximum of the absolute values of the elements).

Truncated Fourier Basis Estimator We also considered the Truncated Fourier basis as our estimator. Suppose the domain we interested $\Omega \subseteq [0,1]^d$. For any $z \in \mathbb{N}^d$, we consider the corresponding Fourier basis function $\phi_z(x) := e^{2\pi i \langle z, x\rangle}$ $(x \in \Omega)$. Any function $f \in L^2(\Omega)$ can be represented as weighted sum of the Fourier basis $f(x) := \sum_{z\in\mathbb{N}^d} f_z \phi_z(x)$ where $f_z := \int_\Omega f(x)\overline{\phi_z(x)}dx$ $(\forall\, z \in \mathbb{N}^d)$ is the Fourier coefficient. This inspired us to use the Fourier Basis whose index lies in a truncated set $Z_\xi = \{z \in \mathcal{Z} | \|z\|_\infty \leq \xi\}$ to represent the function class $\boldsymbol{F}$ as $\boldsymbol{F}_\xi = \{\sum_{\|z\|_\infty \leq \xi} a_z \phi_z | a_z \in \mathbb{R}, \|z\|_\infty \leq \xi\}$.

3 Lower Bound

In this section, we aim to consider the statistical limit of learning the solution of a PDE. As discussed in Propositions 2.1 and 2.2, we directly consider the H^1 norm for DRM and H^2 norm for PINN as the evaluation metric. The lower bound shown as follows.

Theorem 3.1 (Lower bound)**.** *We denote $u^*(f)$ to be the solution of the PDE 2.1 and we can access randomly sampled data $\{X_i, f_i\}_{i=1,\cdots,n}$ as described in Section 2.2. We further assume $u^*(f) \in H^\alpha$ for a given $\alpha \in \mathbb{Z}^+$.*

DRM Lower Bound. *For all estimator $H : \left(\mathbb{R}^d\right)^{\otimes n} \times \mathbb{R}^{\otimes n} \to H_\alpha(\Omega)$, we have*

$$\inf_H \sup_{u \in H_\alpha(\Omega)} \mathbb{E}\|H(\{X_i, f_i\}_{i=1,\cdots,n}) - u^*(f)\|_{H^1}^2 \gtrsim n^{-\frac{2\alpha-2}{d+2\alpha-4}}. \tag{3.1}$$

PINN Lower Bound. *For all estimator $H : \left(\mathbb{R}^d\right)^{\otimes n} \times \mathbb{R}^{\otimes n} \to H^1(\Omega)$, we have*

$$\inf_H \sup_{u \in H_\alpha(\Omega)} \mathbb{E}\|H(\{X_i, f_i\}_{i=1,\cdots,n}) - u^*(f)\|_{H^2}^2 \gtrsim n^{-\frac{2\alpha-4}{d+2\alpha-4}}. \tag{3.2}$$

Given that $n^{-\frac{2(\beta-k)}{d+2\beta}}$ is the minimax rate of estimation of the k-th derivative of a β-smooth density in L_2 (Liu & Wang, 2012; Prakasa Rao, 1996; Müller & Gasser, 1979), the lower bound obtained here is the rate of estimating the right hand side function f in terms of the H^{-1} norm. Given the H^{-1} norm error estimate on f, we can achieve estimate of u, which provides an alternative way to understand our upper bound. (See discussion in Appendix D.)

4 UPPER BOUND

To theoretically understand the empirical success of Physics Informed Neural Networks and the Deep Ritz solver, in this section, we aim to prove that the excess risk $\Delta \boldsymbol{E}_n := \boldsymbol{E}(u_n) - \boldsymbol{E}(u^*)$ of a well-trained neural networks on the PINN/DRM loss function will follow a precise power-law scaling relations with the size of the training dataset. Similar to (Xu, 2020; Lu et al., 2021b; Duan et al., 2021; Jiao et al., 2021a;b), we decompose the excess risk into approximation error and generalization error. Different from the concurrent bound (Duan et al., 2021; Jiao et al., 2021a), we provided a fast rate $O(1/n)$ by utilizing the strong convexity of the objective function established in Section 2.1 and achieved a faster and near optimal upper bound. We showed that the generalization error can be bounded by the fixed point of the local Rademacher complexity

$$\psi(r) = R_n(\{\mathcal{I}(u) | \|u - u^*\|_A^2 \leq r\}),$$

where R_n is the Rademacher complexity, $\mathcal{I}(u) = \Delta u + Vu, \|\cdot\|_A = \|\cdot\|_{H^2}$ for PINN and $\mathcal{I}(u) = \|\nabla u\|^2 + Vu, \|\cdot\|_A = \|\cdot\|_{H^1}$ for DRM. We put detailed definition and analysis in Appendix A.5. Then we plug in the approximation and generalization error calculated in Appendix A.4 and Appendix A.3 and finally achieved the following upper bounds.

Physics Informed Neural Network.

Theorem 4.1. *(Informal Upper Bound of PINN with Deep Neural Network Estimator) With proper assumptions, consider the sparse Deep Neural Network function space $\Phi(L, W, S, B)$ with parameters $L = O(1),\ W = O(n^{\frac{d}{d+2s-4}}),\ S = O(n^{\frac{d}{d+2s-4}}),\ B = O(1)$, then the Physics Informed estimator $\hat{u}_{PINN}^{DNN} = \min_{u \in \Phi(L,W,S,B)} \boldsymbol{E}_n^{PINN}(u)$ satisfies the following upper bound with high probability:*

$$\|\hat{u}_{PINN}^{DNN} - u^*\|_{H^2}^2 \lesssim n^{-\frac{2s-4}{d+2s-4}} \log n.$$

Theorem 4.2. *(Informal Upper Bound of PINN with Truncated Fourier Series Estimator) With proper assumptions, consider the Physics Informed Neural Network objective with a plug-in Fourier Series estimator $\hat{u}_{PINN}^{Fourier} = \min_{u \in \boldsymbol{F}_\xi(\Omega)} \boldsymbol{E}_n^{PINN}(u)$ with $\xi = \Theta(n^{\frac{1}{d+2s-4}})$, then with high probability we have*

$$\|\hat{u}_{PINN}^{Fourier} - u^*\|_{H^2}^2 \lesssim n^{-\frac{2s-4}{d+2s-4}}.$$

Deep Ritz Methods.

Theorem 4.3. *(Final Upper Bound of DRM with Deep Neural Network Estimator) With proper assumptions, consider the sparse Deep Neural Network function space $\Phi(L, W, S, B)$ with parameters $L = O(1),\ W = O(n^{\frac{d}{d+2s-2}}),\ S = O(n^{\frac{d}{d+2s-2}}),\ B = O(1)$, then the Deep ritz estimator $\hat{u}_{DRM}^{DNN} = \min_{u \in \Phi(L,W,S,B)} \boldsymbol{E}_n^{DRM}(u)$ satisfies the following upper bound with high probability:*

$$\|\hat{u}_{DRM}^{DNN} - u^*\|_{H^1}^2 \lesssim n^{-\frac{2s-2}{d+2s-2}} \log n.$$

Theorem 4.4. *(Informal Upper Bound of DRM with Truncated Fourier Series Estimator)With proper assumptions, consider the Deep Ritz objective with a plug in Fourier Series estimator $\hat{u}_{DRM}^{Fourier} = \min_{u \in \boldsymbol{F}_\xi(\Omega)} \boldsymbol{E}_n^{DRM}(u)$ with $\xi = \Theta(n^{\frac{1}{d+2s-2}})$, then with high probability we have*

$$\|\hat{u}_{DRM}^{Fourier} - u^*\|_{H^1}^2 \lesssim n^{-\frac{2s-2}{d+2s-2}}.$$

Remark.

- There is a common belief that Machine learning based PDE solvers can break the curse of dimensionality (E & Yu, 2018; Grohs et al., 2018; Lanthaler et al., 2021). However we obtained an $n^{-\frac{2s-2}{2s-4+d}}$ convergence rate which can become super slow in high dimension. Our analysis showed that it is essential to constrain the function space to break the curse of dimensionality. (Lu et al., 2021b) considered the DRM in Barron spaces. (Ongie et al., 2019) showed that functions in the Barron space enjoy a smoothness s at the same magnitude as d , which will also leads to convergence rate independent of the dimension using our upper bound. Neural network can also approximate mixed sparse grid spaces (Montanelli & Du, 2019; Suzuki, 2018), function on manifold (Nitanda & Suzuki, 2020; Chen et al., 2019b) without curse of dimensionality. Combined with these approximation bounds, we can also achieve a bound that breaks the curse of dimensionality using Theorem A.12 and A.9. In this paper, we aim to consider the statistical power of the loss function in common function spaces and put the curse of dimensionality as a separate topic.

- Our bound is faster than the concurrent bound (Duan et al., 2021; Jiao et al., 2021a) for we provided a fast rate $O(1/n)$ by utilizing the strong convexity of the objective function and improves the convergence rate from $n^{-\frac{2s-2}{d+4s-4}}$ to $n^{-\frac{2s-2}{d+2s-2}}$ for Deep Ritz and from $n^{-\frac{2s-4}{d+4s-8}}$ to $n^{-\frac{2s-4}{d+2s-4}}$ for PINN. Comparing to the lower bound provided in Section3, we show that our bound for PINN is near optimal and we'll let our bound for DRM become near optimal in the next section.
- For upper bound of DRM, due to a technical issue, we assume the observation we access is clean, *i.e* $f_i = f(X_i)$. We conjecture that add noising on observation will not effect the rate and leave this to future work.
- (Singer & Wu, 2017; Calder & Trillos, 2019; Trillos et al., 2020; Hoffmann et al., 2021; Wormell & Reich, 2021; Cheng & Wu, 2021) considered learning the Ritz form using random samples via graph Laplacian. All their upper bound is slower than the lower bound showed in Section 3. One reason is all their upper bounds are slow convergence rate without awareness of the strongly convex geometry of the loss function.

5 Modified Deep Ritz Methods

Comparing the lower bound in Section 3 and the upper bound in Section 4, we find out that the Physics Informed Neural Network achieved min-max optimality while the Deep Ritz Method does not. In this section, we proposed a modified version of deep Ritz which can be statistically optimal.

As discussed in Appendix D, the reason behind the suboptimality of DRM comes from the high complexity introduced via the uniform concentration bound of the gradient term in the variational form. At the same time, we further observed that the $\int \|\nabla u\|^2 dx$ does not require any query from the right hand side function f, which means that we can easily make another splitted sample to approximate the $\int \|\nabla u\|^2 dx$ term more precisely.

$$\boldsymbol{E}_{N,n}(u) = \frac{1}{N}\sum_{j=1}^{N}\left[|\Omega|\cdot\frac{1}{2}\|\nabla u(X_j')\|^2\right] + \frac{1}{n}\sum_{j=1}^{n}\left[|\Omega|\cdot\left(\frac{1}{2}V(X_j)|u(X_j)|^2 - f_j u(X_j)\right)\right] \tag{5.1}$$

Once we sampled more data for approximating $\int |\nabla u|^2 dx$, we can achieve an near optimal bound for the Truncated Fourier Estimator when $\frac{N}{n} \gtrsim n^{\frac{2}{d+2s-4}}$.

Theorem 5.1. *(Informal Upper Bound of DRM with Truncated Fourier Series Estimator)With proper assumptions, consider the Deep Ritz objective with a plug in Fourier Series estimator* $\hat{u}_{DRM}^{Fourier} = \min_{u\in \boldsymbol{F}_\xi(\Omega)} \boldsymbol{E}_n^{DRM}(u)$ *with* $\xi = \Theta(n^{\frac{1}{d+2s-4}})$ *and* $\frac{N}{n} \gtrsim n^{\frac{2}{d+2s-4}}$*, then we have*

$$\|\hat{u}_{DRM}^{Fourier} - u^*\|_{H^1}^2 \lesssim n^{-\frac{2s-2}{d+2s-4}}.$$

Remark. We still cannot achieve optimal rate for neural network even with modified DRM methods. The reason is because the number of neuron is not a good complexity measure for the gradient of the function thus the bound for $\psi(r) = R_n(\{\mathcal{I}(u) | \|u-u^*\|_{H^1}^2 \le r\})$ is not enough for achieving optimal convergence rate. However, following (Schmidt-Hieber et al., 2020; Suzuki, 2018; Imaizumi & Fukumizu, 2020; Chen et al., 2019b; Farrell et al., 2021) using deep networks for estimating functions, we optimize the best neural network with constrained sparsity in our paper. Here we conjecture that there exists a computable complexity measure can makes DRM statistical optimal and put finding the right complexity of the neural network's gradient to be future work.

6 Experiments

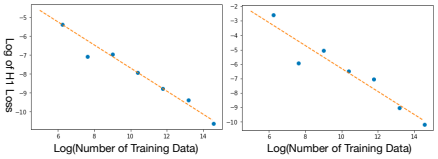
In this section, we conduct several numerical experiments to verify our theory. We follow the neural network and hyper-parameter setting in (Chen et al., 2020). Due to the page limit, we only put the experiments for Deep Ritz Methods here.

6.1 THE MODIFIED DEEP RITZ METHODS

In this section, we conduct experiments which substantiate our theoretical results for modified Deep Ritz methods. For simplicity, we take $V(x) = 1$ in our experiment. We conduct experiment in 2-dimension and select the solution of the PDE as $u^* = \sum_z \|z\|^{-s}\phi_z(x) \in H^s$. We showed the Log-log plot of H^1 loss plotted against the number of sampled data for $s = 4$ in Figure 1. We use an OLS estimator to fit the log-log plot and put the estimated slope and corresponding R2 score in Figure 1. As our theory predicts, the modified Deep Ritz Methods convergences faster than the original one. All the derivation of the two estimators is listed in Appendix D.

6.2 DIMENSION DEPENDENT SCALING LAW.

We conduct experiments to illustrate that the population loss of well-trained and well-tuned deep Ritz method will scales with the d-dimensional training data number N as a power-law $\mathcal{L} \propto \frac{1}{N^\alpha}$. We also scan over a range of d and α and verified an an approximately $\alpha \propto \frac{1}{d}$ scaling law as our theory suggested. We use the same test function in Section 6.1 as the solution of our PDE. For simplicity, we take $V(x) = 1$ in our experiment. We train the deep Ritz method on 20, 80, 320, 1280, 10240 sampled data for 5,6,7,8,9,10 dimensional problems and we plot our result in log–log scale. Result is shown in Figure 2. We discovered the $L \propto n^{\frac{1}{d+2}}$ scaling law in practical situations.



	(a) Deep Ritz Methods	(b) Modified Deep Ritz Methods
Theory	$\frac{2s-2}{d+2s-2} = 0.75$	$\frac{2s-2}{d+2s-4} = 1$
Empirical	0.6097	0.7953
R2 Score	0.91	0.89

Figure 1: The Log-Log plot and estimated convergence slope for Modified DRM and DRM using fourier basis, showing the median error over 5 replicates.

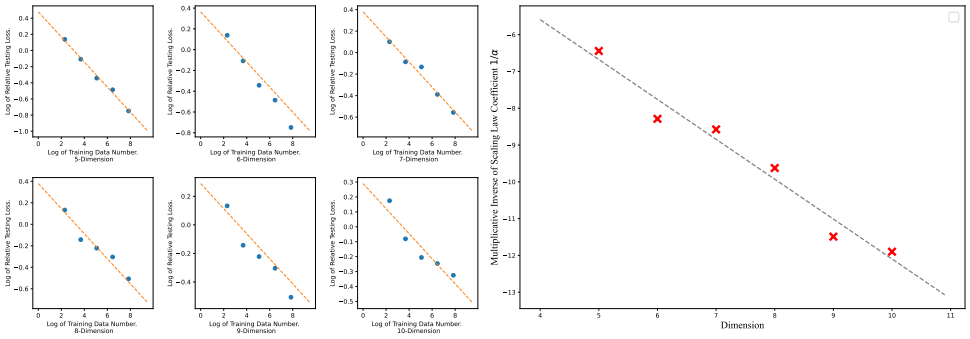


Figure 2: We verify the dimension dependent scaling law empirically and the multiplicative inverse of the scaling law coefficient is highly linear with the dimension d, showing the mean error over 2 replicates.

6.3 ADAPATION TO THE SIMPLER FUNCTIONS.

(Sharma & Kaplan, 2020) have shown that the neural scaling law will adapt to the structure that the target function enjoys, this adaptivity enables the neural network to break the cure of the dimensionality for simple functions in high dimension. (Suzuki & Nitanda, 2019; Chen et al., 2019a) also observed this theortically. For solving PDE, we also observed this adaptivity in practice. Here we tested the following two hypothesis

- **Random Neural Network Teacher.** Following (Sharma & Kaplan, 2020), we also tested random neural network using He initialization (He et al., 2015) as the ground turth solution u^*. (De Palma et al., 2018) has shown that random deep neural networks are biased towards simple functions and in practice we observed a scaling law at the parametric rate. To be specific, we obtained a linear estimate with slope $\alpha = -0.50679429$ in the log-log curve and with a R^2 score 0.96. See Figure3(a).

- **Simple Polynomials.** Neural network can approximate simple polynomials exponentially fast (Wang et al., 2018), thus we select the ground truth solution to be the following simple polynomial in 10 dimensional spaces,

$$u^*(x) = x_1x_2 + \cdots + x_9x_{10}.$$

In this example, we obtained a linear estimate with slope $\alpha = -0.49755418$ in the log-log curve and with a R^2 score 0.99. See Figure3(b). .

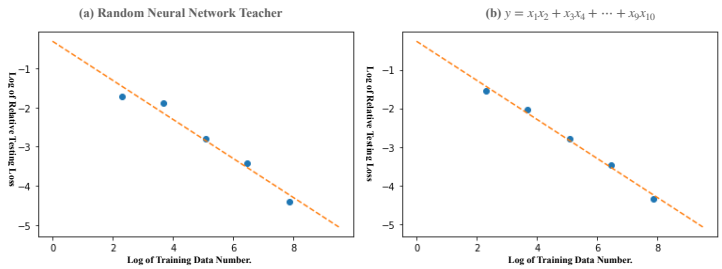


Figure 3: Neural network have the ability to adapt to simple functions and achieves convergence without curse of dimensionality, showing the median error over 5 replicates.

7 Conclusion and Discussion

Conclusion In this paper, We consider the statistical min-max optimality of solving a PDE from random samples. We improve the previous bounds (Xu, 2020; Lu et al., 2021b; Duan et al., 2021; Jiao et al., 2021a) via providing the first fast rate generalization bound for learning PDE solution via the strongly convex nature of the objective function used and achieved the optimal rate via the PINN and a modified deep ritz methods. We verified our theory via numerical experiments and explored the dimensional-dependent scaling laws of Deep PDE solvers.

Discussion and Future Work Here we discuss several drawbacks of our theory

- We restricted our target function and estimators in the W^1_∞ instead of the H^1 due to boundness assumption made in the local Rademacher complexity arguments. However typical functional used in physics is always unbounded, as an example the Newtonian potential $\frac{1}{\|x-y\|^{d-2}}$, which limits the application of our theory.
- This paper did not discuss any optimization aspect of the deep PDE solvers and always assumes achieves global optimum. However, it is important to investigate will the optimization error (Suzuki & Akiyama, 2020; Chizat, 2021) finally dominate.
- Instead of solving a single PDE, recent work(Long et al., 2018; 2019; Li et al., 2020; Lanthaler et al., 2021; Bhattacharya et al., 2020; Fan & Ying, 2020; Feliu-Faba et al., 2020) consider so called "operator learning" which aims to learn a family of PDE/inverse problem using a single network. It is interesting to investigate the generalization bound and neural scaling law there.
- We find out that the sparsity of the neural network is not a good complexity measure of neural network's gradient. We conjecture there exist an oracle complexity measure whose approximation and generalization bounds can lead Modified DRM achieves the optimal convergence rate.

References

Genming Bai, Ujjwal Koley, Siddhartha Mishra, and Roberto Molinaro. Physics informed neural networks (pinns) for approximating nonlinear dispersive pdes. *arXiv preprint arXiv:2104.05584*, 2021.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B Kovachki, and Andrew M Stuart. Model reduction and neural networks for parametric pdes. *arXiv preprint arXiv:2005.03180*, 2020.

Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

Lawrence D Brown and Mark G Low. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics*, 24(6):2384–2398, 1996.

Jeff Calder and Nicolas Garcia Trillos. Improved spectral convergence rates for graph laplacians on epsilon-graphs and k-nn graphs. *arXiv preprint arXiv:1910.13476*, 2019.

Jingrun Chen, Rui Du, and Keke Wu. A comprehensive study of boundary conditions when solving pdes by dnns. *arXiv preprint arXiv:2005.04554*, 2020.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*, 32:8174–8184, 2019a.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep relu networks. *arXiv preprint arXiv:1908.01842*, 2019b.

Yifan Chen, Bamdad Hosseini, Houman Owhadi, and Andrew M Stuart. Solving and learning nonlinear pdes with gaussian processes. *arXiv preprint arXiv:2103.12959*, 2021.

Xiuyuan Cheng and Nan Wu. Eigen-convergence of gaussian kernelized graph laplacian by manifold heat interpolation. *arXiv preprint arXiv:2101.09875*, 2021.

Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *arXiv preprint arXiv:2105.08368*, 2021.

Carl De Boor and Carl De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978.

Giacomo De Palma, Bobak Toussi Kiani, and Seth Lloyd. Random deep neural networks are biased towards simple functions. *arXiv preprint arXiv:1812.10156*, 2018.

Chenguang Duan, Yuling Jiao, Yanming Lai, Xiliang Lu, and Zhijian Yang. Convergence rate analysis for deep ritz method. *arXiv preprint arXiv:2103.13330*, 2021.

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

Lawrence C Evans. Partial differential equations. *Graduate studies in mathematics*, 19(4):7, 1998.

Yuwei Fan and Lexing Ying. Solving electrical impedance tomography with deep learning. *Journal of Computational Physics*, 404:109119, 2020.

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

Jordi Feliu-Faba, Yuwei Fan, and Lexing Ying. Meta-learning pseudo-differential operators with deep neural networks. *Journal of Computational Physics*, 408:109309, 2020.

Sara A Geer and Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Maximilien Germain, Huyên Pham, and Xavier Warin. Neural networks-based algorithms for stochastic control and pdes in finance. *arXiv preprint arXiv:2101.08068*, 2021.

Davis Gilton, Greg Ongie, and Rebecca Willett. Neumann networks for linear inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 6:328–343, 2019.

Philipp Grohs and Lukas Herrmann. Deep neural network approximation for high-dimensional elliptic pdes with boundary conditions. *arXiv preprint arXiv:2007.05384*, 2020.

Philipp Grohs, Fabian Hornung, Arnulf Jentzen, and Philippe Von Wurstemberger. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations. *arXiv preprint arXiv:1809.02362*, 2018.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep relu neural networks in w s, p norms. *Analysis and Applications*, 18(05):803–859, 2020.

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

Jiequn Han, Jianfeng Lu, and Mo Zhou. Solving high-dimensional eigenvalue problems using deep neural networks: A diffusion monte carlo like approach. *Journal of Computational Physics*, 423: 109792, 2020.

Tatsunori Hashimoto. Predicting the impact of dataset composition on model performance, 2021. URL https://openreview.net/forum?id=butEPeLARP_.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

James B Heaton, Nick G Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Franca Hoffmann, Bamdad Hosseini, Assad A Oberai, and Andrew M Stuart. Spectral analysis of weighted laplacians arising in data clustering. *Applied and Computational Harmonic Analysis*, 2021.

Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

Masaaki Imaizumi and Kenji Fukumizu. Advantage of deep neural networks for estimating functions with singularity on curves. *arXiv preprint arXiv:2011.02256*, 2020.

Yuling Jiao, Yanming Lai, Dingwei Li, Xiliang Lu, Yang Wang, and Jerry Zhijian Yang. Convergence analysis for the pinns, 2021a.

Yuling Jiao, Yanming Lai, Yisu Luo, Yang Wang, and Yunfei Yang. Error analysis of deep ritz methods for elliptic equations. *arXiv preprint arXiv:2107.14478*, 2021b.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Yuehaw Khoo, Jianfeng Lu, and Lexing Ying. Solving parametric pde problems with artificial neural networks. *arXiv preprint arXiv:1707.03351*, 2017.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer Science & Business Media, 2011.

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

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Youming Liu and Huiying Wang. Convergence order of wavelet thresholding estimator for differential operators on besov spaces. *Applied and Computational Harmonic Analysis*, 32(3):342–356, 2012.

Zichao Long, Yiping Lu, Xianzhong Ma, and Bin Dong. Pde-net: Learning pdes from data. In *International Conference on Machine Learning*, pp. 3208–3216. PMLR, 2018.

Zichao Long, Yiping Lu, and Bin Dong. Pde-net 2.0: Learning pdes from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019.

Denghui Lu, Han Wang, Mohan Chen, Lin Lin, Roberto Car, E Weinan, Weile Jia, and Linfeng Zhang. 86 pflops deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy. *Computer Physics Communications*, 259:107624, 2021a.

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

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

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps. *arXiv preprint arXiv:2107.12364*, 2021.

Tanya Marwah, Zachary C Lipton, and Andrej Risteski. Parametric complexity bounds for approximating pdes with neural networks. *arXiv preprint arXiv:2103.02138*, 2021.

Hiroaki Mikami, Kenji Fukumizu, Shogo Murai, Shuji Suzuki, Yuta Kikuchi, Taiji Suzuki, Shin-ichi Maeda, and Kohei Hayashi. A scaling law for synthetic-to-real transfer: A measure of pre-training. *arXiv preprint arXiv:2108.11018*, 2021.

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

Hans-Georg Müller and Theo Gasser. Optimal convergence properties of kernel estimates of derivatives of a density function. In *Smoothing techniques for curve estimation*, pp. 144–154. Springer, 1979.

Richard Nickl, Sara van de Geer, and Sven Wang. Convergence rates for penalized least squares estimators in pde constrained regression problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):374–413, 2020.

Atsushi Nitanda and Taiji Suzuki. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. *arXiv preprint arXiv:2006.12297*, 2020.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

BLS Prakasa Rao. Nonparametric estimation of the derivatives of a density by the method of wavelets. *Bulletin of Informatics and Cybernetics*, 28(1):91–100, 1996.

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

Alexander Rakhlin, Karthik Sridharan, and Alexandre B Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

Lorenz Richter, Leon Sallandt, and Nikolas Nüsken. Solving high-dimensional parabolic pdes using the tensor train format. *arXiv preprint arXiv:2102.11830*, 2021.

Jonathan S Rosenfeld. Scaling laws for deep learning. *arXiv preprint arXiv:2108.07686*, 2021.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

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

Larry Schumaker. *Spline functions: basic theory*. Cambridge University Press, 2007.

Utkarsh Sharma and Jared Kaplan. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Neural network approximation: Three hidden layers are enough. *Neural Networks*, 2021.

Yeonjong Shin, Zhongqiang Zhang, and George Em Karniadakis. Error estimates of residual minimization using neural networks for linear pdes. *arXiv preprint arXiv:2010.08019*, 2020.

Amit Singer and Hau-Tieng Wu. Spectral convergence of the connection laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 2017.

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

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in neural information processing systems*, 23, 2010.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Taiji Suzuki and Shunta Akiyama. Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. *arXiv preprint arXiv:2012.03224*, 2020.

Taiji Suzuki and Atsushi Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *arXiv preprint arXiv:1910.12799*, 2019.

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph laplacian on random geometric graphs toward the laplace–beltrami operator. *Foundations of Computational Mathematics*, 20(4):827–887, 2020.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

Sara Van De Geer. A new approach to least-squares estimation, with applications. *The Annals of Statistics*, 15(2):587–602, 1987.

Qingcan Wang et al. Exponential convergence of the deep neural network approximation for analytic functions. *arXiv preprint arXiv:1807.00297*, 2018.

Stephan Wojtowytsch et al. Some observations on partial differential equations in barron and multi-layer spaces. *arXiv preprint arXiv:2012.01484*, 2020.

Caroline L Wormell and Sebastian Reich. Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization. *SIAM Journal on Numerical Analysis*, 59(3):1687–1734, 2021.

Jinchao Xu. The finite neuron method and convergence analysis. *arXiv preprint arXiv:2010.01458*, 2020.

Yunbei Xu and Assaf Zeevi. Towards optimal problem dependent generalization error bounds in statistical learning theory. *arXiv preprint arXiv:2011.06186*, 2020.

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

Yaohua Zang, Gang Bao, Xiaojing Ye, and Haomin Zhou. Weak adversarial networks for high-dimensional partial differential equations. *Journal of Computational Physics*, 411:109409, 2020.

Xiaoshuai Zhang, Yiping Lu, Jiaying Liu, and Bin Dong. Dynamically unfolding recurrent restorer: A moving endpoint control method for image restoration. *arXiv preprint arXiv:1805.07709*, 2018.

Appendix

A Proof of the Upper Bounds

A.1 Notations

In this section, we provide all the notations we need in the proof. Let $\Omega \subset \mathbb{R}^d$ be some open set. We denote $C(\Omega)$ the space of continuous functions on Ω and $C^k(\Omega)$ the space of all functions that are k times continuously differentiable on Ω ($\forall k \in \mathbb{Z}^+$). For any $n \in \mathbb{N}_0$ ($\mathbb{N}_0 := \mathbb{Z}^+ \cup \{0\}$ is the set of all non-negative integers) and $1 \leq p \leq \infty$, we define the Sobolev space $W^{n,p}(\Omega)$ by

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

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

$$\|f\|_{W^{n,p}(\Omega)} := \Big(\sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|^p_{L^p(\Omega)} \Big)^{\frac{1}{p}}.$$

In particular, when $p = \infty$, we have:

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

Consider the Fourier expansion $f := \sum_{z \in \mathbb{N}^d} f_z \phi_z(x)$ of the function $f \in W^{n,p}(\Omega)$. We can equivalently express the Sobolev norm as:

$$\|f\|_{W^{n,p}(\Omega)} = \Big(\sum_z \|z\|^{np} |f_z|^p \Big)^{1/p},$$

where $f_z = \int_\Omega f(x) \overline{\phi_z(x)} dx = \int_\Omega f(x) e^{-2\pi i \langle z, x \rangle} dx$ ($x \in \Omega$) is the $z-$th Fourier coefficient of f. Moreover, we use $W_0^{1,p}(\Omega)$ to denote the closure of $C_c^1(\Omega)$ in $W^{1,p}(\Omega)$. In particular, when $p = 2$, we define $H_0^1(\Omega) := W_0^{1,2}(\Omega)$.
Furthermore, we use $\|\cdot\|$ to present the vector 2 norm and, given a data sample $\{X_i\}_{i=1}^n \subset \Omega$, $\|\cdot\|_{n,p} = (\mathbb{E}_n \cdot^p)^{1/p}$ denote the empirical p norm, where $\mathbb{E}_n : L^2(\Omega) \to \mathbb{R}$ is the corresponding empirical average operator defined as $\mathbb{E}_n f := \frac{1}{n} \sum_{i=1}^n f(X_i)$, $\forall \ f \in L^2(\Omega)$. Given two quantities X and Y, we write $X \lesssim Y$ when the inequality $X \leq CY$ holds, where C is some constant. For two functions f and g mapping from $\mathbb{R}^+$ to $\mathbb{R}$, we write $f = O(g)$ when there exist two constants C' and x_0 independent of f and g, such that the inequality $f(x) \leq C' g(x)$ holds for any $x \geq x_0$.

A.2 Regularity Result For the PDE model.

Regularity Results of the DRM Objective Function

Theorem A.1. *We consider the static Schrödinger equation on the unit hypercube on $\mathbb{R}^d$ with the zero Direchlet boundary condition:*

$$\begin{aligned} -\Delta u + V u &= f \text{ on } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{A.1}$$

where $f \in L^2(\Omega)$ and $V \in L^\infty(\Omega)$ with $0 < V_{\min} \leq V(x) \leq V_{\max} > 0$. There exists a unique weak solution u_S^* to the equivalent variational problem (Evans, 1998):

$$u_S^* = \arg\min_{u \in H_0^1(\Omega)} \boldsymbol{E}_S^{DRM}(u) := \arg\min_{u \in H_0^1(\Omega)} \Big\{ \frac{1}{2} \int_\Omega \Big[\|\nabla u\|^2 + V|u|^2 \Big] dx - \int_\Omega f u dx \Big\}. \tag{A.2}$$

Then for any $u \in H^1(\Omega)$, we have:

$$\frac{\min(1, V_{\min})}{2} \|u - u_S^*\|^2_{H^1(\Omega)} \leq \boldsymbol{E}_S^{DRM}(u) - \boldsymbol{E}_S^{DRM}(u_S^*) \leq \frac{\max(1, V_{\max})}{2} \|u - u_S^*\|^2_{H^1(\Omega)}. \tag{A.3}$$

Proof. To show that u_S^* satisfies estimate A.3, we first claim that for any $u \in H^1(\Omega)$,

$$\boldsymbol{E}_S(u) - \boldsymbol{E}_S(u_S^*) = \frac{1}{2}\int_\Omega \|\nabla u - \nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V(u_S^* - u)^2\, dx. \tag{A.4}$$

In fact, by plugging in the first equation of A.1, one has that

$$\begin{aligned}
\boldsymbol{E}_S(u_S^*) &= \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V|u_S^*|^2 dx - \int_\Omega f u_S^* dx \\
&= \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V|u_S^*|^2 dx + \int_\Omega (\Delta u_S^* - V u_S^*) u_S^* dx \\
&= \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \int_\Omega (\Delta u_S^*) u_S^* dx - \frac{1}{2}\int_\Omega V|u_S^*|^2 dx.
\end{aligned}$$

Furthermore, applying Green's formula to the true solution u_S^* yields:

$$\begin{aligned}
\boldsymbol{E}_S(u_S^*) &= \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \int_\Omega (\Delta u_S^*) u_S^* dx - \frac{1}{2}\int_\Omega V|u_S^*|^2 dx \\
&= \int_{\partial\Omega} \frac{\partial u_S^*}{\partial n} u_S^* dx - \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx - \frac{1}{2}\int_\Omega V|u_S^*|^2 dx \\
&= -\frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx - \frac{1}{2}\int_\Omega V|u_S^*|^2 dx,
\end{aligned}$$

where the last identity above follows from the second equality in A.1. Now for any $u \in H^1(\Omega)$, applying Green's formula to u and the true solution u_S^* implies:

$$\begin{aligned}
\boldsymbol{E}_S(u) - \boldsymbol{E}_S(u_S^*) &= \frac{1}{2}\int_\Omega \|\nabla u\|^2 dx + \frac{1}{2}\int_\Omega V|u|^2 dx - \int_\Omega f u dx + \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V|u_S^*|^2 dx \\
&= \frac{1}{2}\int_\Omega \|\nabla u\|^2 dx + \frac{1}{2}\int_\Omega V|u|^2 dx + \int_\Omega (\Delta u_S^* - V u_S^*) u dx + \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V|u_S^*|^2 dx \\
&= \frac{1}{2}\int_\Omega \|\nabla u\|^2 dx + \int_\Omega (\Delta u_S^*) u dx + \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V\big(u_S^* - u\big)^2 dx \\
&= \frac{1}{2}\int_\Omega \|\nabla u\|^2 dx + \int_{\partial\Omega} \frac{\partial u_S^*}{\partial n} u dx - \int_\Omega \nabla u_S^* \cdot \nabla u dx + \frac{1}{2}\int_\Omega \|\nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V\big(u_S^* - u\big)^2 dx \\
&= \frac{1}{2}\int_\Omega \|\nabla u - \nabla u_S^*\|^2 dx + \frac{1}{2}\int_\Omega V(u_S^* - u)^2\, dx,
\end{aligned}$$

where the last identity above again follows from the second equality in A.1. This completes our proof of identity A.4. Using the assumptions on the potential function V then implies:

$$\begin{aligned}
\boldsymbol{E}_S(u) - \boldsymbol{E}_S(u_S^*) &\le \frac{\max(1, V_{\max})}{2}\Big[\int_\Omega \|\nabla u - \nabla u_S^*\|^2 dx + \int_\Omega (u_S^* - u)^2\, dx\Big] \\
&= \frac{\max(1, V_{\max})}{2}\|u - u_S^*\|^2_{H^1(\Omega)}, \\
\boldsymbol{E}_S(u) - \boldsymbol{E}_S(u_S^*) &\ge \frac{\max(1, V_{\min})}{2}\Big[\int_\Omega \|\nabla u - \nabla u_S^*\|^2 dx + \int_\Omega (u_S^* - u)^2\, dx\Big] \\
&= \frac{\max(1, V_{\min})}{2}\|u - u_S^*\|^2_{H^1(\Omega)}.
\end{aligned}$$

This completes our proof of A.1. □

Regularity Results of the PINN Objective Function

Theorem A.2. *We consider the static Schrödinger equation on the unit hypercube on $\mathbb{R}^d$ with the Neumann boundary condition:*

$$\begin{aligned}
-\Delta u + V u &= f \text{ on } \Omega, \\
u &= 0 \text{ on } \partial\Omega.
\end{aligned} \tag{A.5}$$

where $f \in L^2(\Omega)$ and $V \in L^\infty(\Omega)$ with $V - \frac{1}{2}\Delta V > C_{\min}, 0 < C_{\min} < V(x) \leq V_{\max}$ and $-\Delta V(x) \leq V_{\max}$. Then there exists a unique solution $u_S^* \in H_0^1(\Omega)$ to the following minimization problem (Brezis, 2010):

$$u_S^* = \underset{u \in H_0^1(\Omega)}{\arg\min}\, \boldsymbol{E}_S^{PINN}(u) := \underset{u \in H_0^1(\Omega)}{\arg\min} \Big\{ \int_\Omega |\Delta u - Vu + f|^2 dx \Big\}. \tag{A.6}$$

Then for any $u \in H^1(\Omega)$, we have:

$$\min\{1, C_{\min}\} \|u - u_S^*\|_{H^2(\Omega)}^2 \leq \boldsymbol{E}(u) - \boldsymbol{E}(u_S^*) \leq 2(1 + V_{\max} + V_{\max}^2) \|u - u_S^*\|_{H^2(\Omega)}^2. \tag{A.7}$$

Proof. Let $\tilde{u} = u - u^*$, we have

$$\begin{aligned} \boldsymbol{E}(u) - \boldsymbol{E}(u_S^*) &= \int_\Omega (\Delta \tilde{u})^2 + V^2 \tilde{u}^2 - 2V\tilde{u}\Delta\tilde{u} dx = \int_\Omega (\Delta\tilde{u})^2 + V^2\tilde{u}^2 + 2V\|\nabla\tilde{u}\|^2 + 2\tilde{u}\nabla V \cdot \nabla\tilde{u} dx \\ &= \int (V^2 - \Delta V)\tilde{u}^2 + 2V\|\nabla\tilde{u}\|^2 + (\Delta\tilde{u})^2 \end{aligned} \tag{A.8}$$

For we have assumed $V \in L^\infty(\Omega)$ with $0 < C_{\min} < V^2 - \Delta V, 0 < C_{\min} < V(x) \leq V_{\max}$ and $-\Delta V(x) \leq V_{\max}$, thus we have

$$\min\{1, C_{\min}\} \|u - u_S^*\|_{H^2(\Omega)}^2 \leq \boldsymbol{E}(u) - \boldsymbol{E}(u_S^*) \leq 2(1 + V_{\max} + V_{\max}^2) \|u - u_S^*\|_{H^2(\Omega)}^2. \tag{A.9}$$

□

A.3 Auxiliary definitions and lemmata On Generalization Error

To bound the generalization error, we use the localized Rademacher complexity (Bartlett et al., 2005). Recall that the Rademacher complexity of a function class $\boldsymbol{G}$ is defined by

$$R_n(\boldsymbol{G}) = \mathbb{E}_Z \mathbb{E}_\sigma \Big[\sup_{g \in \boldsymbol{G}} \Big| \frac{1}{n} \sum_{j=1}^n \sigma_j g(Z_j) \Big| \;\Big|\; Z_1, \cdots, Z_n \Big],$$

where Z_i are i.i.d samples according to the data distributions and σ_j are i.i.d Rademacher random variables which take the value 1 with probability $\frac{1}{2}$ and value -1 with probability $\frac{1}{2}$.

The following important symmetrization lemma makes the connection between the uniform law of large numbers and the Rademacher complexity.

Lemma A.1 (Symmetrization Lemma)**.** *Let $\boldsymbol{F}$ be a set of functions. Then*

$$\mathbb{E} \sup_{u \in \boldsymbol{F}} \Big| \frac{1}{n} \sum_{j=1}^n u(X_j) - \mathbb{E}_{X \sim \boldsymbol{P}_\Omega} u(X) \Big| \leq 2R_n(\boldsymbol{F}).$$

Lemma A.2 (Ledoux-Talagrand contraction (Ledoux & Talagrand, 2013, Theorem 4.12))**.** *Assume that $\phi : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz with $\phi(0) = 0$. Let $\{\sigma_i\}_{i=1}^n$ be independent Rademacher random variables. Then for any $T \subset \mathbb{R}^n$*

$$\mathbb{E}_\sigma \Big[\sup_{(t_1, \cdots, t_n) \in T} \sum_{i=1}^n \sigma_i \phi(t_i) \Big] \leq 2L \cdot \mathbb{E}_\sigma \Big[\sup_{(t_1, \cdots, t_n) \in T} \sum_{i=1}^n \sigma_i t_i \Big].$$

Let (E, ρ) be a metric space with metric ρ. A *δ-cover* of a set $A \subset E$ with respect to ρ is a collection of points $\{x_1, \cdots, x_n\} \subset A$ such that for every $x \in A$, there exists $i \in \{1, \cdots, n\}$ such that $\rho(x, x_i) \leq \delta$. The δ-covering number $\boldsymbol{N}(\delta, A, \rho)$ is the cardinality of the smallest δ-cover of the set A with respect to the metric ρ. Equivalently, the δ-covering number $\boldsymbol{N}(\delta, A, \rho)$ is the minimal number of balls $B_\rho(x, \delta)$ of radius δ needed to cover the set A.

Theorem A.3 (Dudley's Integral theorem)**.** *Let $\boldsymbol{F}$ be a function class such that $\sup_{f \in \boldsymbol{F}} \|f\|_{n,2} \leq M$. Then the Rademacher complexity $R_n(\boldsymbol{F})$ satisfies that*

$$R_n(\boldsymbol{F}) \leq \inf_{0 \leq \delta \leq M} \Big\{ 4\delta + \frac{12}{\sqrt{n}} \int_\delta^M \sqrt{\log \boldsymbol{N}(\epsilon, \boldsymbol{F}, \|\cdot\|_{n,2})}\, d\epsilon \Big\}.$$

Lemma A.3 (Talagrand Concentration Inequality)**.** *Consider a function class $\mathcal{F}$ defined on a probability measure μ such that for all $f \in \mathcal{F}$, we have $\|f\|_\infty \leq \beta, \mathbb{E}_\mu[f] = 0, \mathbb{E}_\mu[f^2] \leq \sigma^2$. Then for any $t > 0$, we can have the following concentration results.*

$$\mathbb{P}_{z_1,\cdots,z_n\sim\mu}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(z_i) \geq 2\sup_{f\in\mathcal{F}} \mathbb{E}_{z_1',\cdots,z_n'\sim\mu}\frac{1}{n}\sum_{i=1}^n f(z_i') + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2t\beta}{n}\right] \leq e^{-t}$$

Lemma A.4 (Peeling lemma (Bartlett et al., 2005))**.** *Consider a measurable function class $\mathcal{F}$ defined on a probability measure μ such that for all $f \in \mathcal{F}$, we have $\|f\|_\infty \leq \beta, \mathbb{E}_\mu[f] = 0$. We further have a sub-root function $\phi(r)$ satisfies*

$$R_n(\{f \in \mathcal{F} \mid \mathbf{P}f \leq r\}) \leq \phi(r)\ (\forall\, r > 0). \tag{A.10}$$

Then we have

$$\mathbb{E}_{\sigma_i,z_n}\left[\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)}{\mathbf{P}f + r}\right] \leq \frac{4\phi(r)}{r}$$

Proof. Denote $\mathcal{F}(r) = \{f \in \mathcal{F} \mid \mathbf{P}f \leq r\}$ to be the localized set with radius r. Then for a fixed set of datapoints $\{z_i\}_{i=1}^n$ and a fixed set of Rademacher random variables $\{\sigma_i\}_{i=1}^n$

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)}{\mathbf{P}f + r} &\leq \sup_{f\in\mathcal{F}(r)} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)}{r} + \sum_{j=0}^\infty \sup_{f\in\mathcal{F}(r4^{j+1})\setminus\mathcal{F}(r4^j)} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i f(z_i)}{r4^j + r} \\
&\leq \frac{R_n(\mathcal{F}(r))}{r} + \sum_{j=0}^\infty \frac{R_n(\mathcal{F}(r4^{j+1}))}{r4^j + r} \leq \frac{\phi(r)}{r} + \sum_{j=0}^\infty \frac{\phi(r4^{j+1})}{r4^j + r} \\
&\leq \frac{\phi(r)}{r} + \sum_{j=0}^\infty \frac{2^{j+1}\phi(r)}{r4^j + r} \leq \frac{4\phi(r)}{r}
\end{aligned}$$

□

We also modify the peeling lemma above, as we aim to apply it to derive the upper bound for the Modified Deep Ritz Method (MDRM).

Lemma A.5 (Peeling Lemma For MDRM)**.** *Assume that there exists some function $\phi : [0,\infty) \to [0,\infty)$ and some $r^\star > 0$, such that for any $r > r^\star$, we have:*

$$\phi(4r) \leq 2\phi(r) \text{ and } R_{n,m}(\boldsymbol{F}_r) \leq \phi(r)$$

where the localized set $\boldsymbol{F}_r$ is defined as

$$\boldsymbol{F}_r = \{(g_f, h_f) \in \boldsymbol{F} \mid \mathbb{E}_{x,y}[f(x,y)] = \mathbb{E}_x[g_f(x)] + \mathbb{E}_y[h_f(y)] \leq r\}$$

and the Rademacher Complexity is defined as $R_{n,m}(\boldsymbol{F}_r) := R_n(\{g_f|(g_f,h_f) \in \boldsymbol{F}_r\} + R_m(\{h_f|(g_f,h_f) \in \boldsymbol{F}_r\}$. Then for any $r > r^\star$, we have

$$\mathbb{E}_{\sigma,\tau}\Big[\mathbb{E}_{x,y}[\sup_{f\in\boldsymbol{F}} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i g_f(x_i) + \frac{1}{m}\sum_{j=1}^m \tau_j h_f(y_j)}{\mathbb{E}_x[g_f(x)] + \mathbb{E}_y[h_f(y)] + r}]\Big] \leq \frac{4\phi(r)}{r}$$

Proof. The proof is the same as the original peeling lemma, thus we omit the detailed proof here. □

A.3.1 Local Rademacher Complexity of Truncated Fourier Basis

Definition A.1. *(Fourier Series) Given a domain $\Omega \subseteq [0,1]^d$. For any $z \in \mathbb{N}^d$, we consider the corresponding Fourier basis function $\phi_z(x) := e^{2\pi i\langle z,x\rangle}$ $(x \in \Omega)$. With respect to the Fourier basis, any function $f \in L^2(\Omega)$ can be decomposed as the following sum:*

$$f(x) := \sum_{z\in\mathbb{N}^d} f_z\phi_z(x). \tag{A.11}$$

where for any $z \in \mathbb{N}^d$, the Fourier coefficient $f_z = \int_\Omega f(x)\overline{\phi_z(x)}dx$.

Definition A.2. *(Truncated Fourier Series) For a fixed positive integer $\xi \in \mathbb{Z}^+$, we define the space $F_\xi(\Omega)$ of truncated Fourier series as follows:*

$$F_\xi(\Omega) := \Big\{ f = \sum_{z \in \mathbb{N}^d} f_z \phi_z \;\Big|\; f_z = 0,\ \forall\, \|z\|_\infty > \xi \Big\}. \tag{A.12}$$

Equivalently, we can decompose any $f \in F_\xi(\Omega)$ as $f := \sum_{\|z\|_\infty \le \xi} f_z \phi_z$.

Lemma A.6. *(Local Rademacher Complexity of Localized Truncated Fourier Series) For a fixed $\xi \in \mathbb{Z}^+$, we consider a localized class of functions $\boldsymbol{F}_{\rho,\xi}(\Omega) = \Big\{ f \in F_\xi(\Omega) \;\Big|\; \|f\|^2_{H^1(\Omega)} \le \rho \Big\}$, where $\rho > 0$ is fixed. Then we have the following upper bound on the local Rademacher complexity:*

$$R_n(\boldsymbol{F}_{\rho,\xi}(\Omega)) = \mathbb{E}_X \left[\mathbb{E}_\sigma \Big[\sup_{f \in \boldsymbol{F}_{\rho,\xi}(\Omega)} \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \;\Big|\; X_1, \cdots, X_n \Big] \right] \lesssim \sqrt{\frac{\rho}{n}} \xi^{\frac{d-2}{2}}. \tag{A.13}$$

Proof. Take an arbitrary function $f \in \boldsymbol{F}_{\rho,\xi}(\Omega)$. Let $f = \sum_{\|z\|_\infty \le \xi} f_z \phi_z$ be the Fourier basis expansion of f. On the one hand, substituting the Fourier expansion into the norm restriction and using orthogonality of the Fourier basis imply:

$$\begin{aligned}
\rho \ge \|f\|^2_{H^1(\Omega)} &= \| \sum_{\|z\|_\infty \le \xi} f_z \phi_z \|^2_{H^1(\Omega)} = \| \sum_{\|z\|_\infty \le \xi} f_z \phi_z \|^2_{L^2(\Omega)} + \| \sum_{\|z\|_\infty \le \xi} f_z \nabla \phi_z \|^2_{L^2(\Omega)} \\
&= \sum_{\|z\|_\infty \le \xi} |f_z|^2 \|\phi_z\|^2_{L^2(\Omega)} + \sum_{\|z\|_\infty \le \xi} |f_z|^2 \|\nabla \phi_z\|^2_{L^2(\Omega)} \\
&= |\Omega| \Big(\sum_{\|z\|_\infty \le \xi} |f_z|^2 + 4\pi^2 \sum_{\|z\|_\infty \le \xi} |f_z|^2 \|z\|^2 \Big), \Rightarrow \sum_{\|z\|_\infty \le \xi} |f_z|^2 \|z\|^2 \lesssim \rho.
\end{aligned}$$

On the other hand, substituting the Fourier expansion into the average sum $\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)$ and using Cauchy-Schwarz inequality imply:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) &= \frac{1}{n} \sum_{i=1}^n \sigma_i \sum_{\|z\|_\infty \le \xi} f_z \phi_z(X_i) = \frac{1}{n} \sum_{\|z\|_\infty \le \xi} \sum_{i=1}^n \sigma_i f_z \phi_z(X_i) \\
&\le \frac{1}{n} \Big(\sum_{\|z\|_\infty \le \xi} |f_z|^2 \|z\|^2 \Big)^{\frac{1}{2}} \Big(\sum_{\|z\|_\infty \le \xi} \Big| \sum_{i=1}^n \frac{\sigma_i}{\|z\|} \phi_z(X_i) \Big|^2 \Big)^{\frac{1}{2}} \\
&\lesssim \frac{\sqrt{\rho}}{n} \Big(\sum_{\|z\|_\infty \le \xi} \Big| \sum_{i=1}^n \frac{\sigma_i}{\|z\|} \phi_z(X_i) \Big|^2 \Big)^{\frac{1}{2}}.
\end{aligned}$$

where we have used the constraint $\sum_{\|z\|_\infty \le \xi} |f_z|^2 \|z\|^2 \lesssim \rho$ in the last step above. Moreover, by taking expectation with respect to the i.i.d Rademacher random variables σ_i $(1 \le i \le n)$ and the uniformly sampled data points $\{X_i\}_{i=1}^n$ on both sides and applying Jensen's inequality, we can deduce that:

$$\begin{aligned}
\mathbb{E}_X \mathbb{E}_\sigma \Big[\frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i) \Big] &\lesssim \frac{\sqrt{\rho}}{n} \mathbb{E}_{X,\sigma} \left[\Big(\sum_{\|z\|_\infty \le \xi} \Big| \sum_{i=1}^n \frac{\sigma_i}{\|z\|} \phi_z(X_i) \Big|^2 \Big)^{\frac{1}{2}} \right] \\
&\le \frac{\sqrt{\rho}}{n} \left(\mathbb{E}_{X,\sigma} \Big[\sum_{\|z\|_\infty \le \xi} \Big| \sum_{i=1}^n \frac{\sigma_i}{\|z\|} \phi_z(X_i) \Big|^2 \Big] \right)^{\frac{1}{2}}.
\end{aligned}$$

Using independence between the random variables σ_i $(1 \leq i \leq n)$, we can further simplify the expectation inside the square root above as below:

$$\begin{aligned}
\mathbb{E}_{X,\sigma}\Big[\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\phi_z(X_i)\Big|^2\Big] &= \sum_{\|z\|_\infty \leq \xi}\mathbb{E}_{X,\sigma}\Big[\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\phi_z(X_i)\Big|^2\Big] \\
&= \sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n \mathbb{E}_{X,\sigma}\Big[\frac{\sigma_i^2}{\|z\|^2}\Big|\phi_z(X_i)\Big|^2\Big] \\
&= \sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n \frac{|\Omega|}{\|z\|^2} \lesssim n\sum_{\|z\|_\infty \leq \xi}\frac{1}{\|z\|^2} \lesssim n\frac{\xi^d}{\xi^2} = n\xi^{d-2}.
\end{aligned}$$

Combining the two bounds above yields the desired upper bound:

$$\mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f \in \boldsymbol{F}_{\rho,\xi}(\Omega)}\frac{1}{n}\sum_{i=1}^n \sigma_i f(X_i)\;\Big|\;X_1,\cdots,X_n\Big]\right] \lesssim \frac{\sqrt{\rho}}{n}\sqrt{n\xi^{d-2}} = \sqrt{\frac{\rho}{n}}\xi^{\frac{d-2}{2}}.$$

□

Lemma A.7. *(Local Rademacher Complexity of Localized Truncated Fourier Series' Gradient) For a fixed $\xi \in \mathbb{Z}^+$, we consider a localized class of functions $\boldsymbol{G}_{\rho,\xi}(\Omega) = \{\|\nabla f\| \mid f \in F_{\rho,\xi}(\Omega)\}$, where $\rho > 0$ is fixed. Then for any sample $\{X_i\}_{i=1}^n \subset \Omega$, we have the following upper bound on the local Rademacher complexity:*

$$R_n(\boldsymbol{G}_{\rho,\xi}(\Omega)) = \mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f \in \boldsymbol{F}_{\rho,\xi}(\Omega)}\frac{1}{n}\sum_{i=1}^n \sigma_i\|\nabla f(X_i)\|\;\Big|\;X_1,\cdots,X_n\Big]\right] \lesssim \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}}. \tag{A.14}$$

Proof. Take an arbitrary function $f \in \boldsymbol{F}_{\rho,\xi}(\Omega)$. Let $f = \sum_{\|z\|_\infty \leq \xi} f_z\phi_z$ be the Fourier basis expansion of f. Similarly, the norm restriction condition $\|f\|_{H^1(\Omega)}^2 \leq \rho$ can be reduced to the following condition about Fourier coefficients:

$$\sum_{\|z\|_\infty \leq \xi}|f_z|^2\|z\|^2 \lesssim \rho.$$

Moreover, substituting the Fourier expansion into the average sum $\frac{1}{n}\sum_{i=1}^n \sigma_i\|\nabla f(X_i)\|$ and using Cauchy-Schwarz inequality imply:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \sigma_i\|\nabla f(X_i)\| &= \frac{1}{n}\sum_{i=1}^n \sigma_i\|\sum_{\|z\|_\infty \leq \xi} f_z\nabla\phi_z(X_i)\| \leq \frac{1}{n}\sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n \sigma_i\|f_z\nabla\phi_z(X_i)\| \\
&\leq \frac{1}{n}\Big(\sum_{\|z\|_\infty \leq \xi}|f_z|^2\|z\|^2\Big)^{\frac{1}{2}}\Big(\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\|\nabla\phi_z(X_i)\|\Big|^2\Big)^{\frac{1}{2}} \\
&\lesssim \frac{\sqrt{\rho}}{n}\Big(\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\|\nabla\phi_z(X_i)\|\Big|^2\Big)^{\frac{1}{2}}.
\end{aligned}$$

where we have used the constraint $\sum_{\|z\|_\infty \leq \xi}|f_z|^2\|z\|^2 \lesssim \rho$ in the last step above. Moreover, by taking expectation with respect to the i.i.d Rademacher random variables σ_i $(1 \leq i \leq n)$ and the uniformly sampled data points $\{X_i\}_{i=1}^n$ on both sides and applying Jensen's inequality, we can deduce that:

$$\begin{aligned}
\mathbb{E}_X\mathbb{E}_\sigma\Big[\frac{1}{n}\sum_{i=1}^n \sigma_i\|\nabla f(X_i)\|\Big] &\lesssim \frac{\sqrt{\rho}}{n}\mathbb{E}_{X,\sigma}\left[\Big(\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\|\nabla\phi_z(X_i)\|\Big|^2\Big)^{\frac{1}{2}}\right] \\
&\leq \frac{\sqrt{\rho}}{n}\left(\mathbb{E}_{X,\sigma}\Big[\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\|\nabla\phi_z(X_i)\|\Big|^2\Big]\right)^{\frac{1}{2}}.
\end{aligned}$$

Using independence between the random variables σ_i $(1 \leq i \leq n)$, we can further simplify the expectation inside the square root above as below:

$$\begin{aligned}\mathbb{E}_{X,\sigma}\Big[\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\|\nabla\phi_z(X_i)\|\Big|^2\Big] &= \sum_{\|z\|_\infty \leq \xi}\mathbb{E}_{X,\sigma}\Big[\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|}\|\nabla\phi_z(X_i)\|\Big|^2\Big] \\ &= \sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n \mathbb{E}_{X,\sigma}\Big[\frac{\sigma_i^2}{\|z\|^2}\|\nabla\phi_z(X_i)\|^2\Big] \\ &= \sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n |\Omega|\frac{4\pi^2\|z\|^2}{\|z\|^2} \lesssim n\sum_{\|z\|_\infty \leq \xi} 1 \lesssim n\xi^d.\end{aligned}$$

Combining the two bounds above yields the desired upper bound:

$$\mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f\in \boldsymbol{F}_{\rho,\xi}(\Omega)}\frac{1}{n}\sum_{i=1}^n \sigma_i\|\nabla f(X_i)\| \;\Big|\; X_1,\cdots,X_n\Big]\right] \lesssim \frac{\sqrt{\rho}}{n}\sqrt{n\xi^d} = \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}}.$$

□

Lemma A.8. *(Local Rademacher Complexity of Localized Truncated Fourier Series' Laplacian) For a fixed $\xi \in \mathbb{Z}^+$, we consider a localized class of functions $\boldsymbol{J}_{\rho,\xi}(\Omega) := \Big\{f \in F_\xi(\Omega) \;\Big|\; \|f\|^2_{H^2(\Omega)} \leq \rho\Big\}$, where $\rho > 0$ is fixed. Correspondingly, we define a localized class of Laplacians $\boldsymbol{K}_{\rho,\xi}(\Omega) := \{\Delta f \mid f \in J_{\rho,\xi}(\Omega)\}$. Then for any sample $\{X_i\}_{i=1}^n \subset \Omega$, we have the following upper bound on the local Rademacher complexity:*

$$R_n(\boldsymbol{K}_{\rho,\xi}(\Omega)) = \mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f\in \boldsymbol{F}_{\rho,\xi}(\Omega)}\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta f(X_i) \;\Big|\; X_1,\cdots,X_n\Big]\right] \lesssim \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}}. \tag{A.15}$$

Proof. Take an arbitrary function $f \in \boldsymbol{J}_{\rho,\xi}(\Omega)$. Let $f = \sum_{\|z\|_\infty \leq \xi} f_z\phi_z$ be the Fourier basis expansion of f. Similarly, the norm restriction condition $\|f\|^2_{H^2(\Omega)} \leq \rho$ can be reduced to the following condition about fourier coefficients:

$$\sum_{\|z\|_\infty \leq \xi}|f_z|^2\|z\|^4 \lesssim \rho.$$

Moreover, substituting the Fourier expansion into the average sum $\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta f(X_i)$ and using Cauchy-Schwarz inequality imply:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta f(X_i) &= \frac{1}{n}\sum_{i=1}^n \sigma_i\sum_{\|z\|_\infty \leq \xi} f_z\Delta\phi_z(X_i) = \frac{1}{n}\sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n \sigma_i f_z\Delta\phi_z(X_i) \\ &\leq \frac{1}{n}\Big(\sum_{\|z\|_\infty \leq \xi}|f_z|^2\|z\|^4\Big)^{\frac{1}{2}}\Big(\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|^2}\Delta\phi_z(X_i)\Big|^2\Big)^{\frac{1}{2}} \\ &\lesssim \frac{\sqrt{\rho}}{n}\Big(\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|^2}\Delta\phi_z(X_i)\Big|^2\Big)^{\frac{1}{2}}.\end{aligned}$$

where we have used the constraint $\sum_{\|z\|_\infty \leq \xi}|f_z|^2\|z\|^4 \lesssim \rho$ in the last step above. Moreover, by taking expectation with respect to the i.i.d Rademacher random variables σ_i $(1 \leq i \leq n)$ and the uniformly sampled data points $\{X_i\}_{i=1}^n$ on both sides and applying Jensen's inequality, we can deduce that:

$$\begin{aligned}\mathbb{E}_X\mathbb{E}_\sigma\Big[\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta f(X_i)\Big] &\lesssim \frac{\sqrt{\rho}}{n}\mathbb{E}_{X,\sigma}\left[\Big(\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|^2}\Delta\phi_z(X_i)\Big|^2\Big)^{\frac{1}{2}}\right] \\ &\leq \frac{\sqrt{\rho}}{n}\left(\mathbb{E}_{X,\sigma}\Big[\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|^2}\Delta\phi_z(X_i)\Big|^2\Big]\right)^{\frac{1}{2}}.\end{aligned}$$

Using independence between the random variables σ_i $(1 \leq i \leq n)$, we can further simplify the expectation inside the square root above as below:

$$\begin{aligned}\mathbb{E}_{X,\sigma}\Big[\sum_{\|z\|_\infty \leq \xi}\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|^2}\Delta\phi_z(X_i)\Big|^2\Big] &= \sum_{\|z\|_\infty \leq \xi}\mathbb{E}_{X,\sigma}\Big[\Big|\sum_{i=1}^n \frac{\sigma_i}{\|z\|^2}\Delta\phi_z(X_i)\Big|^2\Big] \\ &= \sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n \mathbb{E}_{X,\sigma}\Big[\frac{\sigma_i^2}{\|z\|^4}|\Delta\phi_z(X_i)|^2\Big] \\ &= \sum_{\|z\|_\infty \leq \xi}\sum_{i=1}^n |\Omega|\frac{16\pi^4\|z\|^4}{\|z\|^4} \lesssim n\sum_{\|z\|_\infty \leq \xi} 1 \lesssim n\xi^d.\end{aligned}$$

Combining the two bounds above yields the desired upper bound:

$$\mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f\in \boldsymbol{F}_{\rho,\xi}(\Omega)}\frac{1}{n}\sum_{i=1}^n \sigma_i\Delta f(X_i)\;\Big|\; X_1,\cdots,X_n\Big]\right] \lesssim \frac{\sqrt{\rho}}{n}\sqrt{n\xi^d} = \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}}.$$

□

A.3.2 Local Rademacher Complexity of the Deep Neural Network Model

In this section we aim to bound the local Rademacher Complexity of a Deep Neural Network. We first bound the covering number of the function space composed by the gradient of all possible neural networks and then apply a Duley Integral to achieve the final bound.

Definition A.3. *Let η_l denote the l-ReLU activiation function. Here we use $\eta_3 := \max\{0, x\}^3$(E & Yu, 2018) as the activation function to ensure smoothness. We can define the space consisting of all neural network models with depth L, width W, sparsity constraint S and norm constraint B as follows:*

$$\Phi(L,W,S,B) := \Big\{(\mathcal{W}^{(L)}\eta_3(\cdot)+b^{(L)})\cdots(\mathcal{W}^{(1)}x+b^{(1)}) \mid \mathcal{W}^{(L)}\in\mathbb{R}^{1\times W}, b^{(L)}\in\mathbb{R}, \tag{A.16}$$

$$\mathcal{W}^{(1)}\in\mathbb{R}^{W\times d}, b^{(1)}\in\mathbb{R}^W, \mathcal{W}^{(l)}\in\mathbb{R}^{W\times W}, b^{(l)}\in\mathbb{R}^W (1<l<L), \tag{A.17}$$

$$\sum_{l=1}^L(\|\mathcal{W}^{(l)}\|_0+\|b^{(l)}\|_0)\leq S, \max_l\|\mathcal{W}^{(l)}\|_{\infty,\infty}\vee\|b^{(l)}\|_\infty\leq B\Big\}. \tag{A.18}$$

where $\|\cdot\|_0$ measures the number of nonzero entries in a matrix and $\|\cdot\|_{\infty,\infty}$ measures the maximum of the absolute values of the entries in a matrix.
For any $d\in\mathbb{Z}^+$, we refer to an arbitrary element in $\Phi(L,W,S,B)$ as a ReLU3 Deep Neural Network. Then for any index $1\leq k\leq L$, we use F_k to denote the $k-$ReLU3 Deep Neural Network composed by the first k layers, i.e:

$$F_k(x) := (\mathcal{W}_F^{(k)}\eta_3(\cdot)+b_F^{(k)})\cdots(\mathcal{W}_F^{(1)}x+b_F^{(1)}).$$

Also, we use $\Phi_k(L,W,S,B)$ to denote the space consisting of all F_k. In particular, when $k=L$, we have:

$$F(x) := F_L(x) = (\mathcal{W}_F^{(L)}\eta_3(\cdot)+b_F^{(L)})\cdots(\mathcal{W}_F^{(1)}x+b_F^{(1)}),\ \textit{and}\ \Phi_L(L,W,S,B)=\Phi(L,W,S,B).$$

Furthermore, given that the domain $\Omega\subset[0,1]^d$ is bounded, we have $\sup_{x\in\Omega}\|x\|_\infty = 1$

Lemma A.9. *(Upper bound on ∞-norm of functions in DNN space) For any $1\leq k\leq L$, the following inequality holds:*

$$\sup_{x\in\Omega,\ F_k\in\Phi_k(L,W,S,B)}\|F_k(x)\|_\infty \leq W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-1}{2}}2^{\frac{3^k-1}{2}-k+1}.$$

Proof. We use induction to prove this claim.
Base cases: When $k=1$, we have that for any $x\in\Omega$ and any $F_1\in\Phi_1(L,W,S,B)$, the following holds:

$$\begin{aligned}\|F_1(x)\|_\infty &= \|\mathcal{W}_F^{(1)}x+b_F^{(1)}\|_\infty \leq \|\mathcal{W}_F^{(1)}\|_\infty\|x\|_\infty + \|b_F^{(1)}\|_\infty \\ &\leq d\|\mathcal{W}_F^{(1)}\|_{\infty,\infty}+B \leq dB+B \leq 2(B\vee d)^2.\end{aligned} \tag{A.19}$$

When $k = 2$, we have that for any $x \in \Omega$ and any $F_2 \in \Phi_2(L, W, S, B)$, the following holds:

$$\|F_2(x)\|_\infty = \|\mathcal{W}_F^{(2)}\eta_3(F_1(x)) + b_F^{(2)}\|_\infty \leq \|\mathcal{W}_F^{(2)}\|_\infty \|\eta_3(F_1(x))\|_\infty + \|b_F^{(2)}\|_\infty \leq W\|\mathcal{W}_F^{(2)}\|_{\infty,\infty}\|F_1(x)\|_\infty^3 + B.$$

By applying the bound proved in the case when $k = 1$, we have:

$$\begin{aligned}\|F_2(x)\|_\infty &\leq WB(dB + B)^3 + B = WB^4(d+1)^3 + B \\ &= WB^4(d^3 + 3d^2 + 3d + 1) + B \leq 8W(B \vee d)^7.\end{aligned}$$

where the last inequality follows from the assumption that $W \geq 2$.
Inductive Step: Now we assume that the claim has been proved for $k-1$, where $3 \leq k \leq L$. Similarly, for any $x \in \Omega$ and any $F_k \in \Phi_k(L, W, S, B)$, we have:

$$\begin{aligned}\|F_k(x)\|_\infty &= \|\mathcal{W}_F^{(k)}\eta_3(F_{k-1}(x)) + b_F^{(k)}\|_\infty \leq \|\mathcal{W}_F^{(k)}\|_\infty \|\eta_3(F_{k-1}(x))\|_\infty + \|b_F^{(k)}\|_\infty \\ &\leq W\|\mathcal{W}_F^{(k)}\|_{\infty,\infty}\|F_{k-1}(x)\|_\infty^3 + B \leq WB\|F_{k-1}(x)\|_\infty^3 + B.\end{aligned}$$

Using inductive hypothesis, we can further deduce that:

$$\begin{aligned}\|F_k(x)\|_\infty &\leq WB \times W^{\frac{3^{k-1}-3}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-3}{2}} 2^{\frac{3^k-3}{2}-3k+6} + B \\ &\leq W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-3}{2}-3k+6} + B \vee d \\ &\leq W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} [2^{\frac{3^k-3}{2}-3k+6} + 1] \\ &\leq W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-3}{2}-k+2} \ (k \geq 3) \\ &= W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1}.\end{aligned}$$

Taking supremum with respect to $x \in \Omega$ and $F_k \in \Phi_k(L, W, S, B)$ on the LHS implies that the given upper bound also holds for k. By induction, the claim is proved. □

We also need to show that the ReLU3 activation function is a Lipschitzness functions over a bounded domain.

Lemma A.10. *For any $k \in \mathbb{Z}^+$, consider the $k-$ReLU activation function η_k defined on some bounded domain $\mathcal{D} \subset \mathbb{R}^d$ (i.e, $\sup_{x\in\mathcal{D}} \|x\|_\infty \leq C$ for some $C > 0$). Then we have that for any $x, y \in \mathcal{D}$, the following inequalities hold:*

$$\begin{aligned}&\|\eta_1(x) - \eta_1(y)\|_\infty \leq \|x - y\|_\infty, \\ &\|\eta_2(x) - \eta_2(y)\|_\infty \leq 2C\|x - y\|_\infty, \\ &\|\eta_3(x) - \eta_3(y)\|_\infty \leq 3C^2\|x - y\|_\infty.\end{aligned}$$

Proof. This is because $|\nabla\eta_1(x)| = |\max\{1, 0\}| = 1$, $|\nabla\eta_2(x)| = |2\max\{x, 0\}| \leq 2C$ and $|\nabla\eta_3(x)| = |3\max\{x, 0\}^2| \leq 3C^2$. □

Lemma A.11. *(Relation between the covering number of DNN space and parameter space) For any $1 \leq k \leq L$, suppose that a pair of different two networks $F_k, G_k \in \Phi_k(L, W, S, B)$ are given by:*

$$\begin{aligned}F_k(x) &:= (\mathcal{W}_F^{(k)}\eta_3(\cdot) + b_F^{(k)}) \cdots (\mathcal{W}_F^{(1)}x + b_F^{(1)}), \\ G_k(x) &:= (\mathcal{W}_G^{(k)}\eta_3(\cdot) + b_G^{(k)}) \cdots (\mathcal{W}_G^{(1)}x + b_G^{(1)}).\end{aligned}$$

Furthermore, assume that the $\| \|_\infty$ norm of the distance between the parameter spaces of F_k and G_k is uniformly upper bounded by δ, i.e

$$\|W_F^{(l)} - W_G^{(l)}\|_{\infty,\infty} \leq \delta, \ \|b_F^{(l)} - b_G^{(l)}\|_\infty \leq \delta, \ (\forall\, 1 \leq l \leq k). \tag{A.20}$$

Then we have:

$$\sup_{x\in\Omega} \|F_k(x) - G_k(x)\|_\infty \leq \delta W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1} 3^{k-1}. \tag{A.21}$$

Proof. Let's prove the claim by using induction on k.
Base Case: When $k = 1$, we have that for any $x \in \Omega$ and any $F_1, G_1 \in \Phi_1(L, W, S, B)$ satisfying constraint A.20, the following holds:

$$\begin{aligned}\|F_1(x) - G_1(x)\|_\infty &= \|\mathcal{W}_F^{(1)} x + b_F^{(1)} - \mathcal{W}_G^{(1)} x - b_G^{(1)}\|_\infty \\ &\leq \|\mathcal{W}_F^{(1)} - \mathcal{W}_G^{(1)}\|_\infty \|x\|_\infty + \|b_F^{(1)} - b_G^{(1)}\|_\infty \\ &\leq \delta d + \delta = \delta(d+1) \leq 2\delta(B \vee d) \leq 2\delta(B \vee d)^2.\end{aligned} \tag{A.22}$$

When $k = 2$, we have that for any $x \in \Omega$ and any $F_2, G_2 \in \Phi_2(L, W, S, B)$ satisfying constraint A.20, the following inequality holds:

$$\begin{aligned}\|F_2(x) - G_2(x)\|_\infty &= \|\mathcal{W}_F^{(2)} \eta_3(F_1(x)) + b_F^{(2)} - \mathcal{W}_G^{(2)} \eta_3(G_1(x)) - b_G^{(2)}\|_\infty \\ &\leq \|\mathcal{W}_F^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(G_1(x))\|_\infty + \|b_F^{(2)} - b_G^{(2)}\|_\infty \\ &\leq \|\mathcal{W}_F^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(F_1(x))\|_\infty + \|\mathcal{W}_G^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(G_1(x))\|_\infty + \delta.\end{aligned}$$

By applying the upper bound proved in equation A.19, we can upper bound the first part $\|\mathcal{W}_F^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(F_1(x))\|_\infty$ by:

$$\begin{aligned}\|\mathcal{W}_F^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(F_1(x))\|_\infty &\leq \|\mathcal{W}_F^{(2)} - \mathcal{W}_G^{(2)}\|_\infty \|\eta_3(F_1(x))\|_\infty \\ &\leq W\delta \|F_1(x)\|_\infty^3 \leq \delta W [2(B \vee d)^2]^3.\end{aligned}$$

By applying the Lipschitz condition proved in Lemma A.10 and the bound proved in equation A.22, we can further upper bound the second part $\|\mathcal{W}_G^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(G_1(x))\|_\infty$ by:

$$\begin{aligned}\|\mathcal{W}_G^{(2)} \eta_3(F_1(x)) - \mathcal{W}_G^{(2)} \eta_3(G_1(x))\|_\infty &\leq \|\mathcal{W}_G^{(2)}\|_\infty \|\eta_3(F_1(x)) - \eta_3(G_1(x))\|_\infty \\ &\leq WB \times 3 \sup_{F_1 \in \Phi_1(L,W,S,B)} \|F_1(x)\|_\infty^2 \times \|F_1(x) - G_1(x)\|_\infty \\ &\leq WB \times 3[2(B \vee d)^2]^2 \times 2\delta(B \vee d) \\ &\leq 24\delta W (B \vee d)^6.\end{aligned}$$

Summing the two upper bounds above yields:

$$\|F_2(x) - G_2(x)\|_\infty \leq 8\delta W (B \vee d)^6 + 24\delta W (B \vee d)^6 + \delta \leq 24\delta W (B \vee d)^7.$$

where we again use the assumption $d \geq 2$ in the last step.
Inductive Step: Now we assume that the claim has been proved for $k - 1$, where $k \geq 3$. For any $x \in \Omega$ and $F_k \in \Phi_k(L, W, S, B)$, we have that:

$$\begin{aligned}\|F_k(x) - G_k(x)\|_\infty &= \|\mathcal{W}_F^{(k)} \eta_3(F_{k-1}(x)) + b_F^{(k)} - \mathcal{W}_G^{(k)} \eta_3(G_{k-1}(x)) - b_G^{(k)}\|_\infty \\ &\leq \|\mathcal{W}_F^{(k)} \eta_3(F_{k-1}(x)) - \mathcal{W}_G^{(k)} \eta_3(G_{k-1}(x))\|_\infty + \|b_F^{(k)} - b_G^{(k)}\|_\infty \\ &\leq \|\mathcal{W}_F^{(k)} \eta_3(F_{k-1}(x)) - \mathcal{W}_G^{(k)} \eta_3(G_{k-1}(x))\|_\infty + \delta.\end{aligned}$$

Applying triangle inequality helps us upper bound the first term above as follows:

$$\begin{aligned}&\|\mathcal{W}_F^{(k)} \eta_3(F_{k-1}(x)) - \mathcal{W}_G^{(k)} \eta_3(G_{k-1}(x))\|_\infty \\ &\leq \|\mathcal{W}_F^{(k)} \eta_3(F_{k-1}(x)) - \mathcal{W}_G^{(k)} \eta_3(F_{k-1}(x))\|_\infty + \|\mathcal{W}_G^{(k)} \eta_3(F_{k-1}(x)) - \mathcal{W}_G^{(k)} \eta_3(G_{k-1}(x))\|_\infty \\ &\leq \|\mathcal{W}_F^{(k)} - \mathcal{W}_G^{(k)}\|_\infty \|\eta_3(F_{k-1}(x))\|_\infty + \|\mathcal{W}_G^{(k)}\|_\infty \|\eta_3(F_{k-1}(x)) - \eta_3(G_{k-1}(x))\|_\infty \\ &\leq \delta W \|F_{k-1}(x)\|_\infty^3 + BW \|\eta_3(F_{k-1}(x)) - \eta_3(G_{k-1}(x))\|_\infty.\end{aligned}$$

From Lemma A.9, we can upper bound the first term $\delta W \|F_{k-1}(x)\|_\infty^3$ by:

$$\delta W \|F_{k-1}(x)\|_\infty^3 \leq \delta W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-3}{2}} 2^{\frac{3^k-3}{2} - 3k + 6}.$$

Moreover, applying Lemma A.10 and the inductive hypothesis let us upper bound the second term $BW\|\eta_3(F_{k-1}(x)) - \eta_3(G_{k-1}(x))\|_\infty$ as follows:

$$
\begin{aligned}
&BW\|\eta_3(F_{k-1}(x)) - \eta_3(G_{k-1}(x))\|_\infty \\
&\leq BW \times 3 \sup_{x\in\Omega,\ F_{k-1}\in\Phi_{k-1}(L,W,S,B)} \|F_{k-1}(x)\|_\infty^2 \times \|F_{k-1}(x) - G_{k-1}(x)\|_\infty \\
&\leq 3BW \times W^{3^{k-2}-1}(B\vee d)^{5\times 3^{k-2}-1} 2^{3^{k-1}-1-2k+4}\|F_{k-1}(x) - G_{k-1}(x)\|_\infty \\
&\leq 3BW \times W^{3^{k-2}-1}(B\vee d)^{5\times 3^{k-2}-1} 2^{3^{k-1}-1-2k+4} \times \delta W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} \\
&\leq 3^{k-1}\delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\times 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-3k+5}.
\end{aligned}
$$

Combining the two upper bounds derived above yields:

$$
\begin{aligned}
\|F_k(x) - G_k(x)\|_\infty &\leq \delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}} 2^{\frac{3^k-3}{2}-3k+6} \\
&\quad + 3^{k-1}\delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\times 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-3k+5} + \delta \\
&\leq \delta 3^{k-1} W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\times 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1},
\end{aligned}
$$

where the last inequality above follows from $k \geq 3$. Taking supremum with respect to $x \in \Omega$ on the LHS implies the given upper bound also holds for k. By induction, the claim is proved. □

Theorem A.4. *(Bounding the DNN space covering number) Fix some sufficiently large $N \in \mathbb{Z}^+$. Consider a Deep Neural Network space $\Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$. Then the* log *value of the covering number of this DNN space with respect to the inf-norm $\|F(x)\|_\infty := \sup_{x\in\Omega}|F(x)|$, which is denoted by $\mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty)$, can be upper bounded by:*

$$
\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty) = O\left(S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]\right). \tag{A.23}
$$

Proof. We firstly fix a sparsity pattern (i.e, the locations of the non-zero entries are fixed). By picking $k = L$ in Lemma A.11, we get the following upper bound on the covering number with respect to $\|\cdot\|_\infty$:

$$
\left(\frac{\delta}{3^{L-1}W^{\frac{3^{L-1}-1}{2}}(B\vee d)^{\frac{5\times 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+2}}\right)^{-S}.
$$

Furthermore, note that the number of feasible configurations is upper bounded by $\binom{(W+1)^L}{S} \leq (W+1)^{LS}$.(Schmidt-Hieber et al., 2020; Farrell et al., 2021) Plug in the previous inequality and yields:

$$
\begin{aligned}
\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty) &\leq \log\left[(W+1)^{LS}\left(\frac{\delta}{3^{L-1}W^{\frac{3^{L-1}-1}{2}}(B\vee d)^{\frac{5\times 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+1}}\right)^{-S}\right] \\
&\leq S\log\left[\delta^{-1}(W+1)^L 3^{L-1} W^{\frac{3^{L-1}-1}{2}}(B\vee d)^{\frac{5\times 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+1}\right] \\
&\lesssim S\Big[\log(\delta^{-1}) + L\log(3W) + 3^L\log(W(B\vee d)) + 3^L\log 2\Big].
\end{aligned}
$$

Note that here the dimension d is some constant. Thus, by plugging in thee given magnitudes $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$, we can further deduce that:

$$
\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty) \lesssim S\Big[\log(\delta^{-1}) + 3^L\log(WB)\Big].
$$

This finishes our proof. □

Now let's consider upper bounding the covering number of the l_2 norm of the sparse Deep Neural Networks' gradients. Note that for any $1 \leq k \leq L-1$, any $k-$ReLU3 Deep Neural Network $F_k \in \Phi_k(L, W, S, B)$ is a map from $\mathbb{R}^d$ to $\mathbb{R}^W$. For any $1 \leq l \leq W$, we use $F_{k,l}(x)$ to denote the

l-th component of the map F_k. This helps us write the map $F_k(x)$ and its Jacobian matrix $J[F_k](x)$ explicitly as:

$$F_k(x) = [F_{k,1}(x), F_{k,2}(x), \cdots, F_{k,W}(x)]^T \in \mathbb{R}^W.$$

$$J[F_k](x) = \begin{bmatrix} \frac{\partial}{\partial x_1} F_{k,1}(x) & \frac{\partial}{\partial x_2} F_{k,1}(x) & \cdots & \frac{\partial}{\partial x_d} F_{k,1}(x) \\ \frac{\partial}{\partial x_1} F_{k,2}(x) & \frac{\partial}{\partial x_2} F_{k,2}(x) & \cdots & \frac{\partial}{\partial x_d} F_{k,2}(x) \\ \cdots & \cdots & \ddots & \\ \frac{\partial}{\partial x_1} F_{k,W}(x) & \frac{\partial}{\partial x_2} F_{k,W}(x) & \cdots & \frac{\partial}{\partial x_d} F_{k,W}(x) \end{bmatrix} \in \mathbb{R}^{W \times d}.$$

In particular, when $k = L$, we have that any $F_L \in \Phi_L(L, W, S, B) = \Phi(L, W, S, B)$ is a map from $\mathbb{R}^d$ to $\mathbb{R}$. Thus, its Jacobian can be explicitly written as the following row vector:

$$J[F_L](x) = [\frac{\partial}{\partial x_1} F_L(x), \frac{\partial}{\partial x_2} F_L(x), \cdots \frac{\partial}{\partial x_d} F_L(x)] \in \mathbb{R}^{1 \times d}.$$

Lemma A.12. *(Upper bound on ∞-norm of Jacobian/Gradient of elements in the DNN space) For any $1 \leq k \leq L$, the following inequality holds:*

$$\sup_{x \in \Omega, F_k \in \Phi_k(L,W,S,B)} \|J[F_k](x)\|_\infty \leq W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1} 3^{k-1}.$$

Proof. We use induction on k to prove the claim.
Base case: $k = 1$. By the definition of Jacobian matrix, we have that for any $x \in \Omega$ and any $F_1 \in \Phi_1(L, W, S, B)$, the following holds:

$$\|J[F_1](x)\|_\infty = \|\mathcal{W}_F^{(1)}\|_\infty \leq dB \leq 2(B \vee d)^2.$$

Inductive Step: Assume that the claim has been proved for $k - 1$, where $2 \leq k \leq L$. For any $x \in \Omega$ and any $F_k \in \Phi_k(L, W, S, B)$, by applying the Chain Rule, we can write the Jacobian matrix $J[F_k](x)$ as $J[F_k](x) = \mathcal{W}_F^{(k)} J[\eta_3 \circ F_{k-1}](x)$, where the ReLU3 activation function η_3 is applied to each component $F_{k-1,l}$ $(1 \leq l \leq W)$ of the map F_k. Then we have the following upper bound:

$$\|J[F_k](x)\|_\infty \leq \|\mathcal{W}_F^{(k)}\|_\infty \|J[\eta_3 \circ F_{k-1}](x)\|_\infty \leq WB \|J[\eta_3 \circ F_{k-1}](x)\|_\infty. \tag{A.24}$$

Note that the composition $\eta_3 \circ F_{k-1}$ is a map from $\mathbb{R}^d$ to $\mathbb{R}^W$. Hence, the Jacobian matrix $J[\eta_3 \circ F_{k-1}](x)$ is of shape $\mathbb{R}^{W \times d}$. Applying the Chain Rule again implies:

$$\|J[\eta_3 \circ F_{k-1}](x)\|_\infty = \sup_{1 \leq l \leq W} (\sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j}|).$$

Furthermore, for any $1 \leq l \leq W$, the summation on the RHS above can be upper bounded by:

$$\sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j}| \leq 3\|F_{k-1}(x)\|_\infty^2 (\sum_{j=1}^d |\frac{\partial}{\partial x_j} F_{k-1,l}(x)|) \leq 3\|F_{k-1}(x)\|_\infty^2 \|J[F_{k-1}](x)\|_\infty.$$

Now let's take supremum with respect to l and apply the inductive hypothesis and Lemma A.9. This yields:

$$\begin{aligned} \|J[\eta_3 \circ F_{k-1}](x)\|_\infty &\leq 3W^{3^{k-2}-1} (B \vee d)^{5 \cdot 3^{k-2}-1} 2^{3^{k-1}-1-2k+4} \times W^{\frac{3^{k-2}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} \\ &= W^{\frac{3^{k-1}-1}{2}-1} (B \vee d)^{\frac{5 \cdot 3^{k-1}-1}{2}-1} 2^{\frac{3^k-1}{2}-3k+5} 3^{k-1}. \end{aligned} \tag{A.25}$$

By substituting equation A.25 into equation A.24, we can derive the final bound:

$$\begin{aligned} \|J[F_k](x)\|_\infty &\leq WB \|J[\eta_3 \circ F_{k-1}](x)\|_\infty \leq W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-3k+5} 3^{k-1} \\ &\leq W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1} 3^{k-1}. \end{aligned}$$

where the last inequality above follows from $k \geq 2$. Taking supremum with respect to $x \in \Omega$ and $F_k \in \Phi_k(L, W, S, B)$ on the LHS implies that the given upper bound also holds for k. By induction, the claim is proved. □

For the convenience of the following proof, we first prove this lemma for vector 2 norm and ∞ norm.

Lemma A.13. *Given any two row vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{1\times d}$, we have:*

$$\Big| \|\boldsymbol{u}\| - \|\boldsymbol{v}\| \Big| \leq \|\boldsymbol{u} - \boldsymbol{v}\|_\infty.$$

Proof. Assume that the two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ can be explicitly written as $\boldsymbol{u} = [u_1, u_2, \cdots, u_d]$ and $v = [v_1, v_2, \cdots, v_d]$, respectively. By applying Cauchy-Schwarz inequality, we have:

$$\begin{aligned}
\Big| \|\boldsymbol{u}\| - \|\boldsymbol{v}\| \Big|^2 &= \left| \sqrt{\sum_{i=1}^d u_i^2} - \sqrt{\sum_{i=1}^d v_i^2} \right|^2 \\
&= \sum_{i=1}^d u_i^2 + \sum_{i=1}^d v_i^2 - 2\sqrt{\sum_{i=1}^d u_i^2}\sqrt{\sum_{i=1}^d v_i^2} \\
&\leq \sum_{i=1}^d u_i^2 + \sum_{i=1}^d v_i^2 - 2\sum_{i=1}^d u_i v_i = \sum_{i=1}^d |u_i - v_i|^2 \\
&\leq \Big(\sum_{i=1}^d |u_i - v_i| \Big)^2 = \|\boldsymbol{u} - \boldsymbol{v}\|_\infty^2.
\end{aligned}$$

Taking the square root on both sides yields the desired inequality. □

Then we upper bound the Lipschitz constant of the gradient of the neural network. Given a DNN space $\Phi(L, W, S, B)$, we define a corresponding DNN Gradient space $\nabla\Phi(L, W, S, B)$ as:

$$\nabla\Phi(L, W, S, B) := \{\|\nabla F\| \mid F \in \Phi(L, W, S, B)\}. \tag{A.26}$$

Lemma A.14. *(Relation between the covering number of the DNN Gradient space and parameter space) For any $1 \leq k \leq L$, suppose that a pair of different two networks $F_k, G_k \in \Phi_k(L, W, S, B)$ are given by:*

$$\begin{aligned}
F_k(x) &:= (\mathcal{W}_F^{(k)}\eta_3(\cdot) + b_F^{(k)}) \cdots (\mathcal{W}_F^{(1)}x + b_F^{(1)}), \\
G_k(x) &:= (\mathcal{W}_G^{(k)}\eta_3(\cdot) + b_G^{(k)}) \cdots (\mathcal{W}_G^{(1)}x + b_G^{(1)}).
\end{aligned}$$

Furthermore, assume that the $\| \ \|_\infty$ norm of the distance between the parameter spaces is uniformly upper bounded by δ, i.e

$$\|W_F^{(l)} - W_G^{(l)}\|_{\infty,\infty} \leq \delta, \ \|b_F^{(l)} - b_G^{(l)}\|_\infty \leq \delta, \ (\forall\, 1 \leq l \leq k). \tag{A.27}$$

Then we have:

$$\sup_{x\in\Omega} \|J[F_k](x) - J[G_k](x)\|_\infty \leq \delta W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1} 3^{2k-2}. \tag{A.28}$$

In particular, when $k = L$, we have:

$$\sup_{x\in\Omega} \Big| \|\nabla F_L(x)\| - \|\nabla G_L(x)\| \Big| \leq \delta W^{\frac{3^{L-1}-1}{2}} (B \vee d)^{\frac{5\cdot 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+1} 3^{2L-2}. \tag{A.29}$$

Proof. We use induction on k to prove the claim.
Base case: When $k = 1$, we have that for any $x \in \Omega$ and any $F_1, G_1 \in \Phi_1(L, W, S, B)$, the following holds:

$$\|J[F_1](x) - J[G_1](x)\|_\infty = \|\mathcal{W}_F^{(1)} - \mathcal{W}_G^{(1)}\|_\infty \leq \delta d \leq 2\delta(B \vee d)^2.$$

Inductive Step: assume that the claim has been proved for $k - 1$, where $2 \leq k \leq L$. Then for any $x \in \Omega$ and $F_k, G_k \in \Phi_k(L, W, S, B)$ satisfying constraint A.27, applying the Chain Rule and

triangle inequality help us upper bound the inf-norm $\|J[F_k](x) - J[G_k](x)\|_\infty$ by:

$$\begin{aligned}
&\|J[F_k](x) - J[G_k](x)\|_\infty = \|\mathcal{W}_F^{(k)} J[\eta_3 \circ F_{k-1}](x) - \mathcal{W}_G^{(k)} J[\eta_3 \circ G_{k-1}](x)\|_\infty \\
&\leq \|\mathcal{W}_F^{(k)} J[\eta_3 \circ F_{k-1}](x) - \mathcal{W}_G^{(k)} J[\eta_3 \circ F_{k-1}](x)\|_\infty + \|\mathcal{W}_G^{(k)} J[\eta_3 \circ F_{k-1}](x) - \mathcal{W}_G^{(k)} J[\eta_3 \circ G_{k-1}](x)\|_\infty \\
&\leq \|\mathcal{W}_F^{(k)} - \mathcal{W}_G^{(k)}\|_\infty \|J[\eta_3 \circ F_{k-1}](x)\|_\infty + \|\mathcal{W}_G^{(k)}\|_\infty \|J[\eta_3 \circ F_{k-1}](x) - J[\eta_3 \circ G_{k-1}](x)\|_\infty \\
&\leq \delta W \|J[\eta_3 \circ F_{k-1}](x)\|_\infty + BW \|J[\eta_3 \circ F_{k-1}](x) - J[\eta_3 \circ G_{k-1}](x)\|_\infty.
\end{aligned} \tag{A.30}$$

Using equation A.25 helps us upper bound the first term by:

$$\delta W \|J[\eta_3 \circ F_{k-1}](x)\|_\infty \leq \delta W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-1}{2}-1} 2^{\frac{3^k-1}{2}-3k+5} 3^{k-1}. \tag{A.31}$$

Note that the two compositions $\eta_3 \circ F_{k-1}$ and $\eta_3 \circ G_{k-1}$ both map from $\mathbb{R}^d$ to $\mathbb{R}^W$. Hence, the two Jacobian matrices $J[\eta_3 \circ F_{k-1}](x)$ and $J[\eta_3 \circ G_{k-1}](x)$ are of shape $\mathbb{R}^{W \times d}$. Applying the Chain Rule again implies:

$$\|J[\eta_3 \circ F_{k-1}](x) - J[\eta_3 \circ G_{k-1}](x)\|_\infty = \sup_{1 \leq l \leq W} \Big(\sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j} - 3\eta_2(G_{k-1,l}(x)) \frac{\partial G_{k-1,l}(x)}{\partial x_j}|\Big).$$

For any $1 \leq l \leq W$, the summation on the RHS above can be upper bounded by:

$$\begin{aligned}
&\sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j} - 3\eta_2(G_{k-1,l}(x)) \frac{\partial G_{k-1,l}(x)}{\partial x_j}| \\
&\leq \sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j} - 3\eta_2(G_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j}| \\
&+ \sum_{j=1}^d |3\eta_2(G_{k-1,l}(x)) \frac{\partial F_{k-1,l}(x)}{\partial x_j} - 3\eta_2(G_{k-1,l}(x)) \frac{\partial G_{k-1,l}(x)}{\partial x_j}| \\
&\leq \sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) - 3\eta_2(G_{k-1,l}(x))| \frac{\partial F_{k-1,l}(x)}{\partial x_j}| + \sum_{j=1}^d |3\eta_2(G_{k-1,l}(x))| \frac{\partial F_{k-1,l}(x)}{\partial x_j} - \frac{\partial G_{k-1,l}(x)}{\partial x_j}|.
\end{aligned}$$

We denote the two summations above by T_1 and T_2, respectively:

$$\begin{aligned}
T_1 &:= \sum_{j=1}^d |3\eta_2(F_{k-1,l}(x)) - 3\eta_2(G_{k-1,l}(x))| \frac{\partial F_{k-1,l}(x)}{\partial x_j}|, \\
T_2 &:= \sum_{j=1}^d |3\eta_2(G_{k-1,l}(x))| \frac{\partial F_{k-1,l}(x)}{\partial x_j} - \frac{\partial G_{k-1,l}(x)}{\partial x_j}|.
\end{aligned}$$

For the first sum T_1, applying Lemma A.9, Lemma A.10, Lemma A.11 and Lemma A.12 yields the following upper bound:

$$\begin{aligned}
T_1 &\leq 6\Big(\sup_{x \in \Omega,\ F_{k-1} \in \Phi_{k-1}(L,W,S,B)} \|F_{k-1}(x)\|_\infty\Big) \|F_{k-1}(x) - G_{k-1}(x)\|_\infty \sum_{j=1}^d |\frac{\partial F_{k-1,l}(x)}{\partial x_j}| \\
&\leq 6\Big(\sup_{x \in \Omega,\ F_{k-1} \in \Phi_{k-1}(L,W,S,B)} \|F_{k-1}(x)\|_\infty\Big) \|F_{k-1}(x) - G_{k-1}(x)\|_\infty \|J[F_{k-1}](x)\|_\infty \\
&\leq 3 \times W^{\frac{3^{k-2}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} \times \delta W^{\frac{3^{k-2}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} \\
&\times W^{\frac{3^{k-2}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} = \delta W^{\frac{3^{k-1}-3}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-3}{2}} 2^{\frac{3^k-3}{2}-3k+6} 3^{2k-3}.
\end{aligned}$$

For the second sum T_2, applying Lemma A.9 and inductive hypothesis yields:

$$\begin{aligned}
T_2 &\leq 3\Big(\sup_{x \in \Omega,\ G_{k-1} \in \Phi_{k-1}(L,W,S,B)} \|G_{k-1}(x)\|_\infty\Big)^2 \|J[F_{k-1}](x) - J[G_{k-1}](x)\|_\infty \\
&\leq 3 \times W^{3^{k-2}-1} (B \vee d)^{5 \cdot 3^{k-2}-1} 2^{3^{k-1}-1-2k+4} \times \delta W^{\frac{3^{k-2}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{2k-4} \\
&= \delta W^{\frac{3^{k-1}-3}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-3}{2}} 2^{\frac{3^k-3}{2}-3k+6} 3^{2k-3}.
\end{aligned}$$

Combining the two upper bounds on T_1 and T_2 yields:

$$\sum_{j=1}^{d} |3\eta_2(F_{k-1,l}(x))\frac{\partial F_{k-1,l}(x)}{\partial x_j} - 3\eta_2(G_{k-1,l}(x))\frac{\partial G_{k-1,l}(x)}{\partial x_j}|$$
$$\leq T_1 + T_2 \leq 2 \times \delta W^{\frac{3^{k-1}-3}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-3}{2}} 2^{\frac{3^k-3}{2}-3k+6} 3^{2k-3}.$$

By taking supremum with respect to $1 \leq l \leq W$ on the LHS yields:

$$BW\|J[\eta_3 \circ F_{k-1}](x) - J[\eta_3 \circ G_{k-1}](x)\|_\infty \leq \delta W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-3k+6} 3^{2k-3}. \tag{A.32}$$

By adding the two upper bounds in A.31 and A.32, we can deduce that:

$$\begin{aligned}\|J[F_k](x) - J[G_k](x)\|_\infty &\leq \delta W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}-1} 2^{\frac{3^k-1}{2}-3k+5} 3^{k-1} \\ &+ \delta W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-3k+6} 3^{2k-3} \\ &\leq \delta W^{\frac{3^{k-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} 2^{\frac{3^k-1}{2}-k+1} 3^{2k-2}.\end{aligned}$$

where the last inequality above follows from $k \geq 2$. Taking supremum with respect to $x \in \Omega$ on the LHS implies the given upper bound also holds for k. By induction, the claim is proved.
In particular, when $k = L$, we have $\nabla F_L(x) = J[F_L](x)^T$ for any $x \in \Omega$. Applying Lemma A.13 then yields:

$$\begin{aligned}\sup_{x\in\Omega}\Big|\|\nabla F_L(x)\| - \|\nabla G_L(x)\|\Big| &= \sup_{x\in\Omega}\Big|\|\nabla J[F_L](x)^T\| - \|\nabla J[G_L](x)^T\|\Big| \\ &\leq \sup_{x\in\Omega}\|J[F_L](x) - J[G_L](x)\|_\infty \\ &\leq \delta W^{\frac{3^{L-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-k+1} 3^{2L-2}.\end{aligned}$$

This finishes our proof of the Lemma. □

Theorem A.5. *(Bounding the DNN Gradient space covering number) Fix some sufficiently large $N \in \mathbb{Z}^+$. Consider a Deep Neural Network space $\Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$. Then the* log *value of the covering number of the DNN Gradient space with respect to the $\|\cdot\|_\infty$ norm $\|F(x)\|_\infty := \sup_{x\in\Omega}|F(x)|$, which is denoted by $\mathcal{N}(\delta, \nabla\Phi(L, W, S, B), \|\cdot\|_\infty)$, can be upper bounded by:*

$$\log \mathcal{N}(\delta, \nabla\Phi(L, W, S, B), \|\cdot\|_\infty) = O\left(S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]\right). \tag{A.33}$$

Proof. We firstly fix a sparsity pattern (i.e, the locations of the non-zero entries are fixed). Using equation A.29 in Lemma A.14, yields the following upper bound on the covering number with respect to $\|\cdot\|_\infty$:

$$\Big(\frac{\delta}{W^{\frac{3^{L-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+1} 3^{2L-2}}\Big)^{-S}.$$

Furthermore, note that the number of feasible configurations is upper bounded by: $\binom{(W+1)^L}{S} \leq (W+1)^{LS}$.(Schmidt-Hieber et al., 2020; Farrell et al., 2021). Plug this inequality into the previous estimation then yields:

$$\begin{aligned}\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty) &\leq \log\left[(W+1)^{LS}\Big(\frac{\delta}{W^{\frac{3^{L-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+1} 3^{2L-2}}\Big)^{-S}\right] \\ &\leq S\log\left[\delta^{-1}(W+1)^L 3^{2L-2} W^{\frac{3^{L-1}-1}{2}}(B \vee d)^{\frac{5\cdot 3^{L-1}-1}{2}} 2^{\frac{3^L-1}{2}-L+1}\right] \\ &\lesssim S\Big[\log(\delta^{-1}) + 2L\log(3W) + 3^L\log(W(B \vee d)) + 3^L\log 2\Big].\end{aligned}$$

Note that here the dimension d is some constant. Thus, by plugging in thee given magnitudes $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$, we can further deduce that:

$$\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty) \lesssim S\Big[\log(\delta^{-1}) + 3^L\log(WB)\Big].$$

This finishes our proof. □

Now let's consider upper bounding the covering number of the Laplacian of the sparse Deep Neural Networks. Note that for any $1 \leq k \leq L-1$, any $k-$ReLU3 Deep Neural Network $F_k \in \Phi_k(L, W, S, B)$ is a vector-valued function mapping from $\mathbb{R}^d$ to $\mathbb{R}^W$. Moreover, we define the Laplacian of $F_k(x)$, which is denoted by $\Delta[F_k](x)$, as follows:

$$\Delta[F_k](x) = [\Delta F_{k,1}(x), \Delta F_{k,2}(x), \cdots, \Delta F_{k,W}(x)]^T \in \mathbb{R}^W,$$

where for any $1 \leq l \leq W$, we have:

$$\Delta F_{k,l}(x) = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} F_{k,l}(x).$$

In particular, when $k = L$, we have that any $F_L \in \Phi_L(L, W, S, B) = \Phi(L, W, S, B)$ is a scalar-valued function mapping from $\mathbb{R}^d$ to $\mathbb{R}$. Thus, its Laplacian can be explicitly written as:

$$\Delta[F_L](x) = \Delta F_L(x) = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} F_L(x).$$

For both Lemma A.15 and Lemma A.16 below, we consider a fixed Deep Neural Network space $\Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$, where $N \in \mathbb{Z}^+$ is fixed and sufficiently large.

Lemma A.15. *(Upper bound on ∞-norm of Laplacian of elements in the DNN space) For any $1 \leq k \leq L$, we have the following upper bound:*

$$\sup_{x \in \Omega, F_k \in \Phi_k(L,W,S,B)} \|\Delta[F_k](x)\|_\infty = O\Big(W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-1}{2}}\Big).$$

Proof. We use induction on k to prove the claim.
Base case: $k = 1$. Note that any $F_1 \in \Phi_1(L, W, S, B)$ is a linear transform, so the Laplacian $\Delta[F_1](x)$ must be the zero vector for any $x \in \Omega$. This implies:

$$\|\Delta[F_1](x)\|_\infty = 0 \lesssim (B \vee d)^2.$$

Inductive Step: Assume that the claim has been proved for $k-1$, where $2 \leq k \leq L$. For any $x \in \Omega$ and any $F_k \in \Phi_k(L, W, S, B)$, using linearity of the Laplacian operator implies:

$$\Delta[F_k](x) = \mathcal{W}_F^{(k)} \Delta[\eta_3 \circ F_{k-1}](x).$$

Taking the inf-norm on both sides of the identity above implies:

$$\|\Delta[F_k](x)\|_\infty \leq \|\mathcal{W}_F^{(k)}\|_\infty \|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty \leq WB \|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty.$$

It now remains to upper bound the term $\|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty$. For any $1 \leq l \leq W$, we will use the Chain Rule to write the l-th component $\Big(\Delta[\eta_3 \circ F_{k-1}](x)\Big)_l$ in an explicit form. For any $1 \leq j \leq d$, we have:

$$\frac{\partial}{\partial x_j} \eta_3[F_{k-1,l}(x)] = 3\eta_2[F_{k-1,l}(x)] \frac{\partial}{\partial x_j} F_{k-1,l}(x).$$

Differentiating with respect to x_j on both sides above yields:

$$\frac{\partial^2}{\partial x_j^2} \eta_3[F_{k-1,l}(x)] = 6\eta_1[F_{k-1,l}(x)] \Big(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\Big)^2 + 3\eta_2[F_{k-1,l}(x)] \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x). \quad \text{(A.34)}$$

Summing the expression above from $j = 1$ to $j = d$ implies:

$$\begin{aligned}
\Big|\Big(\Delta[\eta_3 \circ F_{k-1}](x)\Big)_l\Big| &= \left|\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} \eta_3[F_{k-1,l}(x)]\right| \\
&= \left|6\eta_1[F_{k-1,l}(x)] \sum_{j=1}^{d} \Big(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\Big)^2 + 3\eta_2[F_{k-1,l}(x)] \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x)\right| \\
&\leq 6\Big|\eta_1[F_{k-1,l}(x)]\Big| \left(\sum_{j=1}^{d} \Big|\frac{\partial}{\partial x_j} F_{k-1,l}(x)\Big|\right)^2 + 3\Big|\eta_2[F_{k-1,l}(x)]\Big| \left|\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x)\right|.
\end{aligned}$$

We denote the two summations above by U_1 and U_2, respectively:

$$U_1 := 6\Big|\eta_1[F_{k-1,l}(x)]\Big| \left(\sum_{j=1}^{d} \Big|\frac{\partial}{\partial x_j} F_{k-1,l}(x)\Big|\right)^2.$$

$$U_2 := 3\Big|\eta_2[F_{k-1,l}(x)]\Big| \left|\sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x)\right|.$$

On the one hand, by applying Lemma A.9 and Lemma A.12, we can upper bound U_1 by:

$$\begin{aligned} U_1 &\leq 6\Big(\sup_{x\in\Omega,\ F_{k-1}\in\Phi_{k-1}(L,W,S,B)} \|F_{k-1}(x)\|_\infty\Big)\Big(\sup_{x\in\Omega, F_{k-1}\in\Phi_{k-1}(L,W,S,B)} \|J[F_{k-1}](x)\|_\infty\Big)^2 \\ &\leq 6\times W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} \times W^{3^{k-2}-1}(B\vee d)^{5\cdot 3^{k-2}-1} 2^{3^{k-1}-1-2k+4} 3^{2k-4} \\ &\lesssim W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}, \end{aligned}$$

where the last step above follows from $k \leq L$ and $L = O(1)$.
On the other hand, by applying Lemma A.9 and the inductive hypothesis, we have:

$$\begin{aligned} U_2 &\leq 3\Big(\sup_{x\in\Omega,\ F_{k-1}\in\Phi_{k-1}(L,W,S,B)} \|F_{k-1}(x)\|_\infty\Big)^2 \|\Delta[F_{k-1}](x)\|_\infty \\ &\lesssim 3\times W^{3^{k-2}-1}(B\vee d)^{5\cdot 3^{k-2}-1} 2^{3^{k-1}-1-2k+4} \times W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} \\ &\lesssim W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}, \end{aligned}$$

where the last step above follows from $k \leq L$ and $L = O(1)$.
Summing the two bounds on U_1 and U_2 implies that for any $1 \leq l \leq W$, we have:

$$\Big|\Big(\Delta[\eta_3 \circ F_{k-1}](x)\Big)_l\Big| \leq U_1 + U_2 \lesssim W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}. \tag{A.35}$$

Taking supremum with respect to $1 \leq l \leq W$ then yields:

$$\|\Delta[F_k](x)\|_\infty \leq WB\|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty \lesssim W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-1}{2}}.$$

Taking supremum with respect to $x \in \Omega$ and $F_k \in \Phi_k(L,W,S,B)$ on the LHS implies that the given upper bound also holds for k. By induction, the claim is proved. □

Lemma A.16. *(Relation between the covering number of the DNN Laplacian space and parameter space) For any $1 \leq k \leq L$, suppose that a pair of different two networks $F_k, G_k \in \Phi_k(L,W,S,B)$ are given by:*

$$F_k(x) := (\mathcal{W}_F^{(k)}\eta_3(\cdot) + b_F^{(k)})\cdots(\mathcal{W}_F^{(1)}x + b_F^{(1)}),$$
$$G_k(x) := (\mathcal{W}_G^{(k)}\eta_3(\cdot) + b_G^{(k)})\cdots(\mathcal{W}_G^{(1)}x + b_G^{(1)}).$$

Furthermore, assume that the $\|\ \|_\infty$ norm of the distance between the parameter spaces is uniformly upper bounded by δ, i.e

$$\|W_F^{(l)} - W_G^{(l)}\|_{\infty,\infty} \leq \delta,\ \|b_F^{(l)} - b_G^{(l)}\|_\infty \leq \delta,\ (\forall\, 1 \leq l \leq k). \tag{A.36}$$

Then we have:

$$\sup_{x\in\Omega} \|\Delta[F_k](x) - \Delta[G_k](x)\|_\infty = O\Big(\delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-1}{2}}\Big). \tag{A.37}$$

Proof. We use induction on k to prove the claim.
Base case: $k = 1$. Note that any $F_1 \in \Phi_1(L,W,S,B)$ is a linear transform, so the Laplacian $\Delta[F_1](x)$ must be the zero vector for any $x \in \Omega$. Hence, for any $x \in \Omega$ and any $F_1, G_1 \in \Phi_1(L,W,S,B)$, we have:

$$\|\Delta[F_1](x) - \Delta[G_1](x)\|_\infty = 0 \lesssim \delta(B\vee d)^2.$$

Inductive Step: assume that the claim has been proved for $k-1$, where $2 \leq k \leq L$. Then for any $x \in \Omega$ and $F_k, G_k \in \Phi_k(L, W, S, B)$ satisfying constraint A.36, applying linearity of the Laplacian operator indicates:

$$\begin{aligned}
\|\Delta[F_k](x) - \Delta[G_k](x)\|_\infty &= \|\mathcal{W}_F^{(k)}\Delta[\eta_3 \circ F_{k-1}](x) - \mathcal{W}_G^{(k)}\Delta[\eta_3 \circ G_{k-1}](x)\|_\infty \\
&= \left\|\left(\mathcal{W}_F^{(k)} - \mathcal{W}_G^{(k)}\right)\Delta[\eta_3 \circ F_{k-1}](x)\right\|_\infty \\
&+ \left\|\mathcal{W}_G^{(k)}\left(\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\right)\right\|_\infty \\
&\leq \|\mathcal{W}_F^{(k)} - \mathcal{W}_G^{(k)}\|_\infty \|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty \\
&+ \|\mathcal{W}_G^{(k)}\|_\infty \|\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\|_\infty.
\end{aligned}$$

For the first term $\|\mathcal{W}_F^{(k)} - \mathcal{W}_G^{(k)}\|_\infty \|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty$, applying the bound in equation A.35 and equation A.36 yields:

$$\begin{aligned}
\|\mathcal{W}_F^{(k)} - \mathcal{W}_G^{(k)}\|_\infty \|\Delta[\eta_3 \circ F_{k-1}](x)\|_\infty &\lesssim \delta W \times W^{\frac{3^{k-1}-3}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-3}{2}} \\
&= \delta W^{\frac{3^{k-1}-1}{2}} (B \vee d)^{\frac{5 \cdot 3^{k-1}-3}{2}}.
\end{aligned} \tag{A.38}$$

For the second term $\|\mathcal{W}_G^{(k)}\|_\infty \|\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\|_\infty$, we need to upper bound the norm $\|\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\|_\infty$ at first. Note that for any $1 \leq l \leq W$, we can use equation A.34 to write the l-th component of $\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)$ as:

$$\begin{aligned}
\left(\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\right)_l &= \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}\eta_3[F_{k-1,l}(x)] - \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2}\eta_3[G_{k-1,l}(x)] \\
&= 6\eta_1[F_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2 - 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} G_{k-1,l}(x)\right)^2 \\
&+ 3\eta_2[F_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x) - 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} G_{k-1,l}(x) \\
&= 6\eta_1[F_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2 - 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2 \\
&+ 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2 - 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} G_{k-1,l}(x)\right)^2 \\
&+ 3\eta_2[F_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x) - 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x) \\
&+ 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x) - 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} G_{k-1,l}(x).
\end{aligned}$$

We denote the four summations above by V_1, V_2, V_3 and V_4, respectively:

$$\begin{aligned}
V_1 &:= 6\eta_1[F_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2 - 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2, \\
V_2 &:= 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} F_{k-1,l}(x)\right)^2 - 6\eta_1[G_{k-1,l}(x)] \sum_{j=1}^d \left(\frac{\partial}{\partial x_j} G_{k-1,l}(x)\right)^2, \\
V_3 &:= 3\eta_2[F_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x) - 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x), \\
V_4 &:= 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} F_{k-1,l}(x) - 3\eta_2[G_{k-1,l}(x)] \sum_{j=1}^d \frac{\partial^2}{\partial x_j^2} G_{k-1,l}(x).
\end{aligned}$$

By applying Lemma A.9, Lemma A.10, Lemma A.11 and Lemma A.12, we can upper bound V_1 by:

$$
\begin{aligned}
V_1 &= 6\Big(\eta_1[F_{k-1,l}(x)] - \eta_1[G_{k-1,l}(x)]\Big)\sum_{j=1}^{d}\Big(\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big)^2 \\
&\leq 6|F_{k-1,l}(x) - G_{k-1,l}(x)|\left(\sum_{j=1}^{d}\Big|\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big|\right)^2 \leq 6\|F_{k-1}(x) - G_{k-1}(x)\|_\infty \|J[F_{k-1}](x)\|_\infty^2 \\
&\lesssim \delta W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} \times W^{3^{k-2}-1}(B\vee d)^{5\cdot 3^{k-2}-1} 2^{3^{k-1}-1-2k+4} 3^{2k-4} \\
&\lesssim \delta W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}.
\end{aligned}
$$

where the last step above follows from $k \leq L$ and $L = O(1)$.

Furthermore, note that for any $1 \leq j \leq d$, we can upper bound the difference $\Big(\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big)^2 - \Big(\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big)^2$ as follows:

$$
\begin{aligned}
&\Big(\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big)^2 - \Big(\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big)^2 \leq \left|\Big(\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big)^2 - \Big(\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big)^2\right| \\
&= \left|\frac{\partial}{\partial x_j}F_{k-1,l}(x) + \frac{\partial}{\partial x_j}G_{k-1,l}(x)\right|\left|\frac{\partial}{\partial x_j}F_{k-1,l}(x) - \frac{\partial}{\partial x_j}G_{k-1,l}(x)\right| \qquad \text{(A.39)} \\
&\leq \left(\Big|\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big| + \Big|\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big|\right)\left|\frac{\partial}{\partial x_j}F_{k-1,l}(x) - \frac{\partial}{\partial x_j}G_{k-1,l}(x)\right|.
\end{aligned}
$$

Note that $\eta_1(G_{k-1,l}(x)) \geq 0$. Combining the non-negativity with equation A.39, Lemma A.9, Lemma A.12 and Lemma A.14 helps us upper bound V_2 by:

$$
\begin{aligned}
V_2 &= 6\eta_1[G_{k-1,l}(x)]\sum_{j=1}^{d}\left[\Big(\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big)^2 - \Big(\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big)^2\right] \\
&\leq 6\|G_{k-1}(x)\|_\infty \sum_{j=1}^{d}\left(\Big|\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big| + \Big|\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big|\right)\left|\frac{\partial}{\partial x_j}F_{k-1,l}(x) - \frac{\partial}{\partial x_j}G_{k-1,l}(x)\right| \\
&\leq 6\|G_{k-1}(x)\|_\infty \left(\sum_{j=1}^{d}\Big|\frac{\partial}{\partial x_j}F_{k-1,l}(x)\Big| + \sum_{j=1}^{d}\Big|\frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big|\right)\left(\sum_{j=1}^{d}\Big|\frac{\partial}{\partial x_j}F_{k-1,l}(x) - \frac{\partial}{\partial x_j}G_{k-1,l}(x)\Big|\right) \\
&\leq 6\|G_{k-1}(x)\|_\infty \Big(\|J[F_{k-1}](x)\|_\infty + \|J[G_{k-1}](x)\|_\infty\Big)\Big\|J[F_{k-1}](x) - J[G_{k-1}](x)\Big\|_\infty \\
&\leq 6W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} \times 2W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} \\
&\times \delta W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+1} 3^{2k-4} \lesssim \delta W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}.
\end{aligned}
$$

where the last step above follows from $k \leq L$ and $L = O(1)$.

Moreover, using Lemma A.9, Lemma A.10 and Lemma A.15 helps us upper bound V_3 by:

$$
\begin{aligned}
V_3 &= \Big(3\eta_2[F_{k-1,l}(x)] - 3\eta_2[G_{k-1,l}(x)]\Big)\sum_{j=1}^{d}\frac{\partial^2}{\partial x_j^2}F_{k-1,l}(x) \\
&\leq \Big|3\eta_2[F_{k-1,l}(x)] - 3\eta_2[G_{k-1,l}(x)]\Big|\Big|\sum_{j=1}^{d}\frac{\partial^2}{\partial x_j^2}F_{k-1,l}(x)\Big| \\
&\leq 6\Big(\sup_{x\in\Omega,\ F_{k-1}\in\Phi_{k-1}(L,W,S,B)}\|F_{k-1}(x)\|_\infty\Big)\|F_{k-1}(x) - G_{k-1}(x)\|_\infty \|\Delta[F_{k-1}](x)\|_\infty \\
&\lesssim 6W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} \times \delta W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} 2^{\frac{3^{k-1}-1}{2}-k+2} 3^{k-2} \\
&\times W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} \lesssim \delta W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}
\end{aligned}
$$

where the last step above follows from $k \leq L$ and $L = O(1)$.
Finally, applying Lemma A.9 and inductive hypothesis helps us upper bound V_4 by:

$$\begin{aligned} V_4 &= 3\eta_2[G_{k-1,l}(x)]\Big(\sum_{j=1}^{d}\frac{\partial^2}{\partial x_j^2}F_{k-1,l}(x) - \sum_{j=1}^{d}\frac{\partial^2}{\partial x_j^2}G_{k-1,l}(x)\Big) \\ &\leq 3\|G_{k-1}(x)\|_\infty^2\|\Delta[F_{k-1}](x) - \Delta[G_{k-1}](x)\|_\infty \\ &\lesssim 3W^{3^{k-2}-1}(B\vee d)^{5\cdot 3^{k-2}-1}2^{3^{k-1}-1-2k+4} \times \delta W^{\frac{3^{k-2}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-2}-1}{2}} \\ &\lesssim \delta W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}} \end{aligned}$$

where the last step above follows from $k \leq L$ and $L = O(1)$.
Combining the four bounds on V_1, V_2, V_3 and V_4 implies:

$$\Big(\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\Big)_l = \sum_{i=1}^{4} V_i \lesssim \delta W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}}$$

Taking supremum with respect to $1 \leq l \leq W$ gives us an upper bound on the second term $\|\mathcal{W}_G^{(k)}\|_\infty\|\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\|_\infty$:

$$\begin{aligned} \|\mathcal{W}_G^{(k)}\|_\infty\|\Delta[\eta_3 \circ F_{k-1}](x) - \Delta[\eta_3 \circ G_{k-1}](x)\|_\infty &\lesssim WB \times \delta W^{\frac{3^{k-1}-3}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}} \\ &= \delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} \end{aligned} \tag{A.40}$$

Combining the two bounds derived in equation A.38 and equation A.40 then implies:

$$\begin{aligned} \|\Delta[F_k](x) - \Delta[G_k](x)\|_\infty &\lesssim \delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-3}{2}} + \delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} \\ &\lesssim \delta W^{\frac{3^{k-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{k-1}-1}{2}} \end{aligned}$$

Taking supremum with respect to $x \in \Omega$ on the LHS implies that the given upper bound also holds for k. By induction, the claim is proved. □

Given a Neural Network function space $\Phi(L, W, S, B)$, we define a corresponding Neural Network Laplacian space $\Delta\Phi(L, W, S, B)$ as:

$$\Delta\Phi(L, W, S, B) := \{\Delta F \mid F \in \Phi(L, W, S, B)\}. \tag{A.41}$$

Theorem A.6. *(Bounding the Neural Network Laplacian space covering number) Fix some sufficiently large $N \in \mathbb{Z}^+$. Consider a Deep Neural Network space $\Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$. Then the $\log$ value of the covering number of the DNN Laplacian space with respect to the $\|\cdot\|_\infty$ norm $\|F(x)\|_\infty := \sup_{x\in\Omega}|F(x)|$, which is denoted by $\mathcal{N}(\delta, \Delta\Phi(L, W, S, B), \|\cdot\|_\infty)$, can be upper bounded by:*

$$\log\mathcal{N}(\delta, \Delta\Phi(L, W, S, B), \|\cdot\|_\infty) = O\left(S\Big[\log(\delta^{-1}) + 3^L\log(WB)\Big]\right) \tag{A.42}$$

Proof. We firstly fix a sparsity pattern (i.e, the locations of the non-zero entries are fixed). Applying Lemma A.16 yields that there exists some constant $C = O(1)$, such that the covering number with respect to $\|\cdot\|_\infty$ can be upper bounded by:

$$\Big(\frac{\delta}{CW^{\frac{3^{L-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{L-1}-1}{2}}}\Big)^{-S}$$

Furthermore, note that the number of feasible configurations is upper bounded by $\binom{(W+1)^L}{S} \leq (W+1)^{LS}$(Schmidt-Hieber et al., 2020; Farrell et al., 2021). Then we plug this into the pervious estimation and yields:

$$\begin{aligned} \log\mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty) &\leq \log\left[(W+1)^{LS}\Big(\frac{\delta}{CW^{\frac{3^{L-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{L-1}-1}{2}}}\Big)^{-S}\right] \\ &\leq S\log\left[\delta^{-1}(W+1)^L W^{\frac{3^{L-1}-1}{2}}(B\vee d)^{\frac{5\cdot 3^{L-1}-1}{2}}\right] \\ &\lesssim S\Big[\log(\delta^{-1}) + L\log(W) + 3^L\log(W(B\vee d))\Big] \end{aligned}$$

Note that here the dimension d is some constant. Thus, by plugging in thee given magnitudes $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$, we can further deduce that:

$$\log \mathcal{N}(\delta, \Delta\Phi(L, W, S, B), \|\cdot\|_\infty) \lesssim S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]$$

This finishes our proof. □

Lemma A.17 (Local Rademacher Complexity Bound for Deep Ritz Method)**.** *Consider a Deep Neural Network space $\boldsymbol{F}(\Omega) = \Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$, where $N \in \mathbb{Z}^+$ is fixed to be sufficiently large. Moreover, assume that the gradients and function value of $\boldsymbol{F}(\Omega), V$ and f are uniformly bounded*

$$\max\Big\{\sup_{u\in\boldsymbol{F}(\Omega)} \|u\|_{L^\infty(\Omega)}, \sup_{u\in\boldsymbol{F}(\Omega)} \|\nabla u\|_{L^\infty(\Omega)}, \|u^*\|_{L^\infty(\Omega)}, \|\nabla u^*\|_{L^\infty(\Omega)}, V_{max}, \|f\|_{L^\infty(\Omega)}\Big\} \leq C. \tag{A.43}$$

For any $\rho > 0$, we consider a localized set L_ρ defined by:

$$\boldsymbol{L}_\rho(\Omega) := \{u : u \in \boldsymbol{F}(\Omega), \|u - u^*\|_{H^1}^2 \leq \rho\}.$$

Then for any $\rho \gtrsim n^{-2}$, the Rademacher complexity of a localized function space $\boldsymbol{S}_\rho(\Omega) := \Big\{h := |\Omega| \cdot \Big[\frac{1}{2}\Big(\|\nabla u\|^2 - \|\nabla u^*\|^2\Big) + \frac{1}{2}V(|u|^2 - |u^*|^2) - f(u - u^*)\Big] \;\Big|\; u \in L_\rho(\Omega)\Big\}$ can be upper bounded by a sub-root function

$$\phi(\rho) := O\left(\sqrt{\frac{S3^L\rho}{n}\log(BWn)}\right).$$

i.e. *we have*

$$\phi(4\rho) \leq 2\phi(\rho) \text{ and } R_n(\boldsymbol{S}_\rho(\Omega)) \leq \phi(\rho). \tag{A.44}$$

holds for all $\rho \gtrsim n^{-2}$.

Proof. Firstly, we will check that for any $u \in L_\rho(\Omega)$, the corresponding function h in $\boldsymbol{S}_\rho(\Omega)$ is Lipschitz with respect to $u - u^*$ and $\|\nabla u\| - \|\nabla u^*\|$. Note that for any $u_1, u_2 \in L_\rho(\Omega)$ with corresponding functions $h_1, h_2 \in \boldsymbol{S}_\rho(\Omega)$, applying boundedness condition A.43 yields:

$$\begin{aligned}
|h_1(x) - h_2(x)| &\leq \frac{1}{2}\Big|\|\nabla u_1(x)\|^2 - \|\nabla u_2(x)\|^2\Big| + \frac{1}{2}|V(x)||u_1(x)^2 - u_2(x)^2| + |f(x)||u_1(x) - u_2(x)| \\
&\leq C\Big|\|\nabla u_1(x)\| - \|\nabla u_2(x)\|\Big| + (C^2 + C)|u_1(x) - u_2(x)| \\
&= C\Big|\Big(\|\nabla u_1(x)\| - \|\nabla u^*(x)\|\Big) - \Big(\|\nabla u_2(x)\| - \|\nabla u^*(x)\|\Big)\Big| \\
&+ (C^2 + C)\Big|(u_1(x) - u^*(x)) - (u_2(x) - u^*(x))\Big|.
\end{aligned}$$

Let's pick $L = C^2 + C > C$. Applying the Talagrand Contraction Lemma A.2 helps us upper bound the local Rademacher complexity $R_n(\boldsymbol{S}_\rho(\Omega))$ by

$$\begin{aligned}
R_n(\boldsymbol{S}_\rho(\Omega)) &= \mathbb{E}_x\mathbb{E}_\sigma\left[\sup_{u\in\boldsymbol{L}_\rho(\Omega)} \frac{1}{n}\sum_{i=1}^n \sigma_i\Big[\frac{1}{2}\Big(\|\nabla u\|^2 - \|\nabla u^*\|^2\Big) + \frac{1}{2}V(|u|^2 - |u^*|^2) - f(u - u^*)\Big]\right] \\
&\leq 2L\mathbb{E}_x\mathbb{E}_\sigma\left[\sup_{u\in\boldsymbol{L}_\rho(\Omega)} \frac{1}{n}\sum_{i=1}^n \sigma_i\Big(u(x_i) - u^*(x_i)\Big)\right] \\
&+ 2L\mathbb{E}_{x'}\mathbb{E}_{\sigma'}\left[\sup_{u\in\boldsymbol{L}_\rho(\Omega)} \frac{1}{n}\sum_{i=1}^n \sigma_i'\Big(\|\nabla u(x_i')\| - \|\nabla u^*(x_i')\|\Big)\right] \\
&\lesssim R_n\left(\Big\{u - u_* : u \in \boldsymbol{L}_\rho\Big\}\right) + R_n\left(\Big\{\|\nabla u\| - \|\nabla u^*\| : u \in \boldsymbol{L}_\rho\Big\}\right)
\end{aligned}$$

From the localization constraint $\rho \geq \|u - u^*\|_{H^1(\Omega)}^2 = \|u - u^*\|_{L^2(\Omega)}^2 + \|\nabla u - \nabla u^*\|_{L^2(\Omega)}^2$, we can deduce that

$$\|u - u^*\|_{L^2(\Omega)} \leq \sqrt{\rho} \text{ and } \|\nabla u - \nabla u^*\|_{L^2(\Omega)} \leq \sqrt{\rho} \tag{A.45}$$

Moreover, note that $\Omega \subset [0,1]^d$. Applying triangle inequality yields:

$$\begin{aligned}\Big\| \|\nabla u\| - \|\nabla u^*\| \Big\|^2_{L^2(\Omega)} &= \int_\Omega \Big| \|\nabla u(x)\| - \|\nabla u^*(x)\| \Big|^2 dx \leq \int_\Omega \|\nabla u(x) - \nabla u^*(x)\|^2 dx \\ &= \|\nabla u - \nabla u^*\|^2_{L^2(\Omega)} \leq \rho \Rightarrow \Big\| \|\nabla u\| - \|\nabla u^*\| \Big\|_{L^2(\Omega)} \leq \sqrt{\rho}\end{aligned} \tag{A.46}$$

Using inequality A.45 and inequality A.46, we have:

$$\begin{aligned}R_n(\boldsymbol{S}_\rho(\Omega)) &\lesssim R_n\left(\Big\{u - u_* : u \in \boldsymbol{L}_\rho\Big\}\right) + R_n\left(\Big\{\|\nabla u\| - \|\nabla u^*\| : u \in \boldsymbol{L}_\rho\Big\}\right) \\ &\leq R_n\left(\Big\{u - u^* : u \in \Phi(L,W,S,B), \|u - u^*\|_{L^2(\Omega)} \leq \sqrt{\rho}\Big\}\right) \\ &+ R_n\left(\Big\{\|\nabla u\| - \|\nabla u^*\| : u \in \Phi(L,W,S,B), \Big\|\|\nabla u\| - \|\nabla u^*\|\Big\|_{L^2(\Omega)} \leq \sqrt{\rho}\Big\}\right)\end{aligned}$$

(Geer & van de Geer, 2000; Rakhlin et al., 2017) showed a "upper isometry" property, where the metric $\|\cdot\|_{L_2}$ is equivalent to $\|\cdot\|_{n,2}$ with high probability. Combining this fact with Theorem A.3, we can bound the local Rademacher complexities using Dudley integral:

$$\begin{aligned}R_n(\boldsymbol{S}_\rho(\Omega)) &\lesssim R_n\left(\Big\{u - u^* : u \in \Phi(L,W,S,B), \|u - u^*\|_{L^2(\Omega)} \leq \sqrt{\rho}\Big\}\right) \\ &+ R_n\left(\Big\{\|\nabla u\| - \|\nabla u^*\| : u \in \Phi(L,W,S,B), \Big\|\|\nabla u\| - \|\nabla u^*\|\Big\|_{L^2(\Omega)} \leq \sqrt{\rho}\Big\}\right) \\ &\leq R_n\left(\Big\{u - u^* : u \in \Phi(L,W,S,B), \|u - u^*\|_{n,2} \leq 2\sqrt{\rho}\Big\}\right) \\ &+ R_n\left(\Big\{\|\nabla u\| - \|\nabla u^*\| : u \in \Phi(L,W,S,B), \Big\|\|\nabla u\| - \|\nabla u^*\|\Big\|_{n,2} \leq 2\sqrt{\rho}\Big\}\right) \\ &\lesssim \inf_{0<\alpha<2\sqrt{\rho}} \Big\{4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \Phi(L,W,S,B), \|\cdot\|_n)} d\delta\Big\} \\ &+ \inf_{0<\alpha<2\sqrt{\rho}} \Big\{4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \nabla\Phi(L,W,S,B), \|\cdot\|_{n,2})} d\delta\Big\} \\ &\lesssim \inf_{0<\alpha<2\sqrt{\rho}} \Big\{4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \Phi(L,W,S,B), \|\cdot\|_\infty)} d\delta\Big\} \\ &+ \inf_{0<\alpha<2\sqrt{\rho}} \Big\{4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \nabla\Phi(L,W,S,B), \|\cdot\|_\infty)} d\delta\Big\}\end{aligned}$$

For any $\rho \gtrsim \frac{1}{n^2}$, we pick $\alpha = \frac{1}{n} \lesssim \sqrt{\rho}$ and plug in the upper bounds proved in Theorem A.4 and Theorem A.5, which implies:

$$\begin{aligned}R_n(\boldsymbol{S}_\rho(\Omega)) &\lesssim \frac{1}{n} + \frac{1}{\sqrt{n}} \int_{\frac{1}{n}}^{2\sqrt{\rho}} \sqrt{S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]} d\delta + \frac{1}{\sqrt{n}} \int_{\frac{1}{n}}^{2\sqrt{\rho}} \sqrt{S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]} d\delta \\ &\lesssim \sqrt{\frac{S3^L\rho}{n} \log(BWn)}\end{aligned}$$

□

Lemma A.18 (Local Rademacher Complexity Bound for Physics Informed Neural Network)**.** *Consider a Deep Neural Network space $\boldsymbol{F}(\Omega) = \Phi(L,W,S,B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$, where $N \in \mathbb{Z}^+$ is fixed to be sufficiently large. Moreover, assume that the gradients and function value of $\boldsymbol{F}(\Omega), V$ and f are uniformly bounded*

$$\max\Big\{\sup_{u\in\boldsymbol{F}(\Omega)} \|u\|_{L^\infty(\Omega)}, \sup_{u\in\boldsymbol{F}(\Omega)} \|\Delta u\|_{L^\infty(\Omega)}, \|u^*\|_{L^\infty(\Omega)}, \|\Delta u^*\|_{L^\infty(\Omega)}, V_{max}, \|f\|_{L^\infty(\Omega)}\Big\} \leq C. \tag{A.47}$$

For any $\rho > 0$, we consider a localized set M_ρ defined by:

$$\boldsymbol{M}_\rho(\Omega) := \{u : u \in \boldsymbol{F}(\Omega), \|u - u^*\|^2_{H^2} \leq \rho\}.$$

Then for any $\rho \gtrsim n^{-2}$, the Rademacher complexity of a localized function space $\boldsymbol{T}_\rho(\Omega) := \Big\{ h := |\Omega| \cdot \big[(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2 \big] \;\Big|\; u \in M_\rho(\Omega) \Big\}$ can be upper bounded by a sub-root function

$$\phi(\rho) := O\left(\sqrt{\frac{S3^L \rho}{n} \log(BWn)} \right).$$

i.e. *we have*

$$\phi(4\rho) \leq 2\phi(\rho) \text{ and } R_n(\boldsymbol{T}_\rho(\Omega)) \leq \phi(\rho). \tag{A.48}$$

holds for all $\rho \gtrsim n^{-2}$.

Proof. Firstly, we will check that for any $u \in L_\rho(\Omega)$, the corresponding function h in $\boldsymbol{S}_\rho(\Omega)$ is Lipschitz with respect to $u - u^*$ and $\Delta u - \Delta u^*$. Note that for any $u_1, u_2 \in L_\rho(\Omega)$ with corresponding functions $h_1, h_2 \in \boldsymbol{S}_\rho(\Omega)$, applying boundedness condition A.47 yields:

$$\begin{aligned}
|h_1(x) - h_2(x)| &\leq |\Delta u_1 - \Delta u_2 - V(u_1 - u_2)||\Delta u_1 - Vu_1 + \Delta u_2 - Vu_2 + 2f| \\
&\leq (2C^2 + 4C)\left(|\Delta u_1 - \Delta u_2| + C|u_1 - u_2|\right) \\
&= (2C^2 + 4C)\Big|(\Delta u_1(x) - \Delta u^*(x)) - (\Delta u_2(x) - \Delta u^*(x))\Big| \\
&+ (2C^3 + 4C^2)\Big|(u_1(x) - u^*(x)) - (u_2(x) - u^*(x))\Big|
\end{aligned}$$

Let's pick $L = \max\{2C^2 + 4C, 2C^3 + 4C^2\}$. Applying the Talagrand Contraction Lemma A.2 helps us upper bound the local Rademacher complexity $R_n(\boldsymbol{T}_\rho(\Omega))$ by

$$\begin{aligned}
R_n(\boldsymbol{T}_\rho(\Omega)) &= \mathbb{E}_x \mathbb{E}_\sigma \Big[\sup_{u \in \boldsymbol{M}_\rho(\Omega)} \frac{1}{n} \sum_{i=1}^n \sigma_i \left[(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2 \right] \Big] \\
&\leq 2L \mathbb{E}_x \mathbb{E}_\sigma \left[\sup_{u \in \boldsymbol{M}_\rho(\Omega)} \frac{1}{n} \sum_{i=1}^n \sigma_i \Big(u(x_i) - u^*(x_i) \Big) \right] \\
&+ 2L \mathbb{E}_{x'} \mathbb{E}_{\sigma'} \left[\sup_{u \in \boldsymbol{M}_\rho(\Omega)} \frac{1}{n} \sum_{i=1}^n \sigma_i' \Big(\Delta u(x_i') - \Delta u^*(x_i') \Big) \right] \\
&\lesssim R_n \left(\Big\{ u - u_* : u \in \boldsymbol{M}_\rho \Big\} \right) + R_n \left(\Big\{ \Delta u - \Delta u^* : u \in \boldsymbol{M}_\rho \Big\} \right) \\
&\lesssim R_n \left(\Big\{ u - u^* : u \in \Phi(L, W, S, B), \|u - u^*\|_{L^2(\Omega)} \leq \sqrt{\rho} \Big\} \right) \\
&+ R_n \left(\Big\{ \Delta u - \Delta u^* : u \in \Phi(L, W, S, B), \|\Delta u - \Delta u^*\|_{L^2(\Omega)} \leq \sqrt{\rho} \Big\} \right) \\
&\leq R_n \left(\Big\{ u - u^* : u \in \Phi(L, W, S, B), \|u - u^*\|_{n,2} \leq 2\sqrt{\rho} \Big\} \right) \\
&+ R_n \left(\Big\{ \Delta u - \Delta u^* : u \in \Phi(L, W, S, B), \|\Delta u - \Delta u^*\|_{n,2} \leq 2\sqrt{\rho} \Big\} \right) \\
&\lesssim \inf_{0 < \alpha < 2\sqrt{\rho}} \Big\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_n)} d\delta \Big\} \\
&+ \inf_{0 < \alpha < 2\sqrt{\rho}} \Big\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \Delta\Phi(L, W, S, B), \|\cdot\|_{n,2})} d\delta \Big\} \\
&\lesssim \inf_{0 < \alpha < 2\sqrt{\rho}} \Big\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \Phi(L, W, S, B), \|\cdot\|_\infty)} d\delta \Big\} \\
&+ \inf_{0 < \alpha < 2\sqrt{\rho}} \Big\{ 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^{2\sqrt{\rho}} \sqrt{\log \mathcal{N}(\delta, \Delta\Phi(L, W, S, B), \|\cdot\|_\infty)} d\delta \Big\}
\end{aligned}$$

For any $\rho \gtrsim \frac{1}{n^2}$, we pick $\alpha = \frac{1}{n} \lesssim \sqrt{\rho}$ and plug in the upper bounds proved in Theorem A.4 and Theorem A.5, which implies:

$$R_n(\boldsymbol{T}_\rho(\Omega)) \lesssim \frac{1}{n} + \frac{1}{\sqrt{n}} \int_{\frac{1}{n}}^{2\sqrt{\rho}} \sqrt{S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]} d\delta + \frac{1}{\sqrt{n}} \int_{\frac{1}{n}}^{2\sqrt{\rho}} \sqrt{S\Big[\log(\delta^{-1}) + 3^L \log(WB)\Big]} d\delta$$
$$\lesssim \sqrt{\frac{S3^L \rho}{n} \log(BWn)}$$

□

A.4 Auxiliary definitions and lemmata On Approximation Error

A.4.1 Approximation using Truncated Fourier Basis

Lemma A.19. *Given $\alpha > 0$ and a fixed integer $\xi \in \mathbb{Z}^+$. For any function $f \in H^\alpha(\Omega)$, we let $f_\xi = \sum_{\|z\|_\infty \le \xi} f_z \phi_z$ be the best approximation of f in the space $F_\xi(\Omega)$. Then for any $0 < \beta \le \alpha$, we have the following inequality:*

$$\|f - f_\xi\|^2_{H^\beta(\Omega)} \le \xi^{-2(\alpha-\beta)} \|f\|^2_{H^\alpha}.$$

Proof. For $f \in H^\alpha(\Omega)$, we know the Fourier coefficient satisfies

$$\sum_{\|z\|_\infty \ge \xi} |f_z|^2 \|z\|^{2\alpha} \lesssim \|f\|^2_{H^\alpha}.$$

We directly construct $f_\xi = \sum_{\|z\|_\infty \le \xi} f_z \phi_z$ to be the truncated Fourier series of the function f, then we have

$$\|f - f_\xi\|^2_{H^\beta(\Omega)} \lesssim \sum_{\|z\|_\infty \ge \xi} |f_z|^2 \|z\|^{2\beta} \le \xi^{-2(\alpha-\beta)} \sum_{\|z\|_\infty \ge \xi} |f_z|^2 \|z\|^{2\alpha} \le \xi^{-2(\alpha-\beta)} \|f\|^2_{H^\alpha}.$$

□

A.4.2 Approximation using Neural Network

In this section, we aim to provide approximation bound for deep neural network. Our proof of the approximation upper bound is based on the observation that the B-spline approximation(De Boor & De Boor, 1978; Schumaker, 2007) can be formulated as a ReLU3 neural network efficiently(Suzuki, 2018; Gühring et al., 2020; Duan et al., 2021; Jiao et al., 2021a). Although the proof of the approximation of the neural network to the Sobolev spaces is a standard approach, we still demonstrate the proof sketch here.

Definition A.4. *(Univariate and Multivariate B-splines) Fix an arbitrary integer $l \in \mathbb{Z}^+$. Consider a corresponding uniform partition π_l of $[0, 1]$:*

$$\pi_l : 0 = t_0^{(l)} < t_1^{(l)} < \cdots < t_{l-1}^{(l)} < t_l^{(l)} = 1,$$

where $t_i^{(l)} = \frac{i}{l}$ ($\forall\, 0 \le i \le l$). Now for any $k \in \mathbb{Z}^+$, we can define an extended partition $\pi_{l,k}$ as:

$$\pi_{l,k} : t_{-k+1}^{(l)} = \cdots t_{-1}^{(l)} = 0 = t_0^{(l)} < t_1^{(l)} < \cdots < t_{l-1}^{(l)} < t_l^{(l)} = 1 = t_{l+1}^{(l)} = \cdots = t_{l+k-1}^{(l)}$$

Based on the extended partition $\pi_{l,k}$, the univariate B-splines of order k with respect to partition π_l are defined by:

$$N_{l,i}^{(k)}(x) := (-1)^k (t_{i+k}^{(l)} - t_i^{(l)}) \cdot \Big[t_i^{(l)}, \cdots, t_{i+k}^{(l)}\Big] \max\{(x-t), 0\}^{k-1},\ x \in [0,1],\ i \in I_{l,k} \quad \text{(A.49)}$$

where $I_{l,k} = \{-k+1, -k+2, \cdots, l-1\}$ and $\Big[t_i^{(l)}, \cdots, t_{i+k}^{(l)}\Big]$ denotes the divided difference operator.

Equivalently, for any $x \in [0,1]$, we can rewrite the univariate B-splines $N_{l,i}^{(k)}(x)$ in an explicit form:

$$N_{l,i}^{(k)}(x) = \begin{cases} \frac{l^{k-1}}{(k-1)!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \max\left\{x - \frac{i+j}{l}, 0\right\}^{k-1}, \ (0 \le i \le l-k+1) \\ \sum_{j=0}^{k-1} a_{ij} \max\left\{x - \frac{j}{l}, 0\right\}^{k-1} + \sum_{n=1}^{k-2} b_{in} x^n + b_{i0}, \ (-k+1 \le i \le 0) \\ \sum_{j=l-k+1}^{l} c_{ij} \max\left\{x - \frac{j}{l}, 0\right\}^{k-1}, \ (l-k+1 \le i \le l-1) \end{cases} \tag{A.50}$$

where $\{a_{ij} \mid -k+1 \le i \le 0, \ 0 \le j \le k-1\}$, $\{b_{in} \mid -k+1 \le i \le 0, \ 1 \le n \le k-2\}$ and $\{c_{ij} \mid l-k+1 \le i \le l-1, \ l-k+1 \le j \le l-1\}$ are some fixed constants.
For any index vector $\boldsymbol{i} = (i_1, i_2, \cdots, i_d) \in I_{l,k}^d$, we can define a corresponding multivariate B-spline as a product of univariate B-splines:

$$N_{l,\boldsymbol{i}}^{(k)}(\boldsymbol{x}) := \Pi_{j=1}^d N_{l,i_j}^{(k)}(x_j). \tag{A.51}$$

Definition A.5. *(Interpolation Operator(Schumaker, 2007)) Take some domain $\Omega \subset [0,1]^d$ and two arbitrary integers $k, l \in \mathbb{Z}^+$. Consider the extended partition $\pi_{l,k}$ and the corresponding set of multivariate B-splines $\{N_{l,\boldsymbol{i}}^{(k)}(x)\}_{\boldsymbol{i} \in I_{l,k}^d}$ defined in Definition A.4. For any $\boldsymbol{i} \in I_{l,k}^d$, we define the domain $\Omega_{\boldsymbol{i}} := \{\boldsymbol{x} \in \Omega : x_j \in [t_{i_j}, t_{i_j+k}], \ 1 \le j \le d\}$. There exists a set of linear functionals $\{\lambda_{\boldsymbol{i}}\}_{\boldsymbol{i} \in I_{k,l}^d}$, where $\lambda_{\boldsymbol{i}} : L^1(\Omega) \to \mathbb{R}$ ($\forall \ \boldsymbol{i} \in I_{k,l}^d$), such that for any $\boldsymbol{i} \in I_{k,l}^d$ and $p \in [1, \infty]$, we have:*

$$\lambda_{\boldsymbol{i}}(N_{l,\boldsymbol{j}}^{(k)}) = \delta_{\boldsymbol{i},\boldsymbol{j}} \text{ and } |\lambda_{\boldsymbol{i}}(f)| \le 9^{d(k-1)}(2k+1)^d \left(\frac{k}{l}\right)^{-\frac{d}{p}} \|f\|_{L^p(\Omega_{\boldsymbol{i}})}, \ \forall \ f \in L^p(\Omega). \tag{A.52}$$

The corresponding interpolation operator $Q_{k,l}$ is defined as:

$$Q_{k,l} f := \sum_{\boldsymbol{i} \in I_{k,l}^d} \lambda_i(f) N_{l,\boldsymbol{i}}^{(k)}, \ \forall \ f \in L^1(\Omega).$$

Theorem A.7. *[(Schumaker, 2007)] Fix $f \in W^s(\Omega)$ with $\Omega \subseteq [0,1]^d, s \in \mathbb{Z}^+$ and $p \in [1, \infty)$. Then for any $k, l, r \in \mathbb{Z}^+$ with $k \ge s$ and $0 \le r \le s$, we have that there exists some constant $C = C(k, s, r, p, d)$, such that:*

$$\|f - Q_{k,l} f\|_{H^r(\Omega)} \le C \left(\frac{1}{l}\right)^{s-r} \|f\|_{H^s(\Omega)}.$$

Theorem A.8. *(Approximation result of Deep Neural Network) Fix some dimension $d \in \mathbb{Z}^+$, some domain $\Omega \subseteq [0,1]^d$. We pick some $l = N^{\frac{1}{d}} \ge 2$, for any $s, r \in \mathbb{Z}^+$ with $0 \le r \le s$ and any function $u^* \in H^s(\Omega)$, there exists some sparse Deep Neural Network $u_{DNN} \in \Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N), B = O(N)$, such that:*

$$\|u_{DNN} - u^*\|_{H^r(\Omega)} \lesssim N^{-\frac{s-r}{d}} \|u^*\|_{H^s(\Omega)}. \tag{A.53}$$

Proof. We firstly show that the given function u^* can be approximated well by some linear combination of multivariate splines, which is denoted by u_{sp}. Note that N is assumed to be sufficiently large. Hence, we may pick $l = \lceil N^{\frac{1}{d}} \rceil = \Theta(N^{\frac{1}{d}}) \in \mathbb{Z}^+$ to be the partition size of the B-splines. Moreover, by picking $k = 4$ and $p = 2$ in Theorem A.7, we have that the linear combination $u_{\text{sp}} := Q_{4,l} u^* = \sum_{\boldsymbol{i} \in I_{4,l}^d} \lambda_i(u^*) N_{l,\boldsymbol{i}}^{(4)}$ satisfies:

$$\|u^* - u_{\text{sp}}\|_{H^r(\Omega)} = \|u^* - Q_{4,l} u^*\|_{H^r(\Omega)} \le C \left(\frac{1}{l}\right)^{s-r} \|u^*\|_{H^s(\Omega)} = C N^{-\frac{s-r}{d}} \|u^*\|_{H^s(\Omega)}.$$

We will then show that the linear combination $u_{\text{sp}} = \sum_{\boldsymbol{i} \in I_{4,l}^d} \lambda_i(f) N_{l,\boldsymbol{i}}^{(4)}$ can be implemented by some Deep Neural Network $u_{\text{DNN}} \in \Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(\log N)$. Firstly, note that for $x \ge 0$, both x and x^2 can be expressed in terms of the ReLU3 activation function η_3 with no error:

$$x = -\frac{1}{12}[\eta_3(x+3) - 5\eta_3(x+2) + 7\eta_3(x+1) - 3\eta_3(x) + 6]$$
$$x^2 = -\frac{1}{6}[\eta_3(x+2) - 4\eta_3(x+1) + 3\eta_3(x) - 4]$$

Applying the explicit formula listed in equation A.50 implies that for any $-3 \leq i \leq l-1$, the univariate B-spline function $N_{l,i}^{(4)}(x)$ $(x \in [0,1])$ can be implemented by some ReLU3 Deep Neural Network v_{DNN} with both scalar input and scalar output. We have that for v_{DNN}, the depth L_v is 2 and the maximum width W_v is upper bounded by 11.
Secondly, for any $x, y \geq 0$, we have that the product operation $x \cdot y$ can be expressed in terms of the ReLU3 activation function η_3 with no error:

$$\begin{aligned} x \cdot y &= \frac{1}{2}[(x+y)^2 - x^2 - y^2] \\ &= -\frac{1}{12}\Big[\eta_3(x+y+2) - 4\eta_3(x+y+1) + 3\eta_3(x+y) \\ &- \eta_3(x+2) + 4\eta_3(x+1) - 3\eta_3(x) - \eta_3(y+2) + 4\eta_3(y+1) - 3\eta_3(y) + 4\Big] \end{aligned}$$

In (Schumaker, 2007), it has been proved that the B-splines are always non-negative, i.e $N_{l,i}^{(4)}(x) \geq 0,\ \forall\, x \in [0,1]$. Therefore, by multiplying the non-negative univariate B-splines, we can implement any multivariate B-spline $N_{l,\boldsymbol{i}}^{(4)} = \Pi_{j=1}^{d} N_{l,i_j}^{(4)}(x_j)$ with some ReLU3 Deep Neural Network p_{DNN}. We have that for p_{DNN}, the depth $L_p = \lceil \log_2 d \rceil + 2$ and the maximum width $W_p = \max\{11d, \frac{9}{2}d\}$. Hence, we can further claim that $u^* = \sum_{\boldsymbol{i} \in I_{4,l}^d} \lambda_i(u^*) N_{l,\boldsymbol{i}}^{(4)}$, which is a linear combination of the multivariate B-splines $N_{l,\boldsymbol{i}}^{(4)}$, can be implemented by some ReLU3 Deep Neural Network u_{DNN}. It remains to check that $u_{\text{DNN}} \in \Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N)$ and $B = O(N)$. Note that we can ensure that the hidden layers of u_{DNN} are of the same dimension W by adding inactive neurons.
For the depth L of u_{DNN}, we have that L is equal to $L_p + 1$, where L_p denotes the depth of the ReLU3 Deep Neural Network p_{DNN}. Thus, we have $L = L_p + 1 = \lceil \log_2 d \rceil + 3$, which implies that $L = O(1)$.
For the width W of u_{DNN}, we have that $W \leq |I_{k,l}^d| W_p$, where W_p denotes the width of the ReLU3 Deep Neural Network p_{DNN}. This implies:

$$W \leq |I_{k,l}^d| \times 11d = 11d(l+k)^d = 11d(l+4)^d = O(l^d) \Rightarrow W = O(N)$$

For the sparsity constraint S of u_{DNN}, starting from the third layer, the number of activated neurons is half of the number of activated neurons at previous layer. This yields the following upper bound on S:

$$S \leq 2(W + W + \sum_{j=0}^{L-2} \frac{W}{2^j}) \leq 8W \Rightarrow S = O(W) = O(N)$$

For the norm constraint B of u_{DNN}, we have the following upper bound on B from equation A.50 and equation A.52:

$$B = O(\max\{l^{k-1}, \sup_{\boldsymbol{i} \in I_{k,l}^d} \lambda_{\boldsymbol{i}}(u^*)\}) = O(\max\{l^3, l^d\}) = O(N)$$

Now we have shown that parameters L, W, S, B of the Deep Neural Network u_{DNN} are of the desired magnitude, which completes our proof. □

A.5 Final Upper Bound

In this subsection, we provide the proof of upper bounds for PINN and DRM. For both estimator, we first provide a meta-theorem to illustrate the approximation and generalization decomposition with a $O(1/n)$ fast rate generalization bound(Bartlett et al., 2005; Xu, 2020). Then we use truncated fourier basis estimator and neural network estimator as example to obtain the final rate.

A.5.1 Deep Ritz Methods

Theorem A.9 (Meta-theorem for Upper Bounds of Deep Ritz Methods)**.** *Let $u^* \in H^s(\Omega)$ denote the true solution to the PDE model with Dirichlet boundary condition:*

$$\begin{aligned} -\Delta u + Vu &= f \text{ on } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{A.54}$$

where $f \in L^2(\Omega)$ and $V \in L^\infty(\Omega)$ with $0 < V_{\min} \leq V(x) \leq V_{\max} > 0$. For a fixed function space $\boldsymbol{F}(\Omega)$, consider the empirical loss induced by the Deep Ritz Method:

$$\boldsymbol{E}_n(u) = \frac{1}{n}\sum_{j=1}^{n}\Big[|\Omega| \cdot \Big(\frac{1}{2}|\nabla u(X_j)|^2 + \frac{1}{2}V(X_j)|u(X_j)|^2 - f(X_j)u(X_j)\Big)\Big], \tag{A.55}$$

where $\{X_j\}_{j=1}^n$ are datapoints uniformly sampled from the domain Ω. Then the Deep Ritz estimator associated with function space $\boldsymbol{F}(\Omega)$ is defined as the minimizer of $\boldsymbol{E}_n(u)$ over the function space $\boldsymbol{F}(\Omega)$:

$$\hat{u}_{DRM} = \min_{u \in \boldsymbol{F}(\Omega)} \boldsymbol{E}_n(u)$$

Moreover, we assume that there exists some constant $C > 0$ such that all function u in the function space $\boldsymbol{F}(\Omega)$, the real solution u^* and f, V satisfy the following two conditions.

- *The gradients and function value are uniformly bounded*

$$\max\Big\{\sup_{u\in\boldsymbol{F}(\Omega)}\|u\|_{L^\infty(\Omega)}, \sup_{u\in\boldsymbol{F}(\Omega)}\|\nabla u\|_{L^\infty(\Omega)}, \|u^*\|_{L^\infty(\Omega)}, \|\nabla u^*\|_{L^\infty(\Omega)}, V_{max}, \|f\|_{L^\infty(\Omega)}\Big\} \leq C. \tag{A.56}$$

- *All the functions in the function space $\boldsymbol{F}(\Omega)$ satisfies the boundary condition*

$$u = 0 \text{ on } \partial\Omega.$$

At the the same time, for any $\rho > 0$, we assume the Rademacher complexity of a localized function space $\boldsymbol{S}_\rho(\Omega) := \Big\{h := |\Omega| \cdot \Big[\frac{1}{2}\Big(|\nabla u|^2 - |\nabla u^*|^2\Big) + \frac{1}{2}V(|u|^2 - |u^*|^2) - f(u-u^*)\Big] \;\Big|\; \|u - u^*\|_{H^1}^2 \leq \rho\Big\}$ can be upper bounded by a sub-root function $\phi = \phi(\rho) : [0,\infty) \to [0,\infty)$, i.e.

$$\phi(4\rho) \leq 2\phi(\rho) \text{ and } R_n(\boldsymbol{S}_\rho(\Omega)) \leq \phi(\rho)\ (\forall\, \rho > 0). \tag{A.57}$$

For all constant $t > 0$. We denote r^* to be the solution of the fix point equation of local Rademacher complexity $r = \phi(r)$. There exists a constant C_p such that for probability $1 - C_p\exp(-t)$, we have the following upper bound for the Deep Ritz Estimator

$$\|\hat{u}_{DRM} - u^*\|_{H^1}^2 \lesssim \inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\Big\{r^*, \frac{t}{n}\Big\}.$$

Proof. To upper bound the excess risk $\Delta\boldsymbol{E}^{(n)} := \boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}(u^*)$, following(Xu, 2020; Lu et al., 2021b; Duan et al., 2021), we decompose the excess risk into approximation error and generalization error with probability $1 - e^{-t}$:

$$\begin{aligned}
\Delta\boldsymbol{E}^{(n)}(\hat{u}_{\text{DRM}}) = \boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}(u^\star) &= \big[\boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}_n(\hat{u}_{\text{DRM}})\big] + \big[\boldsymbol{E}_n(\hat{u}_{\text{DRM}}) - \boldsymbol{E}_n(u_{\boldsymbol{F}})\big] \\
&\quad + \big[\boldsymbol{E}_n(u_{\boldsymbol{F}}) - \boldsymbol{E}(u_{\boldsymbol{F}})\big] + \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] \\
&\leq \big[\boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}_n(\hat{u}_{\text{DRM}})\big] + \big[\boldsymbol{E}_n(u_{\boldsymbol{F}}) - \boldsymbol{E}(u_{\boldsymbol{F}})\big] + \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] \\
&\leq \big[\boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_n(u^*) - \boldsymbol{E}_n(u)\big] \\
&\quad + \frac{3}{2}\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] + \frac{t}{2n},
\end{aligned} \tag{A.58}$$

where the expectation is on all sampled data. The inequality of the third line is because the u is the minimizer of the empirical loss $\boldsymbol{E}_n$ in the solution set $\boldsymbol{F}(\Omega)$, so we have $\boldsymbol{E}_n(u) \leq \boldsymbol{E}_n(u_{\boldsymbol{F}})$. The last inequality is based on the Bernstein inequality. The variance of $h = |\Omega| \cdot \Big[\frac{1}{2}\Big(|\nabla u|^2 - |\nabla u^*|^2\Big) + \frac{1}{2}V(|u|^2 - |u^*|^2) - f(u-u^*)\Big]$ can be bounded by $\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big]$ due to the strong convexity of the variation objective (A.60). According to the Bernstein inequality, we know with probability $1 - e^{-t}$ we have

$$\boldsymbol{E}_n(u_{\boldsymbol{F}}) - \boldsymbol{E}_n(u^*) - \boldsymbol{E}(u_{\boldsymbol{F}}) + \boldsymbol{E}(u^*) \leq \sqrt{\frac{t\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big]}{n}} \leq \frac{1}{2}\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] + \frac{t}{2n}.$$

Note that A.58 holds for all function lies in the function space $\boldsymbol{F}$. Thus, we can take $u_{\boldsymbol{F}} := \arg\min_{u_0 \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_0) - \boldsymbol{E}(u^\star)\Big)$ and finally get

$$\Delta \boldsymbol{E}^{(n)} \leq \underbrace{\boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_n(u^*) - \boldsymbol{E}_n(u)}_{\Delta \boldsymbol{E}_{\text{gen}}} + \frac{3}{2} \underbrace{\inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big)}_{\Delta \boldsymbol{E}_{\text{app}}} + \frac{t}{2n}.$$

This inequality decompose the excess risk to the generalization error $\Delta \boldsymbol{E}_{\text{gen}} := \boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_n(u^*) - \boldsymbol{E}_n(\hat{u}_{\text{DRM}})$ and the approximation error $\Delta \boldsymbol{E}_{\text{app}} = \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big)$. From the lemmata proved in Section A.4, we already have an estimation of the approximation error's convergence rate. So now we'll focus on providing fast rate upper bounds of the generalization error for the two estimators using the localization technique(Bartlett et al., 2005; Xu, 2020). To achieve the fast generalization bound, we focus on the following normalized empirical process

$$\tilde{\boldsymbol{S}}_r(\Omega) := \big\{\tilde{h}(x) := \frac{\mathbb{E}[h] - h(x)}{\mathbb{E}[h] + r} \mid h \in \boldsymbol{S}(\Omega)\big\} \ (r > 0).$$

First, we try to bound the expectation of the normalized empirical process. Applying the Symmetrization Lemma A.1, we can first bound the expectation as

$$\sup_{\tilde{h} \in \tilde{S}_r(\Omega)} \mathbb{E}_{x'} \left[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i')\right] \leq \mathbb{E}_{x'} \left[\sup_{h \in S(\Omega)} \Big|\frac{1}{n} \sum_{i=1}^n \frac{h(x_i') - \mathbb{E}[h]}{\mathbb{E}[h] + r}\Big|\right] \leq 2R_n(\hat{\boldsymbol{S}}_r(\Omega)).$$

where the function class $\hat{\boldsymbol{S}}_r(\Omega)$ is defined as:

$$\hat{\boldsymbol{S}}_r(\Omega) := \big\{\hat{h}(x) := \frac{h(x)}{\mathbb{E}[h] + r} \mid h \in \boldsymbol{S}(\Omega)\big\},$$

where $\boldsymbol{S}(\Omega) = \Big\{h := |\Omega| \cdot \Big[\frac{1}{2}\Big(|\nabla u|^2 - |\nabla u^*|^2\Big) + \frac{1}{2}V(|u|^2 - |u^*|^2) - f(u - u^*)\Big]\Big\}$. Then Applying the Peeling Lemma to any function $h \in \boldsymbol{S}(\Omega)$ helps us upper bound the local Rademacher complexity $R_n(\hat{\boldsymbol{S}}_r(\Omega))$ with the function ϕ defined in equation A.57:

$$R_n(\hat{\boldsymbol{S}}_r(\Omega)) = \mathbb{E}_\sigma \left[\mathbb{E}_x\Big[\sup_{h \in \boldsymbol{S}(\Omega)} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)}{\mathbb{E}[h] + r}\Big]\right] \leq \frac{4\phi(r)}{r}.$$

Combining all inequalities derived above yields:

$$\sup_{\tilde{h} \in \tilde{S}_r(\Omega)} \mathbb{E}_{x'} \left[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i')\right] \leq 2R_n(\hat{\boldsymbol{S}}_r(\Omega)) \leq \frac{8\phi(r)}{r} \ (r > 0). \tag{A.59}$$

Secondly we'll apply the Talagrand concentration inequality, which requires us to verify the condition needed. We will first check that the expectation value $\mathbb{E}[h]$ is always non-negative for any $h \in \boldsymbol{S}(\Omega)$:

$$\begin{aligned}
\mathbb{E}[h] &= \frac{1}{|\Omega|} \int_\Omega |\Omega| \cdot (\frac{1}{2}|\nabla u(x)|^2 + \frac{1}{2}V(x)|u(x)|^2 - f(x)u(x))dx \\
&- \frac{1}{|\Omega|} \int_\Omega |\Omega| \cdot (\frac{1}{2}|\nabla u^\star(x)|^2 + \frac{1}{2}V(x)|u^\star(x)|^2 - f(x)u^\star(x))dx \\
&= \boldsymbol{E}(u) - \boldsymbol{E}(u^\star) \geq 0 \Rightarrow \mathbb{E}[h] \geq 0.
\end{aligned}$$

We will proceed to verify that any $\tilde{h} = \frac{\mathbb{E}[h] - h}{\mathbb{E}[h] + r} \in \tilde{\boldsymbol{S}}_r(\Omega)$ is of bounded inf-norm. We need to prove that any $h \in \boldsymbol{S}(\Omega)$ is of bounded inf-norm beforehand. Using boundedness condition listed in

equation A.56 implies:

$$\begin{aligned}\|h\|_\infty &= |\Omega|\|\frac{1}{2}\Big(|\nabla u|^2 - |\nabla u^*|^2\Big) + \frac{1}{2}V(|u|^2 - |u^*|^2) - f(u-u^*)\|_\infty \\ &\le \frac{|\Omega|}{2}\Big(\|\nabla u\|_\infty^2 + \|\nabla u^*\|_\infty^2\Big) + \frac{|\Omega|}{2}V_{\max}\Big(\|u\|_\infty^2 + \|u^*\|_\infty^2\Big) + |\Omega|\|f\|_\infty\Big(\|u\|_\infty + \|u^*\|_\infty\Big) \\ &\le \frac{|\Omega|}{2}\times 2C^2 + \frac{|\Omega|}{2}V_{\max}\times 2C^2 + 2|\Omega|C^2 = |\Omega|(V_{\max}+3)C^2\end{aligned}$$

By taking $M := |\Omega|(V_{\max}+3)C^2$, we then have $\|h\|_\infty \le M$ for all $h \in \boldsymbol{S}(\Omega)$. Note that the denominator can be lower bounded by $|\mathbb{E}[h]+r| \ge r > 0$. Combining these two inequalities help us upper bound the inf-norm $\|\tilde{h}\|_\infty = \sup_{x\in\Omega}|\tilde{h}(x)|$ as follows:

$$\|\tilde{h}\|_\infty = \frac{\|\mathbb{E}[h]-h\|_\infty}{|\mathbb{E}[h]+r|} \le \frac{2\|h\|_\infty}{r} \le \frac{2M}{r} =: \beta.$$

We will then check the normalized functions $\frac{\mathbb{E}[h]-h(x)}{\mathbb{E}[h]+r}$ in $\tilde{S}_r(\Omega)$ have bounded second moment, which is satisfied because of the regularity results of the PDE. We aim to show that there exist some constants $\alpha, \alpha' > 0$, such that for any $h \in \boldsymbol{S}(\Omega)$, the following inequality holds:

$$\alpha\mathbb{E}[h^2] \le \|u-u^*\|_{H^1(\Omega)}^2 \le \alpha'\mathbb{E}[h]. \tag{A.60}$$

The RHS of the inequality follows from strong convexity of the DRM objective function proved in Theorem A.1:

$$\mathbb{E}[h] = \boldsymbol{E}(u) - \boldsymbol{E}(u^*) \ge \frac{\min\{1, V_{\min}\}}{4}\|u-u^*\|_{H^1(\Omega)}^2$$

The LHS of the inequality follows from boundedness condition listed in equation A.56 and the QM-AM inequality:

$$\begin{aligned}\mathbb{E}[h^2] &= \int_\Omega \left[\frac{1}{2}\Big(|\nabla u|^2 - |\nabla u^*|^2\Big) + \frac{1}{2}V(|u|^2-|u^*|^2) - f(u-u^*)\right]^2 dx \\ &\le \frac{3}{4}\int_\Omega \Big(|\nabla u|^2 - |\nabla u^*|^2\Big)^2 dx + \frac{3}{4}\int_\Omega V^2(|u|^2-|u^*|^2)^2 dx + 3\int_\Omega f^2(u-u^*)^2 dx \\ &\le \frac{3}{4}\int_\Omega \Big||\nabla u| - |\nabla u^*|\Big|^2(|\nabla u| + |\nabla u^*|)^2 dx + \frac{3}{4}V_{\max}^2\int_\Omega \Big||u|-|u^*|\Big|^2(|u|+|u^*|)^2 dx \\ &+ 3C^2\int_\Omega (u-u^*)^2 dx \le 3C^2\int_\Omega |\nabla u - \nabla u^*|^2 dx + 3C^2(1+V_{\max}^2)\int_\Omega |u-u^*|^2 dx \\ &\le 3C^2(1+V_{\max}^2)\|u-u^*\|_{H^1(\Omega)}^2\end{aligned}$$

By picking $\alpha' = \frac{4}{\min\{1,V_{\min}\}}$ and $\alpha = \frac{1}{3C^2(1+V_{\max}^2)}$, we have finished proving inequality A.60. Then we can can upper bound the expectation $\mathbb{E}[\tilde{h}^2]$ as:

$$\mathbb{E}[\tilde{h}^2] = \frac{\mathbb{E}[(h-\mathbb{E}[h])^2]}{|\mathbb{E}[h]+r|^2} = \frac{\mathbb{E}[h^2]-\mathbb{E}[h]^2}{|\mathbb{E}[h]+r|^2} \le \frac{\mathbb{E}[h^2]}{|\mathbb{E}[h]+r|^2}.$$

Using the fact that $\mathbb{E}[h] \ge 0$ and inequality A.60, we can lower bound the denominator $|\mathbb{E}[h]+r|^2$ as follows:

$$|\mathbb{E}[h]+r|^2 \ge 2\mathbb{E}[h]r \ge \frac{2r\alpha}{\alpha'}\mathbb{E}[h^2].$$

Therefore, we can deduce that:

$$\mathbb{E}[\tilde{h}^2] \le \frac{\mathbb{E}[h^2]}{|\mathbb{E}[h]+r|^2} \le \frac{\mathbb{E}[h^2]}{\frac{2r\alpha}{\alpha'}\mathbb{E}[h^2]} = \frac{\alpha'}{2r\alpha} =: \sigma^2.$$

Hence, any function in the localized class $\tilde{\boldsymbol{S}}_r(\Omega)$ is of bounded second moment.

It is easy to check that for any $\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)$, we have

$$\mathbb{E}[\tilde{h}] = \frac{\mathbb{E}[h] - \mathbb{E}[h]}{\mathbb{E}[h] + r} = 0,$$

i.e. any function in the localized class $\tilde{\boldsymbol{S}}_r(\Omega)$ is of zero mean.

Now we have verified that any function $\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)$ satisfies all the required conditions. By taking μ to be the uniform distribution on the domain Ω and applying Talagrand's Concentration inequality given in Lemma A.3, we have:

$$\mathbb{P}_x \left[\sup_{\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)} \frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) \geq 2 \sup_{\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)} \mathbb{E}_{x'} \Big[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i') \Big] + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2t\beta}{n} \right] \leq e^{-t}.$$

By using the upper bound deduced above and plugging in the expressions of β and σ, we can rewrite Talagrand's Concentration Inequality in the following way. With probability at least $1 - e^{-t}$, the inequality below holds:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) \leq \sup_{\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)} \frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) &\leq 2 \sup_{\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)} \mathbb{E}_{x'} \Big[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i') \Big] + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2t\beta}{n} \\ &\leq \frac{16\phi(r)}{r} + \sqrt{\frac{t\alpha'}{n\alpha r}} + \frac{4Mt}{nr} =: \psi(r). \end{aligned}$$

Let's pick the threshold radius r_0 to be:

$$r_0 = \max\{2^{14} r^*, \frac{24Mt}{n}, \frac{36\alpha' t}{\alpha n}\}. \tag{A.61}$$

Note that concavity of the function ϕ implies that $\phi(r) \leq r$ for any $r \geq r^*$. Combining this with the first inequality listed in A.57 yields:

$$\frac{16\phi(r)}{r} \leq \frac{2^{11}\phi(\frac{r_0}{2^{14}})}{2^{14}\frac{r_0}{2^{14}}} = \frac{1}{8} \times \frac{\phi(\frac{r_0}{2^{14}})}{\frac{r_)}{2^{14}}} \leq \frac{1}{8}.$$

On the other hand, applying equation A.61 yields:

$$\begin{aligned} \sqrt{\frac{\alpha' t}{n\alpha r_0}} &\leq \sqrt{\frac{\alpha' t}{n\alpha} \frac{\alpha n}{36\alpha' t}} = \frac{1}{6}, \\ \frac{4Mt}{nr_0} &\leq \frac{4Mt}{n} \times \frac{n}{24Mt} = \frac{1}{6}. \end{aligned}$$

Summing the three inequalities above implies:

$$\psi(r_0) = \frac{16\phi(r_0)}{r_0} + \sqrt{\frac{t\alpha'}{n\alpha r_0}} + \frac{4Mt}{nr_0} \leq \frac{1}{8} + \frac{1}{6} + \frac{1}{6} < \frac{1}{2}.$$

By picking $r = r_0$, we can further deduce that for any function $u \in \boldsymbol{F}(\Omega)$, the following inequality holds with probability $1 - e^{-t}$:

$$\frac{\boldsymbol{E}(u) - \boldsymbol{E}(u^*) - \boldsymbol{E}_n(u) + \boldsymbol{E}_n(u^*)}{\boldsymbol{E}(u) - \boldsymbol{E}(u^*) + r_0} = \frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) \leq \psi(r_0) < \frac{1}{2}.$$

Multiplying the denominator on both sides indicates:

$$\Delta \boldsymbol{E}_{\text{gen}} = \boldsymbol{E}(u) - \boldsymbol{E}(u^*) - \boldsymbol{E}_n(u) + \boldsymbol{E}_n(u^*) \leq \frac{1}{2} \Big[\boldsymbol{E}(u) - \boldsymbol{E}(u^*) \Big] + \frac{1}{2} r_0 = \frac{1}{2} \Delta \boldsymbol{E}^{(n)} + \frac{1}{2} r_0.$$

Substituting the upper bound above into the decomposition $\Delta \boldsymbol{E}^{(n)} \leq \Delta E_{\text{gen}} + \frac{3}{2} \Delta E_{\text{app}} + \frac{t}{2n}$ yields that with probability $1 - e^{-t}$, we have:

$$\Delta \boldsymbol{E}^{(n)} \leq \Delta \boldsymbol{E}_{\text{gen}} + \frac{3}{2} \Delta \boldsymbol{E}_{\text{app}} + \frac{t}{2n} \leq \frac{1}{2} \Delta \boldsymbol{E}^{(n)} + \frac{1}{2} r_0 + \frac{3}{2} \Delta \boldsymbol{E}_{\text{app}} + \frac{t}{2n}.$$

Simplifying the inequality above yields that with probability $1 - e^{-t}$, we have:

$$\Delta \boldsymbol{E}^{(n)} \leq r_0 + 3\Delta \boldsymbol{E}_{\text{app}} + \frac{t}{n} = 3 \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big) + \max\{2^{14} r^*, 24M \frac{t}{n}, \frac{36\alpha'}{\alpha} \frac{t}{n}\} + \frac{t}{n}$$
$$\lesssim \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big) + \max\Big\{ r^*, \frac{t}{n} \Big\}$$

Moreover, using strong convexity of the DRM objective function proved in Theorem A.1 implies:

$$\Delta \boldsymbol{E}^{(n)} = \boldsymbol{E}(\hat{u}_{\text{DRM}}) - \boldsymbol{E}(u^*) \geq \frac{\min\{1, V_{\min}\}}{4} \|\hat{u}_{\text{DRM}} - u^*\|^2_{H^1(\Omega)}$$

Combining the two bounds above yields that with probability $1 - e^{-t}$, we have:

$$\|\hat{u}_{\text{DRM}} - u^*\|^2_{H^1(\Omega)} \lesssim \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big) + \max\Big\{ r^*, \frac{t}{n} \Big\}$$

□

Deep Neural Network Estimator. For any $N \in \mathbb{Z}^+$, there exists some Deep Neural Network in $\Phi(L, W, S, B)$ with $L = O(1),\ W = O(N),\ S = O(N),\ B = O(N)$, such that the approximation error $\Delta \boldsymbol{E}_{\text{app}} = O(N^{-\frac{2(s-1)}{d}})$ and generalization error $\Delta \boldsymbol{E}_{\text{gen}} = O(\frac{N \log N}{n})$. With optimal selection $N = n^{\frac{d}{d+2s-2}}$ to balance the bias and variance, we can achieve $n^{-\frac{2s-2}{d+2s-2}} \log n$ convergence rate for DRM estimator.

Theorem A.10. *(Final Upper Bound of DRM with Deep Neural Network Estimator) Consider the sparse Deep Neural Network function space $\Phi(L, W, S, B)$ with parameters $L = O(1),\ W = O(n^{\frac{d}{d+2s-2}}),\ S = O(n^{\frac{d}{d+2s-2}}),\ B = O(n^{\frac{d}{d+2s-2}})$, then the Deep ritz estimator $\hat{u}^{DNN}_{DRM} = \min_{u \in \Phi(L,W,S,B)} \boldsymbol{E}_n(u)$ satisfies the following upper bound with high probability*

$$\|\hat{u}^{DNN}_{DRM} - u^*\|^2_{H^1} \lesssim n^{-\frac{2s-2}{d+2s-2}} \log n.$$

Proof. On the one hand, by taking $s = 1$ and $p = 2$ in Theorem A.8 proved above, we have that there exists some Deep Neural Network $u_{\text{DNN}} \in \Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N), B = O(N)$, such that.

$$\|u_{\text{DNN}} - u^*\|^2_{H^1(\Omega)} \leq N^{-\frac{2s-2}{d}} \|u^*\|_{H^s(\Omega)}.$$

Applying strong convexity of the DRM objective function proved in Section 2.1 further implies:

$$\Delta \boldsymbol{E}_{\text{app}} \lesssim \|u_{\text{DNN}} - u^*\|^2_{H^1(\Omega)} \leq N^{-\frac{2s-2}{d}}.$$

On the other hand, from Lemma A.17 proved above, we know that the function $\phi(\rho)$ that upper bounds the local Rademacher complexity of the Deep Neural Network space is of the same magnitude as $\sqrt{\frac{S3^L \rho}{n} \log(BWn)}$. By plugging in the magnitudes of L, W, S, B, we can determine the thresholding localization radius r^*:

$$\sqrt{\frac{r^* 3^L S}{n} \log(BWn)} \simeq \sqrt{\frac{r^* N}{n} (2 \log N + \log n)} \simeq r^* \Rightarrow r^* \simeq \frac{N(\log N + \log n)}{n}.$$

Combining the two bounds above with Theorem A.9 yields that with high probability, we have:

$$\|\hat{u}^{\text{DNN}}_{\text{DRM}} - u^*\|^2_{H^1} \lesssim \Delta \boldsymbol{E}_{\text{app}} + \hat{r} \lesssim N^{-\frac{2(s-1)}{d}} + \frac{N(\log N + \log n)}{n}.$$

By equating the two terms above, we can solve for the optimal N that yields the desired bound:

$$N^{-\frac{2(s-1)}{d}} \simeq \frac{N}{n} \Rightarrow N \simeq n^{\frac{d}{d+2s-2}}.$$

Plugging in the optimal N gives us the magnitudes of the four parameters $L = O(1),\ W = O(n^{\frac{d}{d+2s-2}}),\ S = O(n^{\frac{d}{d+2s-2}}),\ B = O(n^{\frac{d}{d+2s-2}})$, as well as the final rate:

$$\|\hat{u}^{\text{DNN}}_{\text{DRM}} - u^*\|^2_{H^1} \lesssim N^{-\frac{2(s-1)}{d}} + \frac{N \log N}{n} \lesssim n^{-\frac{2(s-1)}{d+2(s-1)}} \log n.$$

□

Truncated Fourier Series Estimator. For any $\xi \in \mathbb{Z}^+$, there exists some Truncated Fourier Series in $\boldsymbol{F}_\xi(\Omega)$ with approximation error $\Delta \boldsymbol{E}_{\text{app}} = O(\xi^{-2(s-1)})$ and generalization error $\Delta \boldsymbol{E}_{\text{gen}} = O(\frac{\xi^d}{n})$

Theorem A.11. *(Final Upper Bound of DRM with Truncated Fourier Series Estimator) Consider the Deep Ritz objective with a plug in Fourier Series estimator $\hat{u}_{DRM}^{Fourier} = \min_{u \in \boldsymbol{F}_\xi(\Omega)} \boldsymbol{E}_n(u)$ with $\xi = \Theta(n^{\frac{1}{d+2s-2}})$, then we have*

$$\|\hat{u}_{DRM}^{Fourier} - u^*\|_{H^1}^2 \lesssim n^{-\frac{2s-2}{d+2s-2}}$$

Proof. On the one hand, from Lemma A.6 and Lemma A.7 proved above, we know that the function $\phi(\rho)$ that upper bounds the local Rademacher complexity for Truncated Fourier Series can be dominated by the term

$$\begin{aligned} R_n(\boldsymbol{S}_\rho(\Omega)) \lesssim & \mathbb{E}_X \left[\mathbb{E}_\sigma \Big[\sup_{f \in \boldsymbol{F}_{\rho,\xi}(\Omega)} \frac{1}{n} \sum_{i=1}^n \sigma_i \|\nabla f(X_i)\| \, \Big| \|f\|_{H^1(\Omega)}^2 \le \rho \Big] \right] \\ & + \mathbb{E}_X \left[\mathbb{E}_\sigma \Big[\sup_{f \in \boldsymbol{F}_{\rho,\xi}(\Omega)} \frac{1}{n} \sum_{i=1}^n \sigma_i \|\nabla \Pi_\xi f(X_i)\| \Big] \right] \\ & \le \sqrt{\frac{\rho}{n}} \xi^{\frac{d}{2}} + \sqrt{\frac{\|\Pi_\xi f\|_{H^1}^2}{n}} \le \sqrt{\frac{\rho}{n}} \xi^{\frac{d}{2}} + \sqrt{\frac{\xi^{-2(s-1)}}{n}} \lesssim \sqrt{\frac{\rho}{n}} \xi^{\frac{d}{2}} + \frac{1}{n} + \xi^{-2(s-1)} \end{aligned} \tag{A.62}$$

Thus, the localization radius $\hat{r}$ can be determined as follows:

$$\sqrt{\frac{\rho}{n}} \xi^{\frac{d}{2}} + \frac{1}{n} + \xi^{-2(s-1)} \simeq \rho \Rightarrow \hat{r} \simeq \frac{\xi^d}{n} + \frac{1}{n} + \xi^{-2(s-1)},$$

On the other hand, by taking $\alpha = s$ and $\beta = 1$ in Lemma A.19 and applying strong convexity of the DRM objective function proved in Section 2.1, we can upper bound the approximation error $\Delta \boldsymbol{E}_{\text{app}}$ as below:

$$\Delta \boldsymbol{E}_{\text{app}} \lesssim \xi^{-2(s-1)},$$

By equating the two terms above, we can solve for ξ that yields the desired bound:

$$\frac{\xi^d}{n} \simeq \xi^{-2(s-1)} \Rightarrow \xi \simeq n^{\frac{1}{d+2s-2}},$$

Plugging in the expression of ξ gives the final upper bound:

$$\mathbb{E}_{x \sim \mu}[\Delta \boldsymbol{E}_n] \lesssim \hat{r} + \Delta \boldsymbol{E}_{\text{app}} \lesssim \frac{\xi^d}{n} + \xi^{-2(s-1)} \simeq n^{-\frac{2s-2}{d+2s-2}}.$$

□

A.5.2 Physics Informed Neural Network

Theorem A.12 (Meta-theorem for Upper Bounds of Physics Informed Neural Network)**.** *Let $u^* \in H^s(\Omega)$ denote the true solution to the PDE model with Dirichlet boundary condition:*

$$\begin{aligned} -\Delta u + V u &= f \text{ on } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{A.63}$$

where $f \in L^2(\Omega)$ and $V \in L^\infty(\Omega)$ with $0 < V_{\min} \le V(x) \le V_{\max} > 0$. For a fixed function space $\boldsymbol{F}(\Omega)$, consider the empirical loss induced by the Physics Informed Neural Network:

$$\boldsymbol{E}_n(u) = \frac{1}{n} \sum_{j=1}^n \Big[|\Omega| \cdot \Big(\Delta u(X_j) - V(X_j) u(X_j) + f(X_j) \Big)^2 \Big], \tag{A.64}$$

where $\{X_j\}_{j=1}^n$ are datapoints uniformly sampled from the domain Ω. Then the Physics Informed Neural Network estimator associated with function space $\boldsymbol{F}(\Omega)$ is defined as the minimizer of $\boldsymbol{E}_n(u)$ over the function space $\boldsymbol{F}(\Omega)$:

$$\hat{u}_{PINN} = \min_{u \in \boldsymbol{F}(\Omega)} \boldsymbol{E}_n(u)$$

Moreover, we assume that there exists some constant $C > 0$ such that all function u in the function space $\boldsymbol{F}(\Omega)$, the real solution u^* and f, V satisfy the following two conditions.

- *The gradients and function value are uniformly bounded*

$$\max\Big\{ \sup_{u \in \boldsymbol{F}(\Omega)} \|u\|_{L^\infty(\Omega)}, \sup_{u \in \boldsymbol{F}(\Omega)} \|\nabla u\|_{L^\infty(\Omega)}, \sup_{u \in \boldsymbol{F}(\Omega)} \|\Delta u\|_{L^\infty(\Omega)}, \\ \|u^*\|_{L^\infty(\Omega)}, \|\nabla u^*\|_{L^\infty(\Omega)}, \|\Delta u^*\|_{L^\infty(\Omega)}, V_{max}, \|f\|_{L^\infty(\Omega)} \Big\} \leq C. \tag{A.65}$$

- *All the functions in the function space $\boldsymbol{F}(\Omega)$ satisfies the boundary condition*

$$u = 0 \text{ on } \partial\Omega.$$

At the the same time, for any $\rho > 0$, we assume the Rademacher complexity of a localized function space $\boldsymbol{T}_\rho(\Omega) := \Big\{ h := |\Omega| \cdot \big[(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2\big] \;\Big|\; \|u - u^*\|_{H^2}^2 \leq \rho \Big\}$ can be upper bounded by a sub-root function $\phi = \phi(\rho) : [0, \infty) \to [0, \infty)$, i.e.

$$\phi(4\rho) \leq 2\phi(\rho) \text{ and } R_n(\boldsymbol{T}_\rho(\Omega)) \leq \phi(\rho) \ (\forall\, \rho > 0). \tag{A.66}$$

For all constant $t > 0$. We denote r^* to be the solution of the fix point equation of local Rademacher complexity $r = \phi(r)$. There exists a constant C_p such that for probability $1 - C_p \exp(-t)$, we have the following upper bound for the Physics Informed Neural Network Estimator

$$\|\hat{u}_{PINN} - u^*\|_{H^2}^2 \lesssim \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big) + \max\Big\{ r^*, \frac{t}{n} \Big\}.$$

Proof. To upper bound the excess risk $\Delta \boldsymbol{E}^{(n)}$, following(Xu, 2020; Lu et al., 2021b; Duan et al., 2021), we decompose the excess risk into approximation error and generalization error with probability $1 - e^{-t}$:

$$\begin{aligned} \Delta \boldsymbol{E}^{(n)}(\hat{u}_{\text{PINN}}) = \boldsymbol{E}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}(u^\star) &= \big[\boldsymbol{E}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}_n(\hat{u}_{\text{PINN}})\big] + \big[\boldsymbol{E}_n(\hat{u}_{\text{PINN}}) - \boldsymbol{E}_n(u_{\boldsymbol{F}})\big] \\ &\quad + \big[\boldsymbol{E}_n(u_{\boldsymbol{F}}) - \boldsymbol{E}(u_{\boldsymbol{F}})\big] + \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] \\ &\leq \big[\boldsymbol{E}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}_n(\hat{u}_{\text{PINN}})\big] + \big[\boldsymbol{E}_n(u_{\boldsymbol{F}}) - \boldsymbol{E}(u_{\boldsymbol{F}})\big] + \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] \\ &\leq \big[\boldsymbol{E}(\hat{\boldsymbol{E}}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_n(u^*) - \boldsymbol{E}_n(\hat{u}_{\text{PINN}})\big] \\ &\quad + \frac{3}{2}\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] + \frac{t}{2n}, \end{aligned} \tag{A.67}$$

where the expectation is on all sampled data. The inequality of the third line is because the u is the minimizer of the empirical loss $\boldsymbol{E}_n$ in the solution set $\boldsymbol{F}(\Omega)$, so we have $\boldsymbol{E}_n(u) \leq \boldsymbol{E}_n(u_{\boldsymbol{F}})$. The last inequality is based on the Bernstein inequality. The variance of $h = |\Omega| \cdot \big[(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2\big]$ can be bounded by $\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big]$ due to the strong convexity of the variation objective (A.69). According to the Brenstein inequality, we know with probability $1 - e^{-t}$ we have

$$\boldsymbol{E}_n(u_{\boldsymbol{F}}) - \boldsymbol{E}_n(u^*) - \boldsymbol{E}(u_{\boldsymbol{F}}) + \boldsymbol{E}(u^*) \leq \sqrt{\frac{t\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big]}{n}} \leq \frac{1}{2}\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] + \frac{t}{2n}.$$

Note that B.5 holds for all function lies in the function space $\boldsymbol{F}$. Thus, we can take $u_{\boldsymbol{F}} := \arg\min_{u_0 \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_0) - \boldsymbol{E}(u^\star) \Big)$ and finally get

$$\Delta \boldsymbol{E}^{(n)} \leq \underbrace{\boldsymbol{E}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_n(u^*) - \boldsymbol{E}_n(\hat{u}_{\text{PINN}})}_{\Delta \boldsymbol{E}_{\text{gen}}} + \frac{3}{2} \underbrace{\inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big)}_{\Delta \boldsymbol{E}_{\text{app}}} + \frac{t}{2n}.$$

This inequality decompose the excess risk to the generalization error $\Delta \boldsymbol{E}_{\text{gen}} := \boldsymbol{E}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_n(u^*) - \boldsymbol{E}_n(\hat{u}_{\text{PINN}})$ and the approximation error $\Delta \boldsymbol{E}_{\text{app}} = \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big)$. From the lemmata proved in Section A.4, we already have an estimation of the approximation error's convergence rate. So now we'll focus on providing fast rate upper bounds of the generalization error for the two estimators using the localization techinque(Bartlett et al., 2005; Xu, 2020). To achieve the fast generalization bound, we focus on the following normalized empirical process

$$\tilde{\boldsymbol{T}}_r(\Omega) := \big\{\tilde{h}(x) := \frac{\mathbb{E}[h] - h(x)}{\mathbb{E}[h] + r} \mid h \in \boldsymbol{T}(\Omega)\big\} \ (r > 0).$$

First, we try to bound the expectation of the normalized empirical process. Applying the Symmetrization Lemma A.1, we can first bound the expectation as

$$\sup_{\tilde{h} \in \tilde{T}_r(\Omega)} \mathbb{E}_{x'} \left[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i')\right] \leq \mathbb{E}_{x'} \left[\sup_{h \in T(\Omega)} \Big|\frac{1}{n} \sum_{i=1}^n \frac{h(x_i') - \mathbb{E}[h]}{\mathbb{E}[h] + r}\Big|\right] \leq 2R_n(\hat{\boldsymbol{T}}_r(\Omega)).$$

where the function class $\hat{\boldsymbol{S}}_r(\Omega)$ is defined as:

$$\hat{\boldsymbol{T}}_r(\Omega) := \big\{\hat{h}(x) := \frac{h(x)}{\mathbb{E}[h] + r} \mid h \in \boldsymbol{T}(\Omega)\big\},$$

where $\boldsymbol{T}(\Omega) = \Big\{h := |\Omega| \cdot \big[(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2\big] \Big\}$. Then Applying the Peeling Lemma to any function $h \in \boldsymbol{T}(\Omega)$ helps us upper bound the local Rademacher complexity $R_n(\hat{\boldsymbol{T}}_r(\Omega))$ with the function ϕ defined in equation A.66:

$$R_n(\hat{\boldsymbol{T}}_r(\Omega)) = \mathbb{E}_\sigma \left[\mathbb{E}_x \Big[\sup_{h \in \boldsymbol{T}(\Omega)} \frac{\frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)}{\mathbb{E}[h] + r}\Big]\right] \leq \frac{4\phi(r)}{r}.$$

Combining all inequalities derived above yields:

$$\sup_{\tilde{h} \in \tilde{T}_r(\Omega)} \mathbb{E}_{x'} \left[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i')\right] \leq 2R_n(\hat{\boldsymbol{T}}_r(\Omega)) \leq \frac{8\phi(r)}{r} \ (r > 0). \tag{A.68}$$

Secondly we'll apply the Talagrand concentration inequality, which requires us to verify the condition needed. We will first check that the expectation value $\mathbb{E}[h]$ is always non-negative for any $h \in \boldsymbol{S}(\Omega)$:

$$\begin{aligned}
\mathbb{E}[h] &= \frac{1}{|\Omega|} \int_\Omega |\Omega| \cdot (\frac{1}{2}|\nabla u(x)|^2 + \frac{1}{2}V(x)|u(x)|^2 - f(x)u(x))dx \\
&- \frac{1}{|\Omega|} \int_\Omega |\Omega| \cdot (\frac{1}{2}|\nabla u^\star(x)|^2 + \frac{1}{2}V(x)|u^\star(x)|^2 - f(x)u^\star(x))dx \\
&= \boldsymbol{E}(u) - \boldsymbol{E}(u^\star) \geq 0 \Rightarrow \mathbb{E}[h] \geq 0.
\end{aligned}$$

We will proceed to verify that any $\tilde{h} = \frac{\mathbb{E}[h] - h}{\mathbb{E}[h] + r} \in \tilde{\boldsymbol{T}}_r(\Omega)$ is of bounded inf-norm. We need to prove that any $h \in \boldsymbol{T}(\Omega)$ is of bounded inf-norm beforehand. Using boundedness condition listed in equation A.65 implies:

$$\begin{aligned}
\|h\|_\infty &= |\Omega| \cdot \|(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2\|_\infty = |\Omega| \cdot \|(\Delta u - Vu + f)^2\|_\infty \\
&\leq |\Omega| \cdot (\|\Delta u\|_\infty + V_{\max}\|u\|_\infty + \|f\|_\infty)^2 \leq |\Omega|(V_{\max} + 2)^2 C^2
\end{aligned}$$

By taking $M := |\Omega|(V_{\max} + 2)^2 C^2$, we then have $\|h\|_\infty \leq M$ for all $h \in \boldsymbol{T}(\Omega)$. Note that the denominator can be lower bounded by $|\mathbb{E}[h] + r| \geq r > 0$. Combining these two inequalities help us upper bound the inf-norm $\|\tilde{h}\|_\infty = \sup_{x \in \Omega} |\tilde{h}(x)|$ as follows:

$$\|\tilde{h}\|_\infty = \frac{\|\mathbb{E}[h] - h\|_\infty}{|\mathbb{E}[h] + r|} \leq \frac{2\|h\|_\infty}{r} \leq \frac{2M}{r} =: \beta.$$

We will then check the normalized functions $\frac{\mathbb{E}[h]-h(x)}{\mathbb{E}[h]+r}$ in $\tilde{T}_r(\Omega)$ have bounded second moment, which is satisfied because of the regularity results of the PDE. We aim to show that there exist some constants $\alpha, \alpha' > 0$, such that for any $h \in \boldsymbol{T}(\Omega)$, the following inequality holds:

$$\alpha \mathbb{E}[h^2] \leq \|u - u^*\|^2_{H^2(\Omega)} \leq \alpha' \mathbb{E}[h]. \tag{A.69}$$

The RHS of the inequality follows from strong convexity of the PINN objective function proved in Theorem A.2:

$$\mathbb{E}[h] = \boldsymbol{E}(u) - \boldsymbol{E}(u^*) \geq \frac{1}{\min\{1, C_{\min}\}} \|u - u^*\|^2_{H^2(\Omega)}$$

The LHS of the inequality follows from boundedness condition listed in equation A.65 and the QM-AM inequality:

$$\begin{aligned}
\mathbb{E}[h^2] &= \int_\Omega \left[(\Delta u - Vu + f)^2 - (\Delta u^* - Vu^* + f)^2\right]^2 dx = \int_\Omega (\Delta u - Vu + f)^4 dx \\
&\leq M^2 \int_\Omega (\Delta u - Vu - \Delta u^* + Vu^*)^2 dx \leq 2M^2 \int_\Omega [(\Delta u - \Delta u^*)^2 + V^2(u - u^*)^2] dx \\
&\leq 2M^2 \max\{1, V^2_{\max}\} \|u - u^*\|^2_{H^2(\Omega)}
\end{aligned}$$

By picking $\alpha' = \frac{1}{\min\{1, C_{\min}\}}$ and $\alpha = \frac{1}{M^2 \max\{1, V^2_{\max}\}}$, we have finished proving inequality A.69. Then we can can upper bound the expectation $\mathbb{E}[\tilde{h}^2]$ as:

$$\mathbb{E}[\tilde{h}^2] = \frac{\mathbb{E}[(h - \mathbb{E}[h])^2]}{|\mathbb{E}[h] + r|^2} = \frac{\mathbb{E}[h^2] - \mathbb{E}[h]^2}{|\mathbb{E}[h] + r|^2} \leq \frac{\mathbb{E}[h^2]}{|\mathbb{E}[h] + r|^2}.$$

Using the fact that $\mathbb{E}[h] \geq 0$ and inequality A.69, we can lower bound the denominator $|\mathbb{E}[h] + r|^2$ as follows:

$$|\mathbb{E}[h] + r|^2 \geq 2\mathbb{E}[h]r \geq \frac{2r\alpha}{\alpha'} \mathbb{E}[h^2].$$

Therefore, we can deduce that:

$$\mathbb{E}[\tilde{h}^2] \leq \frac{\mathbb{E}[h^2]}{|\mathbb{E}[h] + r|^2} \leq \frac{\mathbb{E}[h^2]}{\frac{2r\alpha}{\alpha'} \mathbb{E}[h^2]} = \frac{\alpha'}{2r\alpha} =: \sigma^2.$$

Hence, any function in the localized class $\tilde{\boldsymbol{T}}_r(\Omega)$ is of bounded second moment.

It is easy to check that for any $\tilde{h} \in \tilde{\boldsymbol{T}}_r(\Omega)$, we have

$$\mathbb{E}[\tilde{h}] = \frac{\mathbb{E}[h] - \mathbb{E}[h]}{\mathbb{E}[h] + r} = 0,$$

i.e. any function in the localized class $\tilde{\boldsymbol{S}}_r(\Omega)$ is of zero mean.

Now we have verified that any function $\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)$ satisfies all the required conditions. By taking μ to be the uniform distribution on the domain Ω and applying Talagrand's Concentration inequality given in Lemma A.3, we have:

$$\mathbb{P}_x \left[\sup_{\tilde{h} \in \tilde{\boldsymbol{T}}_r(\Omega)} \frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) \geq 2 \sup_{\tilde{h} \in \tilde{\boldsymbol{T}}_r(\Omega)} \mathbb{E}_{x'} \left[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x'_i) \right] + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2t\beta}{n} \right] \leq e^{-t}.$$

By using the upper bound deduced above and plugging in the expressions of β and σ, we can rewrite Talagrand's Concentration Inequality in the following way. With probability at least $1 - e^{-t}$, the inequality below holds:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) \leq \sup_{\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)} \frac{1}{n} \sum_{i=1}^n \tilde{h}(x_i) &\leq 2 \sup_{\tilde{h} \in \tilde{\boldsymbol{S}}_r(\Omega)} \mathbb{E}_{x'} \left[\frac{1}{n} \sum_{i=1}^n \tilde{h}(x'_i) \right] + \sqrt{\frac{2t\sigma^2}{n}} + \frac{2t\beta}{n} \\
&\leq \frac{16\phi(r)}{r} + \sqrt{\frac{t\alpha'}{n\alpha r}} + \frac{4Mt}{nr} =: \psi(r)
\end{aligned}$$

Let's pick the threshold radius r_0 to be:

$$r_0 = \max\{2^{14}r^*, \frac{24Mt}{n}, \frac{36\alpha' t}{\alpha n}\}. \tag{A.70}$$

Note that concavity of the function ϕ implies that $\phi(r) \leq r$ for any $r \geq r^*$. Combining this with the first inequality listed in A.66 yields:

$$\frac{16\phi(r)}{r} \leq \frac{2^{11}\phi(\frac{r_0}{2^{14}})}{2^{14}\frac{r_0}{2^{14}}} = \frac{1}{8} \times \frac{\phi(\frac{r_0}{2^{14}})}{\frac{r_)}{2^{14}}} \leq \frac{1}{8}.$$

On the other hand, applying equation A.70 yields:

$$\sqrt{\frac{\alpha' t}{n\alpha r_0}} \leq \sqrt{\frac{\alpha' t}{n\alpha}\frac{\alpha n}{36\alpha' t}} = \frac{1}{6},$$
$$\frac{4Mt}{nr_0} \leq \frac{4Mt}{n} \times \frac{n}{24Mt} = \frac{1}{6}.$$

Summing the three inequalities above implies:

$$\psi(r_0) = \frac{16\phi(r_0)}{r_0} + \sqrt{\frac{t\alpha'}{n\alpha r_0}} + \frac{4Mt}{nr_0} \leq \frac{1}{8} + \frac{1}{6} + \frac{1}{6} < \frac{1}{2}.$$

By picking $r = r_0$, we can further deduce that for any function $u \in \boldsymbol{F}(\Omega)$, the following inequality holds with probability $1 - e^{-t}$:

$$\frac{\boldsymbol{E}(u) - \boldsymbol{E}(u^*) - \boldsymbol{E}_n(u) + \boldsymbol{E}_n(u^*)}{\boldsymbol{E}(u) - \boldsymbol{E}(u^*) + r_0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{h}(x_i) \leq \psi(r_0) < \frac{1}{2}.$$

Multiplying the denominator on both sides indicates:

$$\Delta\boldsymbol{E}_{\text{gen}} = \boldsymbol{E}(u) - \boldsymbol{E}(u^*) - \boldsymbol{E}_n(u) + \boldsymbol{E}_n(u^*) \leq \frac{1}{2}\Big[\boldsymbol{E}(u) - \boldsymbol{E}(u^*)\Big] + \frac{1}{2}r_0 = \frac{1}{2}\Delta\boldsymbol{E}^{(n)} + \frac{1}{2}r_0.$$

Substituting the upper bound above into the decomposition $\Delta\boldsymbol{E}^{(n)} \leq \Delta E_{\text{gen}} + \frac{3}{2}\Delta E_{\text{app}} + \frac{t}{2n}$ yields that with probability $1 - e^{-t}$, we have:

$$\Delta\boldsymbol{E}^{(n)} \leq \Delta\boldsymbol{E}_{\text{gen}} + \frac{3}{2}\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{2n} \leq \frac{1}{2}\Delta\boldsymbol{E}^{(n)} + \frac{1}{2}r_0 + \frac{3}{2}\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{2n}.$$

Simplifying the inequality above yields that with probability $1 - e^{-t}$, we have:

$$\begin{aligned}\Delta\boldsymbol{E}^{(n)} &\leq r_0 + 3\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{n} = 3\inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\{2^{14}r^*, 24M\frac{t}{n}, \frac{36\alpha'}{\alpha}\frac{t}{n}\} + \frac{t}{n} \\ &\lesssim \inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\Big\{r^*, \frac{t}{n}\Big\}\end{aligned}$$

Moreover, using strong convexity of the PINN objective function proved in Theorem A.1 implies:

$$\Delta\boldsymbol{E}^{(n)} = \boldsymbol{E}(\hat{u}_{\text{PINN}}) - \boldsymbol{E}(u^*) \geq \frac{1}{\{1, C_{\min}\}}\|\hat{u}_{\text{PINN}} - u^*\|^2_{H^1(\Omega)}$$

Combining the two bounds above yields that with probability $1 - e^{-t}$, we have:

$$\|\hat{u}_{\text{PINN}} - u^*\|^2_{H^1(\Omega)} \lesssim \inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\Big\{r^*, \frac{t}{n}\Big\}$$

□

Deep Neural Network Estimator. For any $N \in \mathbb{Z}^+$, there exists some Deep Neural Network in $\Phi(L, W, S, B)$ with $L = O(1),\ W = O(N),\ S = O(N),\ B = O(1)$, such that the approximation error $\Delta\boldsymbol{E}_{\text{app}} = O(N^{-\frac{2(s-2)}{d}})$ and generalization error $\Delta\boldsymbol{E}_{\text{gen}} = O(\frac{N\log N}{n})$. With optimal selection $N = n^{\frac{d}{d+2s-2}}$ to balance the bias and variance, we can achieve $n^{-\frac{2s-2}{d+2s-2}}\log n$ convergence rate for PINN estimator.

Theorem A.13. *(Final Upper Bound of PINN with Deep Neural Network Estimator) Consider the sparse Deep Neural Network function space $\Phi(L, W, S, B)$ with parameters $L = O(1),\ W = O(n^{\frac{d}{d+2s-4}}),\ S = O(n^{\frac{d}{d+2s-4}}),\ B = O(1)$, then the Physics Informed Neural Network estimator $\hat{u}_{PINN}^{DNN} = \min_{u\in\Phi(L,W,S,B)} \boldsymbol{E}_n(u)$ satisfies the following upper bound with high probability*

$$\|\hat{u}_{PINN}^{DNN} - u^*\|_{H^2}^2 \lesssim n^{-\frac{2s-4}{d+2s-4}} \log n.$$

Proof. On the one hand, by taking $s = 2$ and $p = 2$ in Theorem A.8 proved above, we have that there exists some Deep Neural Network $u_{\text{DNN}} \in \Phi(L, W, S, B)$ with $L = O(1), W = O(N), S = O(N), B = O(1)$, such that.

$$\|u_{\text{DNN}} - u^*\|_{H^2(\Omega)}^2 \leq N^{-\frac{2s-4}{d}} \|u\|_{H^s(\Omega)}.$$

Applying strong convexity of the DRM objective function proved in Section 2.1 further implies:

$$\Delta\boldsymbol{E}_{\text{app}} \lesssim \|u_{\text{DNN}} - u^*\|_{H^2(\Omega)}^2 \leq N^{-\frac{2s-4}{d}}.$$

On the other hand, from lemma A.18 proved above, we know that the function $\phi(\rho)$ that upper bounds the local Rademacher complexity of the Deep Neural Networks u_{DNN} is dominated by the term $\sqrt{\frac{\rho 3^L S}{n}\log(W(B\vee 1)n)}$. By plugging in the magnitudes of L, W, S, B, we can determine the thresholding localization radius $\hat{r}$:

$$\sqrt{\frac{\rho 3^L S}{n}\log(W(B\vee 1)n)} \simeq \sqrt{\frac{\rho N}{n}(\log N + \log n)} \simeq \rho \Rightarrow \hat{r} \simeq \frac{N(\log N + \log n)}{n}.$$

Combining the two bounds above gives us:

$$\mathbb{E}_{x\sim\mu}[\Delta\boldsymbol{E}_n] \lesssim \Delta\boldsymbol{E}_{\text{app}} + \hat{r} \lesssim N^{-\frac{2(s-2)}{d}} + \frac{N(\log N + \log n)}{n}.$$

By equating the two terms above, we can solve for the optimal N that yields the desired bound:

$$N^{-\frac{2(s-2)}{d}} \simeq \frac{N}{n} \Rightarrow N \simeq n^{\frac{d}{d+2s-4}}.$$

Plugging in the optimal N gives us the magnitudes of the four parameters $L = O(1),\ W = O(n^{\frac{d}{d+2s-4}}),\ S = O(n^{\frac{d}{d+2s-4}}),\ B = O(1)$, as well as the final rate:

$$\mathbb{E}_{x\sim\mu}[\Delta\boldsymbol{E}_n] \lesssim N^{-\frac{2(s-2)}{d}} + \frac{N\log N}{n} \lesssim n^{-\frac{2(s-2)}{d+2(s-2)}} \log n.$$

□

Truncated Fourier Series Estimator. For any $\xi \in \mathbb{Z}^+$, there exists some Truncated Fourier Series in $\boldsymbol{F}_\xi(\Omega)$ with approximation error $\Delta\boldsymbol{E}_{\text{app}} = O(\xi^{-2(s-2)})$ and generalization error $\Delta\boldsymbol{E}_{\text{gen}} = O(\frac{\xi^d}{n})$

Theorem A.14. *(Final Upper Bound of PINN with Truncated Fourier Series Estimator) Consider the PINN objective with a plug in Fourier Series estimator $\hat{u}_{PINN}^{Fourier} = \min_{u\in\boldsymbol{F}_\xi(\Omega)} \boldsymbol{E}_n(u)$ with $\xi = \Theta(n^{\frac{1}{d+2s-4}})$, then we have*

$$\|\hat{u}_{PINN}^{Fourier} - u^*\|_{H^2}^2 \lesssim n^{-\frac{2s-4}{d+2s-4}}$$

Proof. On the one hand, from Lemma A.6 and Lemma A.7 proved above, we know that the function $\phi(\rho)$ that upper bounds the local Rademacher complexity for Truncated Fourier Series is dominated by

$$\begin{aligned}
R_n(\boldsymbol{S}_\rho(\Omega)) \lesssim & \mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f\in\boldsymbol{F}_{\rho,\xi}(\Omega)} \frac{1}{n}\sum_{i=1}^n \sigma_i\|\Delta f(X_i)\| \,\Big|\, \|f\|_{H^2(\Omega)}^2 \leq \rho\Big]\right] \\
& + \mathbb{E}_X\left[\mathbb{E}_\sigma\Big[\sup_{f\in\boldsymbol{F}_{\rho,\xi}(\Omega)} \frac{1}{n}\sum_{i=1}^n \sigma_i\|\Delta\Pi_\xi f(X_i)\|\Big]\right] \\
& \leq \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}} + \sqrt{\frac{\|\Pi_\xi f\|_{H^2}^2}{n}} \leq \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}} + \sqrt{\frac{\xi^{-2(s-1)}}{n}} \lesssim \sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}} + \frac{1}{n} + \xi^{-2(s-2)}
\end{aligned} \tag{A.71}$$

Thus, the localization radius $\hat{r}$ can be determined as follows:

$$\sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}} + \frac{1}{n} + \xi^{-2(s-2)} \simeq \rho \Rightarrow \hat{r} \simeq \frac{\xi^d}{n} + \frac{1}{n} + \xi^{-2(s-2)},$$

On the other hand, by taking $\alpha = s$ and $\beta = 1$ in Lemma A.19 and applying strong convexity of the DRM objective function proved in Section 2.1, we can upper bound the approximation error $\Delta \boldsymbol{E}_{\text{app}}$ as below:

$$\Delta \boldsymbol{E}_{\text{app}} \lesssim \xi^{-2(s-2)},$$

By equating the two terms above, we can solve for ξ that yields the desired bound:

$$\frac{\xi^d}{n} \simeq \xi^{-2(s-2)} \Rightarrow \xi \simeq n^{\frac{1}{d+2s-4}},$$

Plugging in the expression of ξ gives the final upper bound:

$$\mathbb{E}_{x\sim\mu}[\Delta \boldsymbol{E}_n] \lesssim \hat{r} + \Delta \boldsymbol{E}_{\text{app}} \lesssim \frac{\xi^d}{n} + \xi^{-2(s-2)} \simeq n^{-\frac{2s-4}{d+2s-4}}.$$

□

B Proof of Modified DRM

In this section, we provide the proof of the modified deep Ritz method here. We first provide a similar meta-theorem as we did for DRM.

Theorem B.1 (Meta-theorem for Upper Bounds of Modified Deep Ritz Method)**.** *Let $u^* \in H^s(\Omega)$ denote the true solution to the PDE model with Dirichlet boundary condition:*

$$\begin{aligned} -\Delta u + Vu &= f \text{ on } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{B.1}$$

where $f \in L^2(\Omega)$ and $V \in L^\infty(\Omega)$ with $0 < V_{\min} \le V(x) \le V_{\max} > 0$. For a fixed function space $\boldsymbol{F}(\Omega)$, consider the empirical loss induced by the Modified Deep Ritz Method $(N \ge n)$:

$$\boldsymbol{E}_{N,n}(u) = \frac{1}{N}\sum_{i=1}^{N}\left[|\Omega| \cdot \frac{1}{2}|\nabla u(X_i')|^2\right] + \frac{1}{n}\sum_{j=1}^{n}\left[|\Omega| \cdot \left(\frac{1}{2}V(X_j)|u(X_j)|^2 - f(X_j)u(X_j)\right)\right], \tag{B.2}$$

where $\{X_i'\}_{i=1}^N$ and $\{X_j\}_{j=1}^n$ are datapoints uniformly and independently sampled from the domain Ω. Then the Modified Deep Ritz estimator associated with function space $\boldsymbol{F}(\Omega)$ is defined as the minimizer of $\boldsymbol{E}_{N,n}(u)$ over the function space $\boldsymbol{F}(\Omega)$:

$$\hat{u}_{MDRM} = \min_{u\in\boldsymbol{F}(\Omega)} \boldsymbol{E}_{N,n}(u)$$

Moreover, we assume that there exists some constant $C > 0$ such that all function u in the function space $\boldsymbol{F}(\Omega)$, the real solution u^* and f, V satisfy the following two conditions.

- *The gradients and function value are uniformly bounded*

$$\max\left\{\sup_{u\in\boldsymbol{F}(\Omega)} \|u\|_{L^\infty(\Omega)}, \sup_{u\in\boldsymbol{F}(\Omega)} \|\nabla u\|_{L^\infty(\Omega)}, \|u^*\|_{L^\infty(\Omega)}, \|\nabla u^*\|_{L^\infty(\Omega)}, V_{max}, \|f\|_{L^\infty(\Omega)}\right\} \le C. \tag{B.3}$$

- *All the functions in the function space $\boldsymbol{F}(\Omega)$ satisfy the boundary condition*

$$u = 0 \text{ on } \partial\Omega.$$

At the the same time, for any $\rho > 0$, we assume the Rademacher complexity of two localized function spaces

$$\boldsymbol{S}_\rho(\Omega) := \left\{(h_1, h_2) \middle| h1 := |\Omega| \cdot \left[\frac{1}{2}\left(|\nabla u|^2 - |\nabla u^*|^2\right)\right],\ h2 := |\Omega| \cdot \left[\frac{1}{2}V(|u|^2 - |u^*|^2) - f(u - u^*)\right], \|u - u^*\|_{H^1}^2 \le \rho\right\}$$

can be upper bounded by a sub-root function $\phi = \phi(\rho) : [0, \infty) \to [0, \infty)$, i.e.

$$\phi(4\rho) \leq 2\phi(\rho) \text{ and } R_{N,n}(\boldsymbol{S}_\rho(\Omega)) \leq \phi(\rho) \ (\forall\, \rho > 0), \tag{B.4}$$

where $R_{N,n}(\boldsymbol{S}) := R_N(\{h_1 | (h_1, h_2) \in \boldsymbol{S}\}) + R_n(\{h_2 | (h_1, h_2) \in \boldsymbol{S}\})$. *For all constant* $t > 0$. *We denote* r^* *to be the solution of the fix point equation of local Rademacher complexity* $r = \phi(r)$. *There exist a constant* C_p *such that for probability* $1 - C_p \exp(-t)$, *we have the following upper bound for the Modified Deep Ritz Estimator*

$$\|\hat{u}_{MDRM} - u^*\|_{H^1}^2 \lesssim \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big) + \max\Big\{ r^*, \frac{t}{n} \Big\}.$$

Proof. To upper bound the excess risk $\Delta \boldsymbol{E}^{(N,n)} := \boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}(u^*)$, following(Xu, 2020; Lu et al., 2021b; Duan et al., 2021), we decompose the excess risk into approximation error and generalization error with probability $1 - e^{-t}$:

$$\begin{aligned}
\Delta \boldsymbol{E}^{(N,n)} = \big[\boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}(u^\star)\big] &= \big[\boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}_{N,n}(\hat{u}_{\text{MDRM}})\big] + \big[\boldsymbol{E}_{N,n}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}_{N,n}(u_{\boldsymbol{F}})\big] \\
&\quad + \big[\boldsymbol{E}_{N,n}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u_{\boldsymbol{F}})\big] + \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] \\
&\leq \big[\boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}_{N,n}(\hat{u}_{\text{MDRM}})\big] + \big[\boldsymbol{E}_{N,n}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u_{\boldsymbol{F}})\big] + \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] \\
&\lesssim \big[\boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_{N,n}(u^*) - \boldsymbol{E}_{N,n}(\hat{u}_{\text{MDRM}})\big]\big] \\
&\quad + 2\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] + \frac{t}{\min\{N, n\}},
\end{aligned} \tag{B.5}$$

where the expectation is on all sampled data. The inequality of the third line is because the u is the minimizer of the empirical loss $\boldsymbol{E}_n$ in the solution set $\boldsymbol{F}(\Omega)$, so we have $\boldsymbol{E}_{N,n}(u) \leq \boldsymbol{E}_{N,n}(u_{\boldsymbol{F}})$. The last inequality is based on the Bernstein inequality. The variance can be bounded by $\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big]$ due to the strong convexity of the variation objective. According to the Brenstein inequality, we know with probability $1 - e^{-t}$ we have

$$\boldsymbol{E}_{N,n}(u_{\boldsymbol{F}}) - \boldsymbol{E}_{N,n}(u^*) - \boldsymbol{E}(u_{\boldsymbol{F}}) + \boldsymbol{E}(u^*) \leq \sqrt{\frac{2t\big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big]}{\min\{N, n\}}} \leq \big[\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\big] + \frac{4t}{\min\{N, n\}}.$$

Note that B.5 holds for all function lies in the function space $\boldsymbol{F}$. Thus, we can take $u_{\boldsymbol{F}} := \arg\min_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big)$ and finally get

$$\Delta \boldsymbol{E}^{N,n} \lesssim \underbrace{\mathbb{E}_{x,x' \sim \mu}[\boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_{N,n}(u^*) - \boldsymbol{E}_{N,n}(\hat{u}_{\text{MDRM}})]}_{\Delta \boldsymbol{E}_{\text{gen}}} + \underbrace{\inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big)}_{\Delta \boldsymbol{E}_{\text{app}}} + \frac{t}{n}.$$

This inequality decomposes the excess risk to the generalization error $\Delta \boldsymbol{E}_{\text{gen}} := \mathbb{E}_{x \sim \mu}[\boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}(u^*) + \boldsymbol{E}_{N,n}(u^*) - \boldsymbol{E}_{N,n}(\hat{u}_{\text{MDRM}})]$ and the approximation error $\Delta \boldsymbol{E}_{\text{app}} = \inf_{u_{\boldsymbol{F}} \in \boldsymbol{F}(\Omega)} \Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star) \Big)$. From the lemmata proved in Section A.4, we already have an estimation of the approximation error's convergence rate. So now we'll focus on providing fast rate upper bounds of the generalization error for the two estimators using the localization techinque(Bartlett et al., 2005; Xu, 2020). To achieve the fast generalization bound, we focus on the following two normalized empirical processes:

$$\begin{aligned}
\tilde{\boldsymbol{S}}_{r,1}(\Omega) &:= \Big\{ \tilde{h}_1(x) := \frac{\mathbb{E}[H^1] - H^1(x)}{\mathbb{E}[H^1] + \mathbb{E}[H^2] + r} \mid h = H^1 + H^2 \in \boldsymbol{S}(\Omega) \Big\} \ (r > 0), \\
\tilde{\boldsymbol{S}}_{r,2}(\Omega) &:= \Big\{ \tilde{h}_2(x) := \frac{\mathbb{E}[H^2] - H^2(x)}{\mathbb{E}[H^1] + \mathbb{E}[H^2] + r} \mid h = H^1 + H^2 \in \boldsymbol{S}(\Omega) \Big\} \ (r > 0).
\end{aligned}$$

First, we try to bound the expectation of the two normalized empirical processes. Applying the Symmetrization Lemma A.1, we can first bound the two expectations as:

$$\sup_{\tilde{h}_1 \in \tilde{S}_{r,1}(\Omega)} \mathbb{E}_{y'}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{h}_1(y_i')\right] \leq \mathbb{E}_{y'}\left[\sup_{H^1 \in S_1(\Omega)}\Big|\frac{1}{N}\sum_{i=1}^{N}\frac{h_1(y_i') - \mathbb{E}[H^1]}{\mathbb{E}[H^1]+\mathbb{E}[H^2]+r}\Big|\right] \leq 2R_N(\hat{\boldsymbol{S}}_{r,1}(\Omega)),$$

$$\sup_{\tilde{h}_2 \in \tilde{S}_{r,2}(\Omega)} \mathbb{E}_{y}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_2(y_j)\right] \leq \mathbb{E}_{y}\left[\sup_{H^2 \in S_2(\Omega)}\Big|\frac{1}{n}\sum_{i=1}^{n}\frac{h_2(y_j) - \mathbb{E}[H^2]}{\mathbb{E}[H^1]+\mathbb{E}[H^2]+r}\Big|\right] \leq 2R_n(\hat{\boldsymbol{S}}_{r,2}(\Omega)).$$

where the function classes $\hat{\boldsymbol{S}}_{r,k}(\Omega)$ $(1 \leq k \leq 2)$ are defined as:

$$\hat{\boldsymbol{S}}_{r,1}(\Omega) := \big\{\hat{h}_1(x) := \frac{H^1(x)}{\mathbb{E}[H^1]+\mathbb{E}[H^2]+r} \mid h = H^1 + H^2 \in \boldsymbol{S}(\Omega)\big\},$$

$$\hat{\boldsymbol{S}}_{r,2}(\Omega) := \big\{\hat{h}_2(x) := \frac{H^2(x)}{\mathbb{E}[H^1]+\mathbb{E}[H^2]+r} \mid h = H^1 + H^2 \in \boldsymbol{S}(\Omega)\big\},$$

Applying the Peeling Lemma to any function $h = (h_1, h_2) \in \boldsymbol{S}(\Omega)$ helps us upper bound the sum of the two local Rademacher complexities $R_N(\hat{\boldsymbol{S}}_{r,1}(\Omega)) + R_n(\hat{\boldsymbol{S}}_{r,2}(\Omega))$ with the function ϕ defined in equation

$$\begin{aligned}
R_N(\hat{\boldsymbol{S}}_{r,1}(\Omega)) + R_n(\hat{\boldsymbol{S}}_{r,2}(\Omega)) &= \mathbb{E}_\sigma\left[\mathbb{E}_y\Big[\sup_{h\in\boldsymbol{S}(\Omega)}\frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_i h_1(y_i)}{\mathbb{E}[h]+r}\Big]\right] + \mathbb{E}_\tau\left[\mathbb{E}_{y'}\Big[\sup_{h\in\boldsymbol{S}(\Omega)}\frac{\frac{1}{n}\sum_{j=1}^{n}\tau_j h_2(y_j')}{\mathbb{E}[h]+r}\Big]\right] \\
&= \mathbb{E}_\sigma\left[\mathbb{E}_{y,y'}\Big[\sup_{h\in\boldsymbol{S}(\Omega)}\frac{\frac{1}{N}\sum_{i=1}^{N}\sigma_i h_1(y_i)}{\mathbb{E}[h]+r} + \sup_{h\in\boldsymbol{S}(\Omega)}\frac{\frac{1}{n}\sum_{j=1}^{n}\tau_j h_2(y_j')}{\mathbb{E}[h]+r}\Big]\right] \\
&= R_{N,n}(\hat{\boldsymbol{S}}_r(\Omega)) \leq \frac{4\phi(r)}{r}.
\end{aligned}$$

Combining all inequalities derived above yields:

$$\begin{aligned}
&\sup_{\tilde{h}_1 \in \tilde{S}_{r,1}(\Omega)} \mathbb{E}_{y'}\left[\frac{1}{N}\sum_{i=1}^{N}\tilde{h}_1(y_i')\right] + \sup_{\tilde{h}_2 \in \tilde{S}_{r,2}(\Omega)} \mathbb{E}_{y}\left[\frac{1}{n}\sum_{j=1}^{n}\tilde{h}_2(y_j)\right] \\
&\leq 2R_N(\hat{\boldsymbol{S}}_{r,1}(\Omega)) + 2R_n(\hat{\boldsymbol{S}}_{r,2}(\Omega)) = 2R_{N,n}(\hat{\boldsymbol{S}}_r(\Omega)) \leq \frac{8\phi(r)}{r}\ (r > 0).
\end{aligned} \tag{B.6}$$

Secondly we'll apply the Talagrand concentration inequality to the two function classes $\tilde{S}_{r,1}(\Omega)$ and $\tilde{S}_{r,2}(\Omega)$, which requires us to verify the conditions needed. We will first check that the expectation value $\mathbb{E}[h]$ is always non-negative for any $h \in \boldsymbol{S}(\Omega)$:

$$\begin{aligned}
\mathbb{E}[h] &= \frac{1}{|\Omega|}\int_\Omega |\Omega| \cdot (\frac{1}{2}|\nabla u(x)|^2 + \frac{1}{2}V(x)|u(x)|^2 - f(x)u(x))dx \\
&- \frac{1}{|\Omega|}\int_\Omega |\Omega| \cdot (\frac{1}{2}|\nabla u^\star(x)|^2 + \frac{1}{2}V(x)|u^\star(x)|^2 - f(x)u^\star(x))dx \\
&= \boldsymbol{E}(u) - \boldsymbol{E}(u^\star) \geq 0 \Rightarrow \mathbb{E}[h] \geq 0.
\end{aligned}$$

Next, We will verify that $\tilde{S}_{r,1}(\Omega)$ satisfies all three requirements. At first, we will show that any $\tilde{h}_1 = \frac{\mathbb{E}[H^1]-H^1}{\mathbb{E}[h]+r} \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)$ is of bounded inf-norm. We need to prove that any $H^1 \in \boldsymbol{S}_1(\Omega)$ is of bounded inf-norm beforehand. Using boundedness condition listed in equation B.3 implies:

$$\|H^1\|_\infty = \|\frac{1}{2}\Big(|\nabla u|^2 - |\nabla u^*|^2\Big)\|_\infty \leq \frac{1}{2}\Big(\|\nabla u\|_\infty^2 + \|\nabla u^*\|_\infty^2\Big) \leq C^2.$$

By taking $M_1 := C^2$, we then have $\|H^1\|_\infty \leq M_1$ for all $H^1 \in \boldsymbol{S}_1(\Omega)$. Note that the denominator of $\tilde{h}_1$ can be lower bounded by $|\mathbb{E}[h] + r| \geq r > 0$. Combining these two inequalities help us upper bound the inf-norm $\|\tilde{h}_1\|_\infty = \sup_{x\in\Omega}|\tilde{h}_1(x)|$ as follows:

$$\|\tilde{h}_1\|_\infty = \frac{\|\mathbb{E}[H^1] - H^1\|_\infty}{|\mathbb{E}[h]+r|} \leq \frac{2\|H^1\|_\infty}{r} \leq \frac{2M_1}{r} =: \beta_1.$$

Also, it is easy to check that for any $\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)$, we have

$$\mathbb{E}[\tilde{h}_1] = \frac{\mathbb{E}[H^1] - \mathbb{E}[H^1]}{\mathbb{E}[h] + r} = 0,$$

i.e. any function in the localized class $\tilde{\boldsymbol{S}}_{r,1}(\Omega)$ is of zero mean.
Moreover, we take $\sigma_1^2 = \sup_{\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)} \mathbb{E}[\tilde{h}_1^2]$ to be the upper bound on the second moment of functions in $\tilde{\boldsymbol{S}}_{r,1}(\Omega)$. Now we have verified that any function $\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)$ satisfies all the required conditions. By taking μ to be the uniform distribution on the domain Ω and applying Talagrand's Concentration inequality given in Lemma A.3, we have:

$$\mathbb{P}_x \left[\sup_{\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)} \frac{1}{N} \sum_{i=1}^{N} \tilde{h}_1(x_i) \geq 2 \sup_{\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)} \mathbb{E}_y \Big[\frac{1}{N} \sum_{i=1}^{N} \tilde{h}_1(y_i) \Big] + \sqrt{\frac{2t\sigma_1^2}{N}} + \frac{2t\beta_1}{N} \right] \leq e^{-t}. \tag{B.7}$$

Moreover, We will verify that $\tilde{S}_{r,2}(\Omega)$ also satisfies all three requirements. At first, we will show that any $\tilde{h}_2 = \frac{\mathbb{E}[H^2] - H^2}{\mathbb{E}[h] + r} \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)$ is of bounded inf-norm. We need to prove that any $H^2 \in \boldsymbol{S}_2(\Omega)$ is of bounded inf-norm beforehand. Using boundedness condition listed in equation B.3 implies:

$$\begin{aligned}
\|H^2\|_\infty &= \|\frac{1}{2} V(|u|^2 - |u^*|^2) - f(u - u^*)\|_\infty \\
&\leq \frac{1}{2} V_{\max} \Big(\|u\|_\infty^2 + \|u^*\|_\infty^2 \Big) + \|f\|_\infty \Big(\|u\|_\infty + \|u^*\|_\infty \Big) \\
&\leq \frac{1}{2} V_{\max} \times 2C^2 + 2C^2 = (V_{\max} + 2)C^2.
\end{aligned}$$

By taking $M_2 := (V_{\max} + 2)C^2$, we then have $\|H^2\|_\infty \leq M_2$ for all $H^2 \in \boldsymbol{S}_2(\Omega)$. Note that the denominator of $\tilde{h}_2$ can be lower bounded by $|\mathbb{E}[h] + r| \geq r > 0$. Combining these two inequalities help us upper bound the inf-norm $\|\tilde{h}_2\|_\infty = \sup_{x \in \Omega} |\tilde{h}_2(x)|$ as follows:

$$\|\tilde{h}_2\|_\infty = \frac{\|\mathbb{E}[H^2] - H^2\|_\infty}{|\mathbb{E}[h] + r|} \leq \frac{2\|H^2\|_\infty}{r} \leq \frac{2M_2}{r} =: \beta_2.$$

Also, it is easy to check that for any $\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)$, we have

$$\mathbb{E}[\tilde{h}_2] = \frac{\mathbb{E}[H^2] - \mathbb{E}[H^2]}{\mathbb{E}[h] + r} = 0,$$

i.e. any function in the localized class $\tilde{\boldsymbol{S}}_{r,2}(\Omega)$ is of zero mean.
Moreover, we take $\sigma_2^2 = \sup_{\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)} \mathbb{E}[\tilde{h}_2^2]$ to be the upper bound on the second moment of functions in $\tilde{\boldsymbol{S}}_{r,2}(\Omega)$. Now we have verified that any function $\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)$ satisfies all the required conditions. By taking μ to be the uniform distribution on the domain Ω and applying Talagrand's Concentration inequality given in Lemma A.3, we have:

$$\mathbb{P}_{x'} \left[\sup_{\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)} \frac{1}{n} \sum_{j=1}^{n} \tilde{h}_2(x'_j) \geq 2 \sup_{\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)} \mathbb{E}_{y'} \Big[\frac{1}{n} \sum_{j=1}^{n} \tilde{h}_2(y'_j) \Big] + \sqrt{\frac{2t\sigma_2^2}{n}} + \frac{2t\beta_2}{n} \right] \leq e^{-t}. \tag{B.8}$$

By applying a union bound to the two inequalities derived in B.7 and B.8, we can derive that with probability at least $1 - 2e^{-t}$, the inequality below holds:

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \tilde{h}_1(x'_i) + \frac{1}{n} \sum_{j=1}^{n} \tilde{h}(x_j) &\leq \sup_{\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)} \frac{1}{N} \sum_{i=1}^{N} \tilde{h}_1(x_i) + \sup_{\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)} \frac{1}{n} \sum_{j=1}^{n} \tilde{h}_2(x'_j) \\
&\leq 2 \sup_{\tilde{h}_1 \in \tilde{\boldsymbol{S}}_{r,1}(\Omega)} \mathbb{E}_y \Big[\frac{1}{N} \sum_{i=1}^{N} \tilde{h}_1(y_i) \Big] + \sqrt{\frac{2t\sigma_1^2}{N}} + \frac{2t\beta_1}{N} \\
&+ 2 \sup_{\tilde{h}_2 \in \tilde{\boldsymbol{S}}_{r,2}(\Omega)} \mathbb{E}_{y'} \Big[\frac{1}{n} \sum_{j=1}^{n} \tilde{h}_2(y'_j) \Big] + \sqrt{\frac{2t\sigma_2^2}{n}} + \frac{2t\beta_2}{n} \\
&\leq \frac{16\phi(r)}{r} + \sqrt{\frac{2t}{n}}(\sigma_1 + \sigma_2) + \frac{2t(\beta_1 + \beta_2)}{n}
\end{aligned}$$

By the definition of β_1 and β_2, we have that the term $\frac{2t(\beta_1+\beta_2)}{n}$ can be upper bounded by:

$$\frac{2t(\beta_1+\beta_2)}{n} = \frac{4t(M_1+M_2)}{nr} \leq \frac{4(V_{\max}+3)C^2 t}{nr}$$

Now we will derive some upper bound on the sum $\sigma_1+\sigma_2$. By definition we have that:

$$\begin{aligned}
(\sigma_1+\sigma_2)^2 \leq 2(\sigma_1^2+\sigma_2^2) &= 2\Big[\sup_{\tilde{h}_1\in\tilde{\boldsymbol{S}}_{r,1}(\Omega)} \mathbb{E}[\tilde{h}_1^2] + \sup_{\tilde{h}_2\in\tilde{\boldsymbol{S}}_{r,2}(\Omega)} \mathbb{E}[\tilde{h}_2^2]\Big] \\
&= 2\Big[\sup_{h\in\boldsymbol{S}(\Omega)} \frac{\mathbb{E}[h_1^2]-\mathbb{E}[H^1]^2}{|\mathbb{E}[h]+r|^2} + \sup_{h\in\boldsymbol{S}(\Omega)} \frac{\mathbb{E}[h_2^2]-\mathbb{E}[H^2]^2}{|\mathbb{E}[h]+r|^2}\Big] \\
&\leq 4 \sup_{h\in\boldsymbol{S}(\Omega)} \frac{\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2]}{|\mathbb{E}[h]+r|^2}
\end{aligned}$$

Now it suffices to derive an upper bound of $\frac{\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2]}{|\mathbb{E}[h]+r|^2}$ for any $h\in\boldsymbol{S}(\Omega)$. The existence of such an upper bound is guaranteed because of the regularity results of the PDE. We aim to show that there exist some constants $\alpha,\alpha'>0$, such that for any $h\in\boldsymbol{S}(\Omega)$, the following inequality holds:

$$\alpha(\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2]) \leq \|u-u^*\|_{H^1(\Omega)}^2 \leq \alpha'\mathbb{E}[h]. \tag{B.9}$$

The RHS of the inequality follows from strong convexity of the DRM objective function proved in Section 2.1:

$$\mathbb{E}[h] = \boldsymbol{E}(u) - \boldsymbol{E}(u^*) \geq \frac{\min\{1, V_{\min}\}}{4}\|u-u^*\|_{H^1(\Omega)}^2$$

The LHS of the inequality follows from boundedness condition listed in equation B.3 and the QM-AM inequality:

$$\begin{aligned}
\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2] &= \int_\Omega \frac{1}{4}\Big(|\nabla u|^2-|\nabla u^*|^2\Big)^2 dx + \int_\Omega \left[\frac{1}{2}V(|u|^2-|u^*|^2) - f(u-u^*)\right]^2 dx \\
&\leq \frac{1}{4}\int_\Omega \Big(|\nabla u|^2-|\nabla u^*|^2\Big)^2 dx + \frac{1}{2}\int_\Omega V^2(|u|^2-|u^*|^2)^2 dx + 2\int_\Omega f^2(u-u^*)^2 dx \\
&\leq \frac{1}{4}\int_\Omega \Big||\nabla u|-|\nabla u^*|\Big|^2 (|\nabla u|+|\nabla u^*|)^2 dx + \frac{1}{2}V_{\max}^2 \int_\Omega \Big||u|-|u^*|\Big|^2 (|u|+|u^*|)^2 dx \\
&+ 2C^2\int_\Omega (u-u^*)^2 dx \leq C^2\int_\Omega |\nabla u-\nabla u^*|^2 dx + 2C^2(1+V_{\max}^2)\int_\Omega |u-u^*|^2 dx \\
&\leq 2C^2(1+V_{\max}^2)\|u-u^*\|_{H^1(\Omega)}^2.
\end{aligned}$$

By picking $\alpha' = \frac{4}{\min\{1,V_{\min}\}}$ and $\alpha = \frac{1}{2C^2(1+V_{\max}^2)}$, we have finished proving inequality B.9. Then we can can upper bound the term $\frac{\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2]}{|\mathbb{E}[h]+r|^2}$ as:

$$\frac{\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2]}{|\mathbb{E}[h]+r|^2} \leq \frac{\frac{\alpha'}{\alpha}\mathbb{E}[h]}{2r\mathbb{E}[h]} \leq \frac{\alpha'}{2\alpha r}.$$

Combining the bounds derived above helps us upper bound the term $\sqrt{\frac{2t}{n}}(\sigma_1+\sigma_2)$ as below:

$$\sqrt{\frac{2t}{n}}(\sigma_1+\sigma_2) \leq \sqrt{\frac{8t}{n}}\sqrt{\sup_{h\in\boldsymbol{S}(\Omega)} \frac{\mathbb{E}[h_1^2]+\mathbb{E}[h_2^2]}{|\mathbb{E}[h]+r|^2}} \leq \sqrt{\frac{4\alpha' t}{n\alpha r}}$$

Thus, using the two upper bounds on $\sqrt{\frac{2t}{n}}(\sigma_1+\sigma_2)$ and $\frac{2t(\beta_1+\beta_2)}{n}$, we have

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \tilde{h}_1(x_i') + \frac{1}{n}\sum_{j=1}^n \tilde{h}(x_j) &\leq \frac{16\phi(r)}{r} + \sqrt{\frac{2t}{n}}(\sigma_1+\sigma_2) + \frac{2t(\beta_1+\beta_2)}{n} \\
&\leq \frac{16\phi(r)}{r} + \sqrt{\frac{4\alpha' t}{n\alpha r}} + \frac{4(V_{\max}+3)C^2 t}{nr} = \psi(r)
\end{aligned}$$

Let's pick the threshold radius r_0 to be:

$$r_0 = \max\{2^{14}r^*, \frac{24Mt}{n}, \frac{144\alpha' t}{\alpha n}\}. \tag{B.10}$$

Note that concavity of the function ϕ implies that $\phi(r) \leq r$ for any $r \geq r^*$. Combining this with the first inequality listed in B.4 yields:

$$\frac{16\phi(r_0)}{r_0} \leq \frac{2^{11}\phi(\frac{r_0}{2^{14}})}{2^{14}\frac{r_0}{2^{14}}} = \frac{1}{8} \times \frac{\phi(\frac{r_0}{2^{14}})}{\frac{r_)}{2^{14}}} \leq \frac{1}{8}.$$

On the other hand, applying equation B.10 yields:

$$\sqrt{\frac{4\alpha' t}{n\alpha r_0}} \leq \sqrt{\frac{4\alpha' t}{n\alpha}\frac{\alpha n}{144\alpha' t}} = \frac{1}{6},$$
$$\frac{4(V_{\max}+3)C^2 t}{nr_0} \leq \frac{4(V_{\max}+3)C^2 t}{n} \times \frac{n}{24(V_{\max}+3)C^2 t} = \frac{1}{6}.$$

Summing the three inequalities above implies:

$$\psi(r_0) = \frac{16\phi(r_0)}{r_0} + \sqrt{\frac{4\alpha' t}{n\alpha r_0}} + \frac{4(V_{\max}+3)C^2 t}{nr_0} \leq \frac{1}{8} + \frac{1}{6} + \frac{1}{6} < \frac{1}{2}.$$

By picking $r = r_0$, we can further deduce that for any function $u \in \boldsymbol{F}(\Omega)$, the following inequality holds with probability $1 - e^{-t}$:

$$\frac{\boldsymbol{E}(u) - \boldsymbol{E}(u^*) - \boldsymbol{E}_n(u) + \boldsymbol{E}_n(u^*)}{\boldsymbol{E}(u) - \boldsymbol{E}(u^*) + r_0} = \frac{1}{n}\sum_{i=1}^{n}\tilde{h}(x_i) \leq \psi(r_0) < \frac{1}{2}.$$

Multiplying the denominator on both sides indicates:

$$\Delta\boldsymbol{E}_{\text{gen}} = \boldsymbol{E}(u) - \boldsymbol{E}(u^*) - \boldsymbol{E}_n(u) + \boldsymbol{E}_n(u^*) \leq \frac{1}{2}\Big[\boldsymbol{E}(u) - \boldsymbol{E}(u^*)\Big] + \frac{1}{2}r_0 = \frac{1}{2}\Delta\boldsymbol{E}^{(n)} + \frac{1}{2}r_0.$$

Substituting the upper bound above into the decomposition $\Delta\boldsymbol{E}^{(n)} \leq \Delta\boldsymbol{E}_{\text{gen}} + \frac{3}{2}\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{2n}$ yields that with probability $1 - e^{-t}$, we have:

$$\Delta\boldsymbol{E}^{(n)} \leq \Delta\boldsymbol{E}_{\text{gen}} + \frac{3}{2}\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{2n} \leq \frac{1}{2}\Delta\boldsymbol{E}^{(n)} + \frac{1}{2}r_0 + \frac{3}{2}\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{2n}.$$

Simplifying the inequality above yields that with probability $1 - e^{-t}$, we have:

$$\begin{aligned}\Delta\boldsymbol{E}^{(n)} &\leq r_0 + 3\Delta\boldsymbol{E}_{\text{app}} + \frac{t}{n} = 3\inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\{2^{14}r^*, 24M\frac{t}{n}, \frac{36\alpha'}{\alpha}\frac{t}{n}\} + \frac{t}{n} \\ &\lesssim \inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\Big\{r^*, \frac{t}{n}\Big\}\end{aligned}$$

Moreover, using strong convexity of the DRM objective function proved in Theorem A.1 implies:

$$\Delta\boldsymbol{E}^{(n)} = \boldsymbol{E}(\hat{u}_{\text{MDRM}}) - \boldsymbol{E}(u^*) \gtrsim \|\hat{u}_{\text{MDRM}} - u^*\|^2_{H^1(\Omega)}$$

Combining the two bounds above yields that with probability $1 - e^{-t}$, we have:

$$\|\hat{u}_{\text{MDRM}} - u^*\|^2_{H^1(\Omega)} \lesssim \inf_{u_{\boldsymbol{F}}\in\boldsymbol{F}(\Omega)}\Big(\boldsymbol{E}(u_{\boldsymbol{F}}) - \boldsymbol{E}(u^\star)\Big) + \max\Big\{r^*, \frac{t}{n}\Big\}$$

□

Truncated Fourier Series Estimator. Next we aim to show that the truncated Fourier series estimator can achieve the min-max optimal rate using the MDRM objective function. For any $\xi \in \mathbb{Z}^+$, there exists some Truncated Fourier Series in $\boldsymbol{F}_\xi(\Omega)$ with approximation error $\Delta\boldsymbol{E}_{\text{app}} = O(\xi^{-2(s-2)})$ and generalization error $\Delta\boldsymbol{E}_{\text{gen}} = O(\frac{\xi^d}{n})$

Theorem B.2. *(Final Upper Bound of Modified Deep Ritz Methods with Truncated Fourier Series Estimator) Consider the Modified deep Ritz objective with a plug in Fourier Series estimator* $\hat{u}^{Fourier}_{MDRM} = \min_{u \in \boldsymbol{F}_\xi(\Omega)} \boldsymbol{E}_n(u)$ *with* $\xi = \Theta(n^{\frac{1}{d+2s-4}})$*. We also assume that* $\frac{\xi^d}{N} < \frac{\xi^{d-2}}{n}$*. Then we have*

$$\|\hat{u}^{Fourier}_{MDRM} - u^*\|^2_{H^1} \lesssim n^{-\frac{2s-2}{d+2s-4}}$$

Proof. On the one hand, from Lemma A.6 and Lemma A.7 proved above, we know that the function $\phi(\rho)$ that upper bounds the local Rademacher complexity is dominated by the term $\sqrt{\frac{\rho}{n}}\xi^{\frac{d}{2}}$ for Truncated Fourier Series in $\boldsymbol{F}_\xi(\Omega)$. Thus, the localization radius $\hat{r}$ can be determined as follows:

$$\sqrt{\frac{\rho}{n}}\xi^{\frac{d-2}{2}} + \sqrt{\frac{\rho}{N}}\xi^{\frac{d}{2}} \simeq \rho.$$

For we have assumed $\frac{\xi^d}{N} < \frac{\xi^{d-2}}{n}$, the solution of the fixed point equation is $\hat{r} \simeq \frac{\xi^d}{n}$. On the other hand, by taking $\alpha = s$ and $\beta = 1$ in Lemma A.19 and applying strong convexity of the DRM objective function proved in Section 2.1, we can upper bound the approximation error $\Delta\boldsymbol{E}_{\text{app}}$ as below:

$$\Delta\boldsymbol{E}_{\text{app}} \lesssim \xi^{-2(s-1)},$$

By equating the two terms above, we can solve for ξ that yields the desired bound:

$$\frac{\xi^{d-2}}{n} \simeq \xi^{-2(s-1)} \Rightarrow \xi \simeq n^{\frac{1}{d+2s-4}},$$

Plugging in the expression of ξ gives the final upper bound:

$$\mathbb{E}_{x\sim\mu}[\Delta\boldsymbol{E}_n] \lesssim \hat{r} + \Delta\boldsymbol{E}_{\text{app}} \lesssim \frac{\xi^{d-2}}{n} + \xi^{-2(s-2)} \simeq n^{-\frac{2s-2}{d+2s-4}}.$$

□

C Proof of the Lower Bounds

C.1 Preliminaries on Tools for Lower Bounds

In this section, we repeat the standard tools we use to establish the lower bound. The main tool we use is the Fano's inequailty and the Varshamov-Gilber Lemma.

Lemma C.1 (Fano's methods)**.** *Assume that V is a unifrom random variable over set $\mathcal{V}$, then for any markov chain $V \to X \to \hat{V}$, we always have*

$$\mathcal{P}(\hat{V} \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)}$$

Lemma C.2 (Varshamov-Gillbert Lemma,(Tsybakov, 2008) Theorem 2.9)**.** *Let $D \geq 8$. There exists a subset $\mathcal{V} = \{\tau^{(0)}, \cdots, \tau^{(2^{D/8})}\}$ of $D-$dimensional hypercube $\mathcal{H}^D = \{0,1\}^D$ such that $\tau^{(0)} = (0, 0, \cdots, 0)$ and the ℓ_1 distance between every two elements is larger than $\frac{D}{8}$*

$$\sum_{l=1}^{D} \|\tau^{(j)} - \tau^{(k)}\|_{\ell_1} \geq \frac{D}{8}, \textit{for all } 0 \leq j, k \leq 2^{D/8}$$

C.2 Proof Of Lower Bound

In this section, we provide the proof of the lower bound for learning a PDE. Our proof uses standard Fano method to establish minimax lower bound but finally leads to a non-standard convergence rate. We state standard results for Fano methods in Appendix C.1. Following is the proof our main lower bound.

Theorem C.1 (Lower bound)**.** *We denote $u^*(f)$ to be the solution of the PDE 2.1 and we can aceess randomly sampled data $\{X_i, Y_i\}_{i=1,\cdots,n}$ as described in Section 2.2.*

DRM Lower Bound. *For all estimator* $H : \left(\mathbb{R}^d\right)^{\otimes n} \times \mathbb{R}^{\otimes n} \to W_1^\infty(\Omega)$*, we have*

$$\inf_H \sup_{u \in W_s^\infty(\Omega)} \mathbb{E}\|H(\{X_i, f_i\}_{i=1,\cdots,n}) - u^*(f)\|_{H^1}^2 \gtrsim n^{-\frac{2\alpha-2}{d+2\alpha-4}}. \tag{C.1}$$

PINN Lower Bound. *For all estimator* $H : \left(\mathbb{R}^d\right)^{\otimes n} \times \mathbb{R}^{\otimes n} \to W_1^\infty(\Omega)$*, we have*

$$\inf_H \sup_{u \in W_s^\infty(\Omega)} \mathbb{E}\|H(\{X_i, f_i\}_{i=1,\cdots,n}) - u^*(f)\|_{H^2}^2 \gtrsim n^{-\frac{2\alpha-4}{d+2\alpha-4}}. \tag{C.2}$$

Proof. We construct the following bump function to construct the multiple hypothesis test used for proving the lower bound. Consider a simple C^∞ bump function supported on $[0,1]^d$

$$g(x) = \prod_{i=1}^d \xi(x_i), x = (x_1, \cdots, x_d),$$

where $\xi : \mathbb{R} \to \mathbb{R}$ be a non-zero funtion in $C^\infty(\mathbb{R})$ with support contained in $[0,1]$ and satisfies $\xi(x) \neq 0$, $\frac{d}{dx}\xi(x) \neq 0$. Then $\nabla g(x) \neq 0$ and the support of function g is $[0,1]^d$.

Next, take $m = [n^{\frac{1}{2\alpha-4+d}}]$ and let's consider a regular gird $x^{(j)}, j \in [m]^d$. According to the Varshamov-Gilbert lemma, we can find $2^{m^d/8}$ $(0,1)$-sequences $\tau_1, \cdots, \tau_{2^{m^d/8}} \in \{0,1\}^{m^d}$ such that $\|\tau_k - \tau_{k'}\|^2 \geq \frac{m^d}{8}$ for all $0 < k \neq k' \leq 2^{m^d/8}$. Then we construct the multiple hypothesis as

$$u_k(x) = \sum_{j \in [m_1]^d} \tau_k(j) \frac{\omega}{m^{\alpha+\frac{d}{2}}} g(m(x - x^{(j)})), k = 1, 2, \cdots, 2^{m^d/8},$$

where ω is a constant to be determined later. It is easy to find out that $u_k \in \mathcal{C}^\alpha$.

Then we reduce solving PDE to the multiple hypothesis testing which considers all mappings from n sampled data to the constructed hypothesis $\Psi : \left(\mathbb{R}^d\right)^{\otimes n} \times \mathbb{R}^{\otimes n} \to \mathcal{V} := \{u_i | i = 1, 2, \cdots, 2^{m^d/8}\}$. Then we apply the local Fano method and check that we can obtain a constant lower bound of $\mathcal{P}(\hat{V} \neq V)$ for any estimator $\hat{V}$. Applying the local Fano methods, we know that

$$I(V; X) \leq \frac{1}{|\mathcal{V}|^2} \sum_z \sum_{v \neq v'} D_{KL}(P_v || P_v')$$

where P_k joint distribution of the sampled data (x, y), in specific, x follows a uniform distribution on $[0,1]^d$ and $y = f(X) + \sigma\epsilon$ where ϵ is independently sampled from a standard Gaussian distribution $N(0,1)$. Then we have

$$KL(P_k || P_{k'}) = \mathbb{E}\log(\frac{dP_k}{dP}) = \|\Delta u_k + u_k\|_{L_2}^2 = \frac{C\omega}{m^{2\alpha-4}}$$

Using Fano inequality, when taking ω large enough, we know that

$$\mathcal{P}(\hat{V} \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)} \geq 1 - \frac{\frac{8C\omega}{m^{2\alpha-4}}}{m^d \log 2} \geq 1/2$$

At the same time, we can estimate the separation of the hypothesis in the two different norms:

- Deep Ritz Method:

$$\int_{[0,1]^d} \|\nabla u_k - \nabla u_{k'}\|^2 dx = \frac{\kappa^2}{m^{2\alpha-2+d}} \sum_{j \in [m]^d} (\tau_j^{(k)} - \tau_j^{(k')})^2 \int_{\mathbb{R}^d} \|\nabla g(x)\|^2 dx \gtrsim \frac{1}{m^{2\alpha-2}}$$

- Physic Informed Neural Network:

$$\int_{[0,1]^d} \|\Delta u_k - \Delta u_{k'}\|^2 dx = \frac{\kappa^2}{m^{2\alpha-4+d}} \sum_{j \in [m]^d} (\tau_j^{(k)} - \tau_j^{(k')})^2 \int_{\mathbb{R}^d} \|\Delta g(x)\|^2 dx \gtrsim \frac{1}{m^{2\alpha-4}}$$

Plug in $m = [n^{\frac{1}{2\alpha-4+d}}]$, we know that with constant probability we have

$$\inf_{H} \sup_{u \in W_s^\infty(\Omega)} \mathbb{E}\|H(\{X_i, Y_i\}_{i=1,\cdots,n}) - u^*(f)\|_{H^1}^2 \gtrsim n^{-\frac{2\alpha-2}{d+2\alpha-4}} \log(n)^2, \tag{C.3}$$

$$\inf_{H} \sup_{u \in W_s^\infty(\Omega)} \mathbb{E}\|H(\{X_i, Y_i\}_{i=1,\cdots,n}) - u^*(f)\|_{H^2}^2 \gtrsim n^{-\frac{2\alpha-4}{d+2\alpha-4}} \log(n)^2. \tag{C.4}$$

□

D Intuition Behind the Sub-optimality of the Unmodified Deep Ritz Methods

In this section, we aim to discuss the intuition behind the sub-optimality of the unmodified DRM via using the truncation Fourier basis. To simplify the notation, in this section we consider the following simplest Poisson equation $\Delta u = f$ on the hypercube with zero Dirichlet boundary condition. To illustrate the necessity of the modification we made, we consider the difference between the following two estimators

- **Estimator 1.** We use the truncated Fourier basis estimator to learn the right hand side function f and then we invert the PDE exactly to get the estimated u.
- **Estimator 2.** We plug in a parametrization of the truncated fourier basis into the empirical DRM objective

We would like to point out that *estimator 1 isn't build for computational consideration* but we use it to consider the statistical limit of our sampled data. We first show that the estimator 1 can achieve the minmax optimal estimation error.

Error Of Estimator 1 Firstly, we show that if one wants to learn the function u in H_0^1 norm, one need to learn the right hand side function f in H_0^{-1} norm. The H_0^{-1} norm is defined as the dual norm of the H^1 norm, *i.e.* $\|u\|_{H_0^{-1}} = \max_{\|v\|_{H_0^1} \le 1} \langle u, v \rangle$. Once we assume we have an estimation $\hat{f}$ of f in H_0^{-1}, we can have the estimate of u via $\hat{u} := (\Delta)^{-1} \hat{f}$ in H^1 norm for

$$\begin{aligned}
\|\nabla u - \nabla \hat{u}\|_{H_0^1} &= \max_{\|v\|_{H_0^1} \le 1} \langle \nabla u - \nabla \hat{u}, \nabla v \rangle \\
&= \max_{\|v\|_{H_0^1} \le 1} \langle \Delta u - \Delta \hat{u}, v \rangle \\
&= \max_{\|v\|_{H_0^1} \le 1} \left\langle f - \hat{f}, v \right\rangle = \|f - \hat{f}\|_{H_{-1}}.
\end{aligned}$$

Estimator 1 using the truncated fourier estimator to estimate the right hand side function f. Suppose we can access a random sample of observed data as $\{x_i, f(x_i)\}_{i=1}^n$, then the Fourier coefficient $f_z := \langle u, \phi_z \rangle$ can be estimated as $\hat{f}_z := \frac{1}{n} \sum_{i=1}^n f(x_i)\phi_z(x_i)$. To bound the estimation error of $\hat{f} := \sum_{\|z\|_\infty \le Z} \hat{f}_z \phi_z$ in H_0^{-1}, we first apply the bias-variance decompoiton as

$$\mathbb{E}\|\hat{f} - f\|_{H_0^{-1}}^2 \le \|\mathbb{E}\hat{f} - \hat{f}\|_{H_0^{-1}}^2 + \mathbb{E}\|f - \mathbb{E}\hat{f}\|_{H_0^{-1}}^2$$

We first bound the bias term $\|\mathbb{E}\hat{f} - f\|_{H^{-1}}^2$. We know that $\mathbb{E}\hat{f} = \sum_{\|z\|_\infty \le Z} f_z \phi_z$. Thus for a truncation set Z to be of the from $\mathcal{Z} := \{z \in \mathbb{N}^d | \|z\|_\infty \le Z\}$, the bias can be controlled by

$$\| \sum_{\|z\|_\infty > Z} f_z \phi_z \|_{H^{-1}}^2 \le C \sum_{\|z\|_\infty > Z} f_z^2 z^{-2} \le \|z\|^{-2(s-1)} \|f\|_{H_{\alpha-2}}^2$$

Next we estimate the variance of the estimator, we decompose the variance in to every term

$$\mathbb{E}\|f-\hat{f}\|_{H_{-1}}^2 \leq \mathbb{E}\sum_{\|z\|_\infty\leq Z}(\hat{f}_z-f_z)^2\|\phi_z\|_{H_{-1}}^2 \leq \sum_{\|z\|_\infty\leq Z}|z|^{-1}\text{Var}(\hat{f}_z).$$

Finally we achieve a $Z^{-2(s-1)}+\frac{Z^{d-2}}{n}$ upper bound for estimator 1 and with optimal selection of Z, we can achieve the min-max optimal convergence rate $n^{-\frac{2s-2}{d+2s-4}}$.

Difference Between Estimator 1 and Estimator 2 Next we aim to understand the Deep Ritz Method objective function via plugging in a truncated Fourier series estimator. We consider our estimator $u=\sum\hat{u}_z\phi_z(x)$ lies in the truncated fourier spaces. Then the empirical DRM objective function then can be expressed as

$$\frac{1}{2n}\sum_{i=1}^{n}\left(\sum_z\hat{u}_z\nabla\phi_z(x_i)\right)^2+\sum_z\hat{u}_z\phi_z(x_i)f(x_i). \tag{D.1}$$

We observe that (D.1) is a quadratic formula respect to the Fourier coefficient $\mathrm{u}:=(u_z)_{\|z\|_\infty\leq Z}$ and we can reformulate it in the following matrix form

$$\min\frac{1}{2}\mathrm{u}^\top\hat{A}u+\mathrm{u}^\top\hat{f},\text{ where }\hat{A}=\left(\frac{1}{n}\sum_{i=1}^{n}\nabla\phi_i(x_i)\nabla\phi_j(x_i)\right)_{\|i\|_\infty\leq Z,\|j\|_\infty\leq Z}. \tag{D.2}$$

Based on the matrix formulation D.2 we can compare the solution for the two estimator

- **Estimator 1:** The Fourier coefficient of the solution of Estimator 1 is
$$\hat{\mathrm{u}_1}=\text{diag}\left(\|z\|^2\right)^{-1}_{\|z\|_\infty\leq Z}\hat{f}. \tag{D.3}$$
- **Estimator 2:** The Fourier coefficient of the solution of Estimator 2 is
$$\hat{\mathrm{u}_2}=\hat{A}^{-1}\hat{f}. \tag{D.4}$$

We notice that $\mathbb{E}\hat{A}=\left(\|z\|^2\right)_{\|z\|_\infty\leq Z}$, thus we discovered that we further introduce another variance from the sampling of A. We directly estimate $\hat{u}_1-\hat{u}_2$ and showed this term will larger than the final convergence rate. Notice that

$$\|\hat{u}_1-\hat{u}_2\|_{H^1}^2=f^\top\left((\mathbb{E}\hat{A})^{-1}-\hat{A}^{-1}\right)^\top\text{diag}\left(\|z\|^2\right)_{\|z\|_\infty\leq Z}\left((\mathbb{E}\hat{A})^{-1}-\hat{A}^{-1}\right)f \tag{D.5}$$

Next we aim to bound $\left((\mathbb{E}\hat{A})^{-1}-\hat{A}^{-1}\right)$. We first use Matrix Bernstein Inequality(Tropp, 2015) to bound the H^1 distance between $\hat{u}_1$ and $\hat{u}_2$. Using Matrix Bernstein Inequality, we know with high probability $1-e^{-t}$

$$\left\|\left((\mathbb{E}\hat{A})-\hat{A}\right)\right\|_{\mathrm{H}}\leq\sqrt{\frac{Z^d}{n}}+\frac{t}{n}, \tag{D.6}$$

where $\|\cdot\|_{\mathrm{H}}$ is the matrix operator norm respect to the vector $\|\cdot\|_{\mathrm{H}}$ defined as $\|z\|_{\mathrm{H}}^2=z^\top\text{diag}\left(\|z\|^2\right)^{-1}_{\|z\|_\infty\leq Z}z$. Note that

$$\left(I+(\mathbb{E}\hat{A})^{-1}\left(\hat{A}-(\mathbb{E}\hat{A})\right)\right)\left((\mathbb{E}\hat{A})^{-1}-\hat{A}^{-1}\right)=(\mathbb{E}\hat{A})^{-1}\left(\hat{A}-(\mathbb{E}\hat{A})\right)(\mathbb{E}\hat{A})^{-1} \tag{D.7}$$

When n is large enough, we know that $\frac{1}{2}I\leqslant I+(\mathbb{E}\hat{A})^{-1}\left(\hat{A}-(\mathbb{E}\hat{A})\right)\leqslant I$ with high probability. Thus the term $\|\hat{u}_1-\hat{u}_2\|_{H^1}^2$ is at the scale of $\left\|\left((\mathbb{E}\hat{A})-\hat{A}\right)\right\|_{\mathrm{H}}^2\approx\frac{Z^d}{n}$, the same magnitude as what we get from the empirical process approach in our main proof and is large than the $\frac{Z^{d-2}}{n}$ variance term for $\hat{u}_1$. Thus here we conjecture that the our bound for DRM itself is tight and leads to the sub-optimal convergence rate.