CONVERGENCE RATE ANALYSIS FOR DEEP RITZ METHOD

CHENGUANG DUAN 1, YULING JIAO 1, YANMING LAI 5, XILIANG LU 5, AND JERRY ZHIJIAN YANG 1

Abstract. Using deep neural networks to solve PDEs has attracted a lot of attentions recently. However, why the deep learning method works is falling far behind its empirical success. In this paper, we provide a rigorous numerical analysis on deep Ritz method (DRM) [48] for second order elliptic equations with Neumann boundary conditions. We establish the first nonasymptotic convergence rate in $H^1$ norm for DRM using deep networks with ReLU$^2$ activation functions. In addition to providing a theoretical justification of DRM, our study also shed light on how to set the hyper-parameter of depth and width to achieve the desired convergence rate in terms of number of training samples. Technically, we derive bound on the approximation error of deep ReLU$^2$ network in $H^1$ norm and bound on the Rademacher complexity of the non-Lipschitz composition of gradient norm and ReLU$^2$ network, both of which are of independent interest.

1. Introduction. Partial differential equations (PDEs) have broad applications in physics, chemistry, biology, geology and engineering. A great deal of efforts have been devoted to studying numerical methods for solving PDEs [5, 7, 22, 44, 16]. However, it is still a challenging task to develop numerical scheme for solving PDEs in high-dimension. Due to the success of deep learning for high-dimensional data analysis in computer vision and natural language processing, people have been paying more attention to using (deep) neural network to solve PDEs in high dimension with may be complex domain, an idea that goes back to 1990’s [21, 19]. In the last few years, there are growing literatures on neural network based numerical methods for PDEs. These works can be roughly classified into two categories.

In the first category, deep neural networks are used to improve classical methods. [10] designs a neural network to estimate artificial viscosity in discontinuous Galerkin schemes, see also [6]. [32] trains a neural network serving as a troubled-cell indicator in high-resolution schemes for conservation laws. [42] proposes a universal discontinuity detector using convolution neural network and applies it in conjunction of solving nonlinear conservation. [47] uses reinforcement learning to find new and potentially better data-driven solvers for conservation laws.

In the second category, deep neural networks are utilized to approximate the solution of the PDEs directly. Being benefit from the excellent approximation power of deep neural networks and SGD training, these methods have been successfully applied to solve PDEs in high-dimension. [3, 9] convert nonlinear parabolic PDEs into backward stochastic differential equations and solve them by deep neural networks, which can deal with high-dimensional problems.
Methods based on the strong form of PDEs [31, 40] are also proposed. In [31], physics-informed neural networks (PINNs) use the squared residuals on the domain as the loss function and treat boundary conditions as penalty term. There are several extensions of PINNs for different types of PDEs, including fractional PINNs [30], nonlocal PINNs [29], conservative PINNs [18], eXtended PINNs [17], among others. A similar method presented in [25] proposes a residual-based adaptive refinement method to improve the training efficiency.

In contrast to minimizing squared residuals of strong form, a natural alternative approach to derive loss functions are based on the variational form of PDEs [48, 51]. Inspired by Ritz method, [48] proposes deep Ritz method (DRM) to solve variational problems arising from PDEs. The idea of Galerkin method has also been used in [51], where, they propose a deep Galerkin method (DGM) via reformulating the problem of finding the weak solution of PDEs into an operator norm minimization problem induced by the weak formulation.

1.1. Related works and contributions. Although there are great empirical achievements in recent years as mentioned above, a challenging and interesting question is that can we give rigorous analysis to guarantee their performances as people has done in the classical counterpart such as finite element method (FEM) [7] and finite difference method [22]? Several recent efforts have been devoted to making processes along this line. [26] consider the optimization and generalization error of second-order linear PDEs with two-layer neural networks in the scenario of over-parametrization. [37, 27, 38] study the convergence of PINNs with deep neural networks. When we were about to finish our draft, we aware that [24] give an error analysis that focuses on analyzing one hidden layer shallow networks with ReLU-Cosine activation functions to solve elliptic PDEs whose solutions are restricted to spectral Barron space, see also [49] for handling general equations with solutions living in spectral Barron space via two layer ReLU-k networks.

Two important questions have not been addressed in the above mentioned related study are those: what is the influence of the topological structure of the networks, say the depth and width, in the quantitative error analysis? How to determine these hyper-parameters to achieve a desired convergence rate? In this paper, we give a firm answers on these questions by studying convergence rate of the deep Ritz method to solve second order elliptic equations with Neumann boundary conditions by using ReLU2 networks with arbitrary depth. As far as we know, we establish the first nonasymptotic bound on DRM. The main contributions of this paper are summarized as follows.

- We derive a bound on the approximation error of deep ReLU2 network in $H^1$ norm, which is of independent interest, i.e., we prove that for any $u^* \in H^2(\Omega)$, there exist a ReLU2 network $\bar{u}_{\phi}$ with depth $\mathcal{D} \leq \lceil \log_2 d \rceil + 3$, width $\mathcal{W} \leq \mathcal{O}(4d\lceil 1 - 4^{\frac{1}{d}} \rceil^d)$ such that

$$\|u^* - \bar{u}_{\phi}\|_{H^1(\Omega)} \leq \epsilon.$$ 

- We establish a bound on the statistical error in DRM with the tools of Pseudo dimension, especially we give an bound on

$$\mathbb{E}_{Z_i, \sigma_i, i=1,...,n} \left[ \sup_{u_{\phi} \in \mathcal{U}} \left\{ \frac{1}{n} \sum_i \sigma_i \|\nabla u_{\phi}(Z_i)\|_2^2 \right\} \right].$$
i.e., the Rademacher complexity of the non-Lipschitz composition of gradient norm and ReLU² network, via calculating the Pseudo dimension of networks with both ReLU and ReLU² activation functions. We believe that the technique we used here is helpful for bounding the statistical errors for other deep PDEs solvers where the Rademacher complexity of non-Lipschitz composition is hard to handle.

• Based on the above to error bounds we establish the first nonasymptotic convergence rate of deep Ritz method. We prove that if we set the depth and width in ReLU² networks to be

\[ D \leq \lceil \log_2 d \rceil + 3, \quad W \leq O\left( 4d^{\lceil n+2+\nu \rceil} \right), \]

the \( H^1 \) norm error of DRM in expectation is

\[ O\left( n^{-1/(2d+4+\nu)} \right), \]

where \( n \) is the number of training samples on both the domain and the boundary, \( \nu \) is a positive number that can be an arbitrary small. Our theory sheds lights on choosing the topological structure of the employed networks to achieve the desired convergence rate in terms of number of training samples.

• By comparing the known results in nonparametric regression, where the optimal convergence rate in \( H^1 \) norm for estimating functions in \( H^2 \) with \( n \) paired samples is \( O(n^{-1/(d+4)}) \) [41], we conjecture that the optimal convergence rate of DRM in \( H^1 \) norm is also \( O(n^{-1/(d+4)}) \).

The rest of the paper are organized as follows. In Section 2, we give some preliminaries. In Section 3, we present the detail analysis on the convergence rate of DRM. We give conclusion and short discussion in Section 4.

2. Preliminaries. Consider the following elliptic equation with Neumann boundary conditions

\[
\begin{aligned}
-\Delta u + wu &= f \text{ in } \Omega \\
\frac{\partial u}{\partial n} &= g \text{ on } \partial \Omega,
\end{aligned}
\]

(2.1)

where, \( \Omega \) is a bounded open subset of \( \mathbb{R}^d, d > 1 \), \( f(x) \in L^2(\Omega), w(x) \in L^\infty(\Omega) \) satisfying \( w(x) \geq c_1 > 0 \) a.e., and \( g(s) \in L^2(\partial \Omega) \). Without loss of generality we assume \( \Omega = (0,1)^d \).

Define

\[
\mathcal{L}(u) = \frac{1}{2} ||u||_{H^1(\Omega)}^2 + \frac{1}{2} ||u||_{L^2(\Omega,w)}^2 - \langle u, f \rangle_{L^2(\Omega)} - \langle Tu, g \rangle_{L^2(\partial \Omega)},
\]

(2.2)

where \( T \) is the trace operator.

**Lemma 2.1.** The unique weak solution \( u^* \in H^1(\Omega) \) of (2.1) is the unique minimizer of \( \mathcal{L}(u) \) over \( H^1(\Omega) \). Moreover, \( u^* \in H^2(\Omega) \).

**Proof.** Well known results, see for example [13]. \( \square \)
A function $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ implemented by a neural network is defined by

$$
 f_0(x) = x,
 f_{\ell}(x) = g_\ell (A_\ell f_{\ell-1} + b_\ell) \quad \text{for } \ell = 1, \ldots, L - 1,
 f = f_L(x) := A_L f_{L-1} + b_L,
$$

where $A_\ell \in \mathbb{R}^{N_{\ell} \times N_{\ell-1}}$, $b_\ell \in \mathbb{R}^{N_\ell}$ and the activation function $g_\ell$ is understood to act component-wise (note that here we allow different activation functions in different layers). $L$ is called the depth of the network and $\max\{N_\ell, \ell = 0, \ldots, L\}$ is called the width of the network. We will use $\mathcal{L}$ and $\mathcal{W}$ to denote the depth and width of neural networks $f$, respectively. $\sum_{\ell=1}^{L} N_\ell$ is called the number of units of $f$ and $\phi = \{A_\ell, b_\ell\}_{\ell}$ are called the weight parameters. For simplicity we also use $f_\phi$ to refer to the network. We use $\mathcal{N}^2_{D, \mathcal{W}, \mathcal{B}}$ to denote the set of neural networks with depth $D$, width $\mathcal{W}$, output bounded by $\mathcal{B}$, activation function $\text{ReLU}(x) = \max\{0, x\}$, and $\mathcal{N}^{1,2}_{D, \mathcal{W}, \mathcal{B}}$ as the set of neural networks with depth $D$, width $\mathcal{W}$, output bounded by $\mathcal{B}$, activation functions $\text{ReLU}(x) = \max\{0, x\}$ and $\text{ReLU}^2(x) = \max\{0, x^2\}$.

Obviously,

$$
\mathcal{L}(u) = |\Omega|E_{X \sim U(\Omega)}[\|\nabla u(X)\|^2/2 + w(X)u^2(X)/2 - u(X)f(X)]
\quad - |\partial \Omega|E_{Y \sim U(\partial \Omega)}[Tu(Y)g(Y)],
$$

where $U(\Omega)$, $U(\partial \Omega)$ are the uniform distributions on $\Omega$ and $\partial \Omega$, respectively. The main idea of deep Ritz method (DRM) [48] is employing a $u_\phi \in \mathcal{N}^2 := \mathcal{N}^2_{D, \mathcal{W}, \mathcal{B}}$ to approximate the minimizer $u^*$ of $\mathcal{L}$, i.e., finding $u_\phi$ such that $\mathcal{L}(u_\phi)$ closes to $\mathcal{L}(u^*)$. To this end, by Lemma 2.1, one may consider the following empirical loss minimization problem

$$
\hat{u}_\phi \in \min_{u_\phi \in \mathcal{N}^2} \hat{\mathcal{L}}(u_\phi), \quad (2.3)
$$

where,

$$
\hat{\mathcal{L}}(u_\phi) = \frac{|\Omega|}{N} \sum_{i=1}^{N} \frac{\|\nabla u_\phi(X_i)\|^2}{2} + \frac{w(X_i)u^2_\phi(X_i)}{2} - u_\phi(X_i)f(X_i)
\quad - \frac{|\partial \Omega|}{M} \sum_{j=1}^{M} [u_\phi(Y_j)g(Y_j)], \quad (2.4)
$$

is a discrete version of the functional $\mathcal{L}(u_\phi)$ with $\{X_i\}_{i=1}^{N}$ being identically and independently distributed (i.i.d.) according to $U(\Omega)$, $\{Y_j\}_{j=1}^{M}$ being identically and independently drawn from $U(\partial \Omega)$. Then, we call a (random) solver $A$, say SGD, to minimize (2.3) and denote the output of $A$, say $u_{\phi, A} \in \mathcal{N}$, as the final solution.

3. Error Analysis. In this section we prove the convergence rate analysis for DRM with deep ReLu^2 networks. The following Lemma play an important role by decoupling the total errors into three types of errors.
Lemma 3.1.

\[ \|u_{\phi, \lambda} - u^*\|_{H^1(\Omega)}^2 \leq \frac{2}{c_1 \wedge 1} \left( \frac{\|w\|_{L^\infty(\Omega)} \vee 1}{2} \inf_{u \in \mathcal{N}^2} \|\tilde{u} - u^*\|_{H^1(\Omega)}^2 + 2 \sup_{u \in \mathcal{N}^2} \|\mathcal{L}(u) - \mathcal{L}(u)\| + \|\mathcal{L}(u_{\phi, \lambda}) - \mathcal{L}(\tilde{u})\|. \right) \]

Proof. For any \( \tilde{u} \in \mathcal{N}^2 \), we have

\[ \mathcal{L}(u_{\phi, \lambda}) - \mathcal{L}(u^*) \]

\[ = \mathcal{L}(u_{\phi, \lambda}) - \tilde{\mathcal{L}}(u_{\phi, \lambda}) + \tilde{\mathcal{L}}(\tilde{u} - u^*) \]

\[ + \tilde{\mathcal{L}}(\tilde{u}) = \mathcal{L}(\tilde{u}) - \mathcal{L}(u^*) \]

where the last step is due to the fact that \( \tilde{\mathcal{L}}(\tilde{u}) - \tilde{\mathcal{L}}(\tilde{u}) \leq 0 \). Since \( \tilde{u} \) can be any element in \( \mathcal{N}^2 \), we take the infimum of \( \tilde{u} \) on both sides of the above display,

\[ \mathcal{L}(u_{\phi, \lambda}) - \mathcal{L}(u^*) \leq \inf_{u \in \mathcal{N}^2} \left[ \mathcal{L}(\tilde{u}) - \mathcal{L}(u^*) \right] + 2 \sup_{u \in \mathcal{N}^2} \left| \mathcal{L}(u) - \tilde{\mathcal{L}}(\tilde{u}) \right| + \left[ \tilde{\mathcal{L}}(u_{\phi, \lambda}) - \tilde{\mathcal{L}}(\tilde{u}) \right]. \]

Now for any \( u \in \mathcal{N}^2 \), set \( v = u - u^* \), then

\[ \mathcal{L}(u) = \mathcal{L}(u^* + v) \]

\[ = \frac{1}{2} \langle \nabla(u^* + v), \nabla(u^* + v) \rangle_{L^2(\Omega)} + \frac{1}{2} \langle u^* + v, u^* + v \rangle_{L^2(\Omega; w)} \]

\[ - \langle u^* + v, f \rangle_{L^2(\Omega)} - T(u^* + v, g)_{L^2(\Omega)} \]

\[ = \frac{1}{2} \langle \nabla u^*, \nabla u^* \rangle_{L^2(\Omega)} + \frac{1}{2} \langle u^*, u^* \rangle_{L^2(\Omega; w)} - \langle u^*, f \rangle_{L^2(\Omega)} - \langle T(u^*), g \rangle_{L^2(\Omega)} \]

\[ + \frac{1}{2} \langle \nabla v, \nabla v \rangle_{L^2(\Omega)} + \frac{1}{2} \langle v, v \rangle_{L^2(\Omega; w)} \]

\[ + \left[ \langle \nabla u^*, \nabla v \rangle_{L^2(\Omega)} + \langle u^*, v \rangle_{L^2(\Omega; w)} - \langle v, f \rangle_{L^2(\Omega)} - \langle T(v), g \rangle_{L^2(\Omega)} \right] \]

\[ = \mathcal{L}(u^*) + \frac{1}{2} \langle \nabla v, \nabla v \rangle_{L^2(\Omega)} + \frac{1}{2} \langle v, v \rangle_{L^2(\Omega; w)}, \]

where the last equality is due to the fact that \( u^* \) is the weak solution of equation (2.1). Hence

\[ \frac{c_1 \wedge 1}{2} \|v\|_{H^1(\Omega)}^2 \leq \mathcal{L}(u) - \mathcal{L}(u^*) = \frac{1}{2} \langle \nabla v, \nabla v \rangle_{L^2(\Omega)} + \frac{1}{2} \langle v, v \rangle_{L^2(\Omega; w)} \]

\[ \leq \frac{\|w\|_{L^\infty(\Omega)} \vee 1}{2} \|v\|_{H^1(\Omega)}^2, \]

that is,

\[ \frac{c_1 \wedge 1}{2} \|u - u^*\|_{H^1(\Omega)}^2 \leq \mathcal{L}(u) - \mathcal{L}(u^*) \leq \frac{\|w\|_{L^\infty(\Omega)} \vee 1}{2} \|u - u^*\|_{H^1(\Omega)}^2. \]
Combining (3.1) and (3.2) yields
\[
\|u_{\phi,A} - u^*\|_{\mathcal{H}^1(\Omega)}^2 \leq \frac{2}{c_1^2} \|w\|_{L^\infty(I)} \vee 1 \inf_{u \in N^2} \|\bar{u} - u^*\|_{\mathcal{H}^1(\Omega)}^2 + 2 \sup_{u \in N^2} \left|\mathcal{L}(u) - \widehat{\mathcal{L}}(\bar{u})\right| + \left[\widehat{\mathcal{L}}(u_{\phi,A}) - \widehat{\mathcal{L}}(\bar{u}_\phi)\right].
\]

\]

The approximation error \(E_{\text{app}}\) describes the expressive power of the ReLU\(^2\) networks \(N^2\) in \(H^1\) norm, which corresponds to the approximation error in FEM known as the Céa’s lemma [7]. The statistical error \(E_{\text{sta}}\) is caused by the Monte Carlo discretization of \(\mathcal{L}(\cdot)\) defined in (2.2) with \(\widehat{\mathcal{L}}(\cdot)\) in (2.4). While, the optimization error \(E_{\text{opt}}\) indicates the performance of the solver \(A\) we utilized. In contrast, this error is corresponding to the error of solving the linear systems in FEM. In this paper we focus on the first two errors, i.e., considering the scenario of perfect training with \(E_{\text{opt}} = 0\).

\[3.1.\, \text{Approximation error.}\] The current literature on network approximation theory are mainly focused on the \(L^p, p \in [1, +\infty]\) norm for deep networks [50, 43, 36, 28, 33, 23, 39]. The approximation error of ReLU network in Sobolev norm are considered in [15, 12]. However, the ReLU network may not be suitable for solving PDEs since the term \(\nabla u_{\phi}\) in the loss function will become piece-wise constant with respect to \(\phi\), which will prohibit using SGD for training. In this section we derive an upper bound on the approximation error ReLU\(^2\) networks \(N^2\) in \(H^1\) norm, which is of independent interest.

**Theorem 3.1.** Assume \(\|u^*\|_{\mathcal{H}^2(I)} \leq c_2\), then there exist an ReLU\(^2\) network \(\bar{u}_\bar{\phi} \in N^2\) with
\[
D \leq \lceil \log_2 d \rceil + 3, \quad W \leq 4d \left[ \frac{C c_2}{\epsilon} - 4 \right]^d
\]
such that
\[
E_{\text{app}} \leq \frac{\|w\|_{L^\infty(I)} \vee 1}{2} \|u^* - \bar{u}_\bar{\phi}\|_{\mathcal{H}^1(I)}^2 \leq \frac{\|w\|_{L^\infty(I)} \vee 1}{2} \epsilon^2,
\]
where \(C\) is a constant depending only on \(d\).

**Proof.** Our proof is based on some classical approximation results of B-splines [35, 8]. Let us recall some notation and useful results. We denote by \(\pi_t\) the dyadic partition of \([0, 1]\), i.e.,
\[
\pi_t : t^{(l)}_0 = 0 < t^{(l)}_1 < \cdots < t^{(l)}_{2^l-1} < t^{(l)}_{2^l} = 1,
\]
where \(t^{(l)}_i = i \cdot 2^{-l}(0 \leq i \leq 2^l)\). The cardinal B-spline of order 3 with respect to partition \(\pi_t\) is defined by
\[
N_{i,3}^{(3)}(x) = (-1)^k [t^{(l)}_i, \ldots, t^{(l)}_{i+3}, (x-t^{(l)}_i) \cdot (t^{(l)}_{i+3} - t^{(l)}_i)], \quad i = -2, \ldots, 2^l - 1
\]
which can be rewritten in the following equivalent form,

\[
N_{l,i}^{(3)}(x) = 2^{2l-1} \sum_{j=0}^{3} (-1)^j \left( \frac{3}{j} \right) (x - i2^{-l} - j2^{-l})^2, \quad i = -2, \ldots, 2^l - 1.
\] (3.3)

The multivariate cardinal B-spline of order 3 is defined by the product of univariate cardinal B-splines of order 3, i.e.,

\[
N_{l,i}^{(3)}(x) = \prod_{j=1}^{d} N_{l,i_j}^{(3)}(x_j), \quad 1 = (i_1, \ldots, i_d), -3 < i_j < 2^l.
\]

Denote

\[
S_{l}^{(3)}([0,1]^d) = \text{span}\{N_{l,i}^{(3)}, -3 < i_j < 2^l, j = 1, 2, \ldots, d\}.
\]

Then, the element \( f \) in \( S_{l}^{(3)}([0,1]^d) \) are piecewise polynomial functions according to to partition \( \pi^d_l \) with each piece being degree 2 and in \( C^1([0,1]^d) \). Since

\[
S_{1}^{(3)} \subset S_{2}^{(3)} \subset S_{3}^{(3)} \subset \cdots,
\]

We can further denote

\[
S^{(3)}([0,1]^d) = \bigcup_{l=1}^{\infty} S_{l}^{(3)}([0,1]^d).
\]

The following approximation result of cardinal B-splines in Sobolev spaces which is a direct consequence of theorem 3.4 in [34] play an important role in the proof of this Theorem.

**Lemma 3.2.** Assume \( u^* \in H^2([0,1]^d) \), there exists \( \{c_j\}_{j=1}^{(2^l-4)^d} \subset \mathbb{R} \) with \( l > 2 \) such that

\[
\|u^* - \sum_{j=1}^{(2^l-4)^d} c_j N_{l,i_j}^{(3)}\|_{H^1(\Omega)} \leq \frac{C}{2^l}\|u^*\|_{H^1(\Omega)},
\]

where \( C \) is a constant only depend on \( d \).

**Lemma 3.3.** The multivariate B-spline \( N_{l,i}^{(3)}(x) \) can be implemented exactly by a ReLU^2 network with depth \( \lceil \log_2 d \rceil + 2 \) and width \( 4d \).

**Proof.** Denote

\[
\sigma(x) = \begin{cases} 
  x^2, & x \geq 0 \\
  0, & \text{else}
\end{cases}
\]

as the activation function in ReLU^2 network. By definition of \( N_{l,i}^{(3)}(x) \) in (3.3), it’s clear that \( N_{l,i}^{(3)}(x) \) can be implemented by ReLU^2 network without any error with depth 2 and width 4. On the other hand ReLU^2 network can also realize multiplication without any error. In fact, for any \( x, y \in \mathbb{R} \),

\[
xy = \frac{1}{4}[(x+y)^2 - (x-y)^2] = \frac{1}{4}[\sigma(x+y) + \sigma(-x-y) - \sigma(x-y) - \sigma(y-x)].
\]
Hence multivariate B-spline of order 3 can be implemented by ReLU network exactly with depth $\lceil \log_2 d \rceil + 2$ and width $4d$.

For any $\epsilon > 0$, by Lemma 3.2 and 3.3 with $\frac{1}{d} \leq \left[ \frac{C\|u^*\|_{L^2}}{\epsilon} \right]$, there exists $\bar{u}_\phi \in \mathcal{N}^2$, such that
\[
\|u^* - \bar{u}_\phi\|_{H^1(\Omega)} \leq \epsilon.
\]
(3.4)

The depth $D$ and width $W$ of $\bar{u}_\phi$ are satisfying $D \leq \lceil \log_2 d \rceil + 3$ and $W \leq 4d \left[ \frac{C\|u^*\|_{L^2}}{\epsilon} - 4 \right]^d$, respectively.

### 3.2. Statistical error.

In this section, we bound the statistical error $E_{sta} = 2 \sup_{u \in \mathcal{N}^2} |\mathcal{L}(u) - \hat{\mathcal{L}}(u)|$.

**Lemma 3.4.**
\[
\sup_{u \in \mathcal{N}^2} |\mathcal{L}(u) - \hat{\mathcal{L}}(u)| \leq \sum_{j=1}^{d} \sup_{u \in \mathcal{N}^2} |\mathcal{L}_j(u) - \hat{\mathcal{L}}_j(u)|,
\]
where,
\[
\mathcal{L}_1(u) = |\Omega|E_{X \sim U(\Omega)}[w(X)u^2(X)/2], \quad \hat{\mathcal{L}}_1(u) = \frac{|\Omega|}{N} \sum_{i=1}^{N} \frac{w(X_i)u^2(X_i)}{2},
\]
\[
\mathcal{L}_2(u) = |\Omega|E_{X \sim U(\Omega)}[u(X)f(X)], \quad \hat{\mathcal{L}}_2(u) = \frac{|\Omega|}{N} \sum_{i=1}^{N} [u(X_i)f(X_i)],
\]
\[
\mathcal{L}_3(u) = |\partial\Omega|E_{Y \sim U(\partial\Omega)}[Tu(Y)g(Y)], \quad \hat{\mathcal{L}}_3(u) = \frac{|\partial\Omega|}{M} \sum_{j=1}^{M} [u(Y_j)g(Y_j)],
\]
\[
\mathcal{L}_4(u) = |\Omega|E_{X \sim U(\Omega)}[\|\nabla u(X)\|_{L^2}^2/2], \quad \hat{\mathcal{L}}_4(u) = \frac{|\Omega|}{N} \sum_{i=1}^{N} \frac{\|\nabla u(X_i)\|_{L^2}^2}{2}.
\]

**Proof.** This Lemma holds by the direct consequence of triangle inequality.

By Lemma 3.4, we have to bound the the maximum value of four random processes indexed by $u \in \mathcal{N}^2$. To this end, we recall tools in empirical process [45, 20]. Denote
\[
\mathcal{B} = \max\{\|\bar{u}_\phi\|_{L^\infty(\Omega)}, \|\nabla \bar{u}_\phi\|_{L^2(\Omega)}^2\},
\]
(3.5)

where $\bar{u}_\phi$ is the best approximation of $u^*$ in Theorem 3.1. Let $f, g, w$ be bounded, say by some
constant $c_3$, i.e, we assume that
\[ \|f\|_{L^\infty(\Omega)} \vee \|w\|_{L^\infty(\Omega)} \vee \|g\|_{L^\infty(\partial \Omega)} \vee B \leq c_3 < \infty. \]

We use $\mu$ to denote $U(\Omega)(U(\partial \Omega))$. Given $n = N(M)$ i.i.d samples $Z_n = \{Z_i\}_{i=1}^n$ from $\mu$, with $Z_i = X_i(Y_i) \sim \mu$, we need the following Rademacher complexity to measure the capacity of the given function class $\mathcal{N}$ restricted on $n$ random samples $Z_n$.

**Definition 3.5.** The Rademacher complexity of a set $A \subseteq \mathbb{R}^n$ is defined as
\[ \mathcal{R}(A) = \mathbb{E}_{Z_n, \Sigma_n} \left[ \sup_{u \in A} \frac{1}{n} \sum_{i=1}^n \sigma_i u_i \right], \]
where, $\Sigma_n = \{\sigma_i\}_{i=1}^n$ are $n$ i.i.d Rademacher variables with $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. The Rademacher complexity of function class $\mathcal{N}$ associate with random sample $Z_n$ is defined as
\[ \mathcal{R}(\mathcal{N}) = \mathbb{E}_{Z_n, \Sigma_n} \left[ \sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i u(Z_i) \right]. \]

**Lemma 3.6.** Let $\Psi_1(x, y) = \frac{w(x)\sigma}{2} : \mathbb{R}^d \times \mathbb{R}, |y| \leq c_3, \Psi_2(x, y) = f(x)y : \mathbb{R}^d \times \mathbb{R}, |y| \leq c_3, \Psi_3(x, y) = g(x)y : \mathbb{R}^d \times \mathbb{R}, |y| \leq c_3$. Then $\Psi_1(x, y), \Psi_2(x, y)$ and $\Psi_3(x, y)$ are $c_3^2, c_3$ and $c_3$-Lipschitz continuous on $y$ for all $x$ and $\Psi_j(x, 0) = 0, j = 1, 2, 3$.

**Proof.** We give the proof for $\Psi_1$ and omit the details for $\Psi_2, \Psi_3$ since they can be shown similarly. For arbitrary $y_1, y_2$ with $|y_i| \leq c_3, i = 1, 2$,
\[ |\Psi_1(x, y_1) - \Psi_1(x, y_2)| = \frac{|w(x)y_1^2 - w(x)y_2^2|}{2} = \frac{|w(x)(y_1 + y_2)|}{2} |y_1 - y_2| \leq c_3^2 |y_1 - y_2|. \]

By Corollary 3.17 in [20] and Lemma 3.6, we have the following Lipschitz contraction results on Rademacher complexity.

**Lemma 3.7.** Let $\mathcal{N} = \{u(x) : \|u\|_{L^\infty(\Omega)} \leq c_3\}$. Define,
\[ \Psi_j \circ \mathcal{N} = \{\text{composition of } \Psi_j \text{ and } \mathcal{N} : x \rightarrow \Psi_j(x, u(x)) : u \in \mathcal{N}\}, j = 1, 2, 3. \]

Then, $\mathcal{R}(\Psi_1 \circ \mathcal{N}) \leq 2c_3^2 \mathcal{R}(\mathcal{N}), \mathcal{R}(\Psi_i \circ \mathcal{N}) \leq 2c_3 \mathcal{R}(\mathcal{N}), i = 2, 3$.

The following symmetrization result shows that the Rademacher complexity $\mathcal{R}(\Psi_j \circ \mathcal{N}^2)$ gives upper bound on $\sup_{u \in \mathcal{N}^2} |\mathcal{L}_j(u) - \mathcal{C}_j(u)|, j = 1, ..., 3$.

**Lemma 3.8.** $\mathbb{E}_{Z_n} \left[ \sup_{u \in \mathcal{N}^2} |\mathcal{L}_j(u) - \mathcal{C}_j(u)| \right] \leq \mathcal{R}(\Psi_j \circ \mathcal{N}^2), j = 1, ..., 3$.

**Proof.** We give the proof for $j = 1$ and omit the proof for $\Psi_2$ and $\Psi_3$ since they can be shown similarly. Let $\tilde{Z}_n = \{\tilde{Z}_i\}_{i=1}^n$ be an i.i.d ghost sample from $\mu$ and $\tilde{Z}_n$ is independent of
Dudley’s entropy formula [11].

where, the first inequality follows from the Jensen’s inequality, and the second equality holds since both \( \sigma_i (\frac{w(\tilde{Z}_i)u^2(\tilde{Z}_i)}{2} - \frac{w(Z_i)u^2(Z_i)}{2}) \) and \( (\frac{w(\tilde{Z}_i)u^2(\tilde{Z}_i)}{2} - \frac{w(Z_i)u^2(Z_i)}{2}) \) are governed by the same law, and the last equality holds since the distribution of the two terms are the same.

Next we give a upper bound of \( \mathfrak{R}(N^2) \) in terms of the covering number of \( N^2 \) by using the Dudley’s entropy formula [11].

**Definition 3.9.** Suppose that \( W \subset \mathbb{R}^n \). For any \( \epsilon > 0 \), let \( V \subset \mathbb{R}^n \) be a \( \epsilon \)-cover of \( W \) with respect to the distance \( d_\infty \), that is, for any \( w \in W \), there exists a \( v \in V \) such that \( d_\infty (u, v) < \epsilon \), where \( d_\infty \) is defined by

\[
d_\infty (u, v) := \| u - v \|_\infty.
\]

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

**Definition 3.10.** Suppose that \( \mathcal{N} \) is a class of functions from \( \Omega \) to \( \mathbb{R} \). Given \( n \) sample \( Z_n = (Z_1, Z_2, \cdots, Z_n) \in \Omega^n \), \( \mathcal{N}|_{Z_n} \subset \mathbb{R}^n \) is defined by

\[
\mathcal{N}|_{Z_n} = \{(u(Z_1), u(Z_2), \cdots, u(Z_n)) : u \in \mathcal{N}\}.
\]

The uniform covering number \( C_\infty(\epsilon, \mathcal{N}, n) \) is defined by

\[
C_\infty(\epsilon, \mathcal{N}, n) = \max_{Z_n \in \Omega^n} C(\epsilon, \mathcal{N}|_{Z_n}, d_\infty)
\]

**Lemma 3.11.** Assume \( 0 \in \mathcal{N} \) and the diameter of \( \mathcal{N} \) is less than \( B \), i.e., \( \| u \|_{L_\infty(\Omega)} \leq B, \forall u \in \mathcal{N} \). Then

\[
\mathfrak{R} (\mathcal{N}) \leq \inf_{0 < \delta < B} \left( 4\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{B} \sqrt{\log(2C(\epsilon, \mathcal{N}, n))} \, d\epsilon \right).
\]
Moreover, we denote the best approximate element of $\mathbf{u}$ can be bounded by $\epsilon$. The third term in the above display vanishes. By Hölder’s inequality, we deduce that the first term

$$E_{\mathcal{N}}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_i \sigma_i u_i] \leq \frac{D}{n} \sqrt{2 \log(2|\mathcal{V}|)}.$$  

By definition

$$\mathcal{R}(\mathcal{N}) = \mathcal{R}(\mathcal{N}|\mathbf{z}_n) = E_{\mathbf{z}_n}[E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_i \sigma_i u(Z_i)]|\mathbf{Z}_n].$$

Thus, it suffice to show

$$E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_i \sigma_i u(Z_i)] \leq \inf_{0 \leq \delta \leq \mathcal{B}} \left( 4\delta + \frac{12}{\sqrt{n}} \int_{\delta}^{\mathcal{B}} \sqrt{\log C(\epsilon, \mathcal{N}^2, n)d\epsilon} \right)$$

by conditioning on $\mathbf{Z}_n$. Given an positive integer $K$, let $\epsilon_k = 2^{-k+1}\mathcal{B}$, $k = 1, \ldots, K$. Let $C_k$ be a cover of $\mathcal{N}|\mathbf{z}_n \subseteq \mathbb{R}^n$ whose covering number is denoted as $C(\epsilon_k, \mathcal{N}_n, d_\infty)$. Then, by definition, $\forall u \in \mathcal{N}$, there $\exists c^k \in C_k$ such that

$$d_\infty(u|\mathbf{z}_n, c^k) = \max\{|u(Z_i) - c^k_i|, i = 1, \ldots, n\} \leq \epsilon_k, k = 1, \ldots, K.$$  

Moreover, we denote the best approximate element of $u$ in $C_k$ with respect to $d_\infty$ as $c^k(u)$. Then,

$$E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i u(Z_i)]$$

$$= E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i (u(Z_i) - c^K_i(u))] + \sum_{j=1}^{K-1} \sum_{i=1}^n \sigma_i (c^j_i(u) - c^{j+1}_i(u)) + \sum_{i=1}^n \sigma_i c^1_i(u)]$$

$$\leq E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i (u(Z_i) - c^K_i(u))] + \sum_{j=1}^{K-1} E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i (c^j_i(u) - c^{j+1}_i(u))]$$

$$+ E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i c^1_i(u)].$$

Since $0 \in \mathcal{N}$, and the diameter of $\mathcal{N}$ is smaller than $\mathcal{B}$, we can choose $C_1 = \{0\}$ such that the third term in the above display vanishes. By Hölder’s inequality, we deduce that the first term can be bounded by $\epsilon_K$ as follows.

$$E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i (u(Z_i) - c^K_i(u))]$$

$$\leq E_{\Sigma}[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n \sigma_i (\sum_{i=1}^n \max_{i=1, \ldots, n} |u(Z_i) - c^K_i(u)|)]$$

$$\leq \epsilon_K.$$
Let \( V_j = \{ c^j(u) - c^{j+1}(u) : u \in \mathcal{N} \} \). Then by definition, the number of elements in \( V_j \) and \( C_j \) satisfying

\[
|V_j| \leq |C_j||C_{j+1}| \leq |C_{j+1}|^2.
\]

And the diameter of \( V_j \) denoted as \( D_j \) can be bounded as

\[
D_j = \sup_{v \in V_j} \|v\|_2 \leq \sqrt{n} \sup_{u \in \mathcal{N}} \|c^j(u) - c^{j+1}(u)\|_\infty
\]
\[
\leq \sqrt{n} \sup_{u \in \mathcal{N}} \|c^j(u) - u\|_\infty + \|u - c^{j+1}(u)\|_\infty
\]
\[
\leq \sqrt{n}(\epsilon_j + \epsilon_{j+1})
\]
\[
\leq 3\sqrt{n}\epsilon_{j+1}.
\]

Then,

\[
\mathbb{E}_\Sigma[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(c^j_i(u) - c^{j+1}_i(u))] \leq \sum_{j=1}^{K-1} \frac{D_j}{n} \sqrt{2 \log(2|V_j|)}
\]
\[
\leq \sum_{j=1}^{K-1} \frac{6\epsilon_{j+1}}{n} \sqrt{\log(2|C_{j+1}|)},
\]

where we use triangle inequality in the first inequality, and use Lemma 3.12 in the second inequality. Putting all the above estimates together, we get

\[
\mathbb{E}_\Sigma[\sup_{u \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i u(Z_i)] \leq \epsilon_K + \sum_{j=1}^{K-1} \frac{6\epsilon_{j+1}}{n} \sqrt{\log(2|C_{j+1}|)}
\]
\[
\leq \epsilon_K + \frac{12}{n} \int_{\epsilon_K+1}^{B} \sqrt{\log(2C(\epsilon, \mathcal{N}, n))} d\epsilon
\]
\[
\leq \inf_{0 < \delta < B} \left( 4\delta + \frac{12}{n} \int_{\delta}^{B} \sqrt{\log(2C(\epsilon, \mathcal{N}, n))} d\epsilon \right).
\]

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

Now we turn to handle the most difficult term \( \sup_{u \in \mathcal{N}^2} |\mathcal{L}_4(u) - \hat{\mathcal{L}}_4(u)| \), where we need to bound the Rademacher complexity of the non-Lipschitz composition of gradient norm and ReLU\(^2\) network. We believe that the technique we used here is helpful for bounding the statistical errors for other deep PDEs solvers where the main difficulties is bounding the Rademacher complexity of non-Lipschitz composition induced by the gradient operator.
**Lemma 3.13.**

\[
\mathbb{E}_{Z_n} \left[ \sup_{u \in \mathcal{N}^2} |L_4(u) - \hat{L}_4(u)| \right] \\
\leq \mathbb{E}_{Z_n, \mathcal{S}_n} \left[ \sup_{u \in \mathcal{N}^2} \frac{1}{n} \sum_i \sigma_i |\nabla u(Z_i)|^2 \right] \\
\leq \mathcal{R}(\mathcal{N}^{1,2}) = \mathbb{E}_{Z_n, \mathcal{S}_n} \left[ \sup_{u \in \mathcal{N}^{1,2}} \frac{1}{n} |\sum_i \sigma_i u(Z_i)| \right] \\
\leq \inf_{0 < \delta < B} \left( 4\delta + \frac{12}{\sqrt{n}} \int_0^B \sqrt{\log(2\mathcal{C} (\epsilon, \mathcal{N}^{1,2}, n))} \, dx \right). \tag{3.8}
\]

**Proof.** The proof of (3.6) is based on the symmetrization method used in the proof of Lemma 3.8, we omit the detail here. The proof of (3.7) is a direct consequence of the following claim: Let \( u \) be a function implemented by a ReLU\(^2 \) network with depth \( D \) and width \( W \). Then \( |\nabla u|^2 \) can be implemented by a ReLU-ReLU\(^2 \) network with depth \( D + 3 \) and width \( d(D + 2)W \).

Denote ReLU and ReLU\(^2 \) as \( \sigma_1 \) and \( \sigma_2 \), respectively. As long as we show that each partial derivative \( D_i u(i = 1, 2, \cdots, d) \) can be implemented by a ReLU-ReLU\(^2 \) network respectively, we can easily obtain the network we desire, since, \( |\nabla u|^2 = \sum_{i=1}^d |D_i u|^2 \) and the square function can be implemented by \( x^2 = \sigma_2(x) + \sigma_2(-x) \).

Now we show that for any \( i = 1, 2, \cdots, d \), \( D_i u \) can be implemented by a ReLU-ReLU\(^2 \) network. We deal with the first two layers in details since there are a little bit difference for the first two layer and apply induction for layers \( k \geq 3 \). For the first layer, since \( \sigma_2'(x) = 2\sigma_1(x) \), we have for any \( q = 1, 2 \cdots, n_1 \)

\[
D_i u_q^{(1)} = D_i \sigma_1 \left( \sum_{j=1}^d a_q^{(1)} x_j + b_q^{(1)} \right) = 2\sigma_1 \left( \sum_{j=1}^d a_q^{(1)} x_j + b_q^{(1)} \right) \cdot a_q^{(1)}
\]

Hence \( D_i u_q^{(1)} \) can be implemented by a ReLU-ReLU\(^2 \) network with depth 2 and width 1. For the second layer,

\[
D_i u_q^{(2)} = D_i \sigma_2 \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} u_j^{(1)} + b_q^{(2)} \right) = 2\sigma_1 \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} u_j^{(1)} + b_q^{(2)} \right) \cdot \sum_{j=1}^{n_1} a_{qj}^{(2)} D_j u_j^{(1)}
\]

Since \( \sigma_1 \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} u_j^{(1)} + b_q^{(2)} \right) \) and \( \sum_{j=1}^{n_1} a_{qj}^{(2)} D_j u_j^{(1)} \) can be implemented by two ReLU-ReLU\(^2 \) subnetworks, respectively, and the multiplication can also be implemented by

\[
x \cdot y = \frac{1}{4} \left[ (x + y)^2 - (x - y)^2 \right] \\
= \frac{1}{4} \left[ \sigma_2(x + y) + \sigma_2(-x - y) - \sigma_2(x - y) - \sigma_2(-x + y) \right],
\]

13
we conclude that $D_i u_q^{(2)}$ can be implemented by a ReLU-ReLU$^2$ network. We have
\[
D \left( \sigma_1 \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} u_j^{(1)} + b_q^{(2)} \right) \right) = 3, W \left( \sigma_1 \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} u_j^{(1)} + b_q^{(2)} \right) \right) \leq W
\]
and
\[
D \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} D_i u_j^{(1)} \right) = 2, W \left( \sum_{j=1}^{n_1} a_{qj}^{(2)} D_i u_j^{(1)} \right) \leq W.
\]
Thus $D \left( D_i u_q^{(2)} \right) = 4, W \left( D_i u_q^{(2)} \right) \leq \max\{2W, 4\}$.

Now we apply induction for layers $k \geq 3$. For the third layer,
\[
D_i u_q^{(3)} = D_i \sigma_2 \left( \sum_{j=1}^{n_2} a_{qj}^{(3)} u_j^{(2)} + b_q^{(3)} \right) = 2 \sigma_1 \left( \sum_{j=1}^{n_2} a_{qj}^{(3)} u_j^{(2)} + b_q^{(3)} \right) \cdot \sum_{j=1}^{n_2} a_{qj}^{(3)} D_i u_j^{(2)}.
\]
Since $D \left( \sigma_1 \left( \sum_{j=1}^{n_2} a_{qj}^{(3)} u_j^{(2)} + b_q^{(3)} \right) \right) = 4, W \left( \sigma_1 \left( \sum_{j=1}^{n_2} a_{qj}^{(3)} u_j^{(2)} + b_q^{(3)} \right) \right) \leq W$ and
\[
D \left( \sum_{j=1}^{n_2} a_{qj}^{(3)} D_i u_j^{(2)} \right) = 4, W \left( \sum_{j=1}^{n_2} a_{qj}^{(3)} D_i u_j^{(2)} \right) \leq \max\{2W, 4W\} = 4W,
\]
we conclude that $D_i u_q^{(3)}$ can be implemented by a ReLU-ReLU$^2$ network and $D \left( D_i u_q^{(3)} \right) = 5, W \left( D_i u_q^{(3)} \right) \leq \max\{5W, 4\} = 5W$.

We assume that $D_i u_q^{(k)} (q = 1, 2, \ldots, n_k)$ can be implemented by a ReLU-ReLU$^2$ network
and $D \left( D_i u_q^{(k)} \right) = k + 2, W \left( D_i u_q^{(k)} \right) \leq (k + 2)W$. For the $(k + 1)$-th layer,
\[
D_i u_q^{(k+1)} = D_i \sigma_2 \left( \sum_{j=1}^{n_k} a_{qj}^{(k+1)} u_j^{(k+1)} + b_q^{(k+1)} \right) = 2 \sigma_1 \left( \sum_{j=1}^{n_k} a_{qj}^{(k+1)} u_j^{(k+1)} + b_q^{(k+1)} \right) \cdot \sum_{j=1}^{n_k} a_{qj}^{(k+1)} D_i u_j^{(k+1)}.
\]
Since $D \left( \sigma_1 \left( \sum_{j=1}^{n_k} a_{qj}^{(k+1)} u_j^{(k+1)} + b_q^{(k+1)} \right) \right) = k + 2, W \left( \sigma_1 \left( \sum_{j=1}^{n_k} a_{qj}^{(k+1)} u_j^{(k+1)} + b_q^{(k+1)} \right) \right) \leq W$ and
\[
D \left( \sum_{j=1}^{n_k} a_{qj}^{(k+1)} D_i u_j^{(k+1)} \right) = k + 2, W \left( \sum_{j=1}^{n_k} a_{qj}^{(k+1)} D_i u_j^{(k+1)} \right) \leq \max\{(k + 2)W, 4W\} = (k + 2)W,
\]
we conclude that $D_i u_q^{(k+1)}$ can be implemented by a ReLU-ReLU$^2$ network and $D \left( D_i u_q^{(k+1)} \right) = k + 3, W \left( D_i u_q^{(k+1)} \right) \leq \max\{(k + 3)W, 4\} = (k + 3)W$.

Hence we derive that $D_i u = D_i u_q^D$ can be implemented by a ReLU-ReLU$^2$ network and
\[
D \left( D_i u \right) = D + 2, W \left( D_i u \right) \leq (D + 2)W.
\]
Finally we obtain that $D \left( \|\nabla u\|^2 \right) = D + 3, W \left( \|\nabla u\|^2 \right) \leq d (D + 2) W$.

The proof of (3.8) follows from Lemma 3.11. □

By Lemma 3.8, Lemma 3.7, Lemma 3.11 and Lemma 3.13, we have to find upper bounds for the converging numbers $C \left( \epsilon, N^2, n \right)$ and $C \left( \epsilon, N^{1.2}, n \right)$ used in the Dudley’s entropy formula.
To this end, we need the VC-dimension [46] and Pseudo-dimension [1].

**Definition 3.14.** Let \( N \) be a set of functions from \( X = \Omega(\partial \Omega) \) to \( \{0, 1\} \). Suppose that \( S = \{x_1, x_2, \ldots, x_n\} \subset X \). We say that \( S \) is shattered by \( N \) if for any \( b \in \{0, 1\}^n \), there exists a \( u \in N \) satisfying

\[
u(x_i) = b_i, \quad i = 1, 2, \ldots, n
\]

**Definition 3.15.** The VC-dimension of \( N \), denoted as \( \text{VCdim}(N) \), is defined to be the maximum cardinality among all sets shattered by \( N \).

VC-dimension reflects the capability of a class of functions to perform binary classification of points. The larger VC-dimension is, the stronger the capability to perform binary classification is. For more discussion of VC-dimension, readers are referred to [1].

For real-valued functions, we can generalize the concept of VC-dimension into pseudo-dimension [1].

**Definition 3.16.** Let \( N \) be a set of functions from \( X \) to \( \mathbb{R} \). Suppose that \( S = \{x_1, x_2, \ldots, x_n\} \subset X \). We say that \( S \) is pseudo-shattered by \( N \) if there exists \( y_1, y_2, \ldots, y_n \) such that for any \( b \in \{0, 1\}^n \), there exists a \( u \in N \) satisfying

\[
\text{sign}(u(x_i) - y_i) = b_i, \quad i = 1, 2, \ldots, n
\]

and we say that \( \{y_i\}_{i=1}^n \) witnesses the shattering.

**Definition 3.17.** The pseudo-dimension of \( N \), denoted as \( \text{Pdim}(N) \), is defined to be the maximum cardinality among all sets pseudo-shattered by \( N \).

The following proposition showing a relation between uniform covering number and pseudo-dimension [1].

**Proposition 3.18 (Theorem 12.2, [1]).** Let \( N \) be a set of real functions from a domain \( X \) to the bounded interval \([0, B]\). Let \( \epsilon > 0 \). Then

\[
C_{\infty}(\epsilon, N, n) \leq \sum_{i=1}^{\text{Pdim}(N)} \binom{n}{i} \left( \frac{B}{\epsilon} \right)^i,
\]

which is less than \( \left( \frac{enB}{\epsilon \cdot \text{Pdim}(N)} \right)^{\text{Pdim}(N)} \) for \( n \geq \text{Pdim}(N) \).

We now present the bound of pseudo-dimension for the \( \mathcal{N}^{1,2} \), the class of network functions with ReLU and ReLU\(^2\) activation functions. We first need a lemma stated below.

**Lemma 3.19.** Let \( p_1, \ldots, p_m \) be polynomials with \( n \) variables of degree at most \( d \). If \( n \leq m \), then

\[
|\{(\text{sign}(p_1(x)), \ldots, \text{sign}(p_m(x)) : x \in \mathbb{R}^n\}| \leq 2 \left( \frac{2emd}{n} \right)^n
\]

**Proof.** See Theorem 8.3 in [1]. \( \square \)
Theorem 3.2. Let
\[ N := \{ u \in [0, 1]^d : u \text{ can be implemented by a neural network} \]
\[ \text{with depth no more than } D \text{ and width no more than } W, \]
\[ \text{and activation function in each unit be the ReLU or the ReLU}^2. \} \]

Then
\[ \text{Pdim}(N) = O(D^2W^2(D + \log W)). \]

Proof. The argument is follows from the proof of Theorem 6 in [2]. The result stated here
is somewhat stronger then Theorem 6 in [2] since VCdim(sign(N)) \leq \text{Pdim}(N).

We consider a new set of functions:
\[ \tilde{N} = \{ \tilde{u}(x, y) = \text{sign}(u(x) - y) : u \in \mathcal{H} \} \]

It is clear that Pdim(N) \leq VCdim(\tilde{N}). We now bound the VC-dimension of \tilde{N}. Denoting \mathcal{M} as the total number of parameters(weights and biases) in the neural network implementing
functions in \mathcal{N}, in our case we want to derive the uniform bound for
\[ K_{\{x_i\}, \{y_i\}}(m) := |\{(\text{sign}(f(x_1, a) - y_1, \ldots, \text{sign}(u(x_m, a) - y_m)) : a \in \mathbb{R}^\mathcal{M}\}| \]

over all \{x_i\}_{i=1}^m \subset X and \{y_i\}_{i=1}^m \subset \mathbb{R}. Actually the maximum of \text{K}_{\{x_i\}, \{y_i\}}(m) over all
\{x_i\}_{i=1}^m \subset X and \{y_i\}_{i=1}^m \subset \mathbb{R} is the growth function \mathcal{G}_{\tilde{N}}(m). In order to apply Lemma 3.19, we
partition the parameter space \mathbb{R}^\mathcal{M} into several subsets to ensure that in each subset \text{u}(x_i, a) - y_i
is a polynomial with respect to \text{a} without any breakpoints. In fact, our partition is exactly the
same as the partition in [2]. Denote the partition as \{P_1, P_2, \ldots, P_N\} with some integer \text{N}
satisfying
\[ N \leq \prod_{i=1}^{D-1} \left( \frac{2emk_i(1 + (i - 1)2^{i-1})}{\mathcal{M}_i} \right)^{\mathcal{M}_i} \tag{3.9} \]

where \text{k}_i and \text{M}_i denotes the number of units at the \text{i}th layer and the total number of parameters
at the inputs to units in all the layers up to layer \text{i} of the neural network implementing functions
in \mathcal{N}, respectively. See [2] for the construction of the partition. Obviously we have
\[ K_{\{x_i\}, \{y_i\}}(m) \leq \sum_{i=1}^{N} |\{(\text{sign}(u(x_1, a) - y_1), \ldots, \text{sign}(u(x_m, a) - y_m)) : a \in P_i\}| \tag{3.10} \]

Note that \text{u}(x_i, a) - y_i is a polynomial with respect to \text{a} with degree the same as the degree of
$u(x_i, a)$, which is equal to $1 + (D - 1)2^{D-1}$ as shown in [2]. Hence by Lemma 3.19, we have

$$|\{(\text{sign}(u(x_1, a) - y_1), \cdots, \text{sign}(u(x_m, a) - y_m)) : a \in P_i\}| \leq 2 \left(\frac{2em(1 + (D - 1)2^{D-1})}{M_D}\right)^{M_D}.$$  \hspace{1cm} (3.11)

Combining (3.9), (3.10), (3.11) yields

$$K_{(x_i), (y_i)}(m) \leq \prod_{i=1}^{D} 2 \left(\frac{2emk_i(1 + (i - 1)2^{i-1})}{M_i}\right)^{M_i}.$$  

We then have

$$G_N(m) \leq \prod_{i=1}^{D} 2 \left(\frac{2emk_i(1 + (i - 1)2^{i-1})}{M_i}\right)^{M_i},$$

since the maximum of $K_{(x_i), (y_i)}(m)$ over all $\{x_i\}_{i=1}^{m} \subset X$ and $\{y_i\}_{i=1}^{m} \subset \mathbb{R}$ is the growth function $G_N(m)$. Some algebras as that of the proof of Theorem 6 in [2], we obtain

$$\text{Pdim}(N) \leq O(2^D W^2 \log U + D^3 W^2) = O(D^2 W^2 (D + \log W))$$

where $U$ refers to the number of units of the neural network implementing functions in $N$.

### 3.3. Main results.

With the above preparation we present the main results of this paper in the scenario $\mathcal{E}_{opt} = 0$.

**Theorem 3.3.** Let $u^*$ be the solution of (2.1) with bounded $f, g, w$. $\hat{u}_\phi$ is the minimizer of deep Ritz method defined in (2.3) with $n = N(M)$ random samples. If we set the network parameters depth and width in the ReLU$^2$ network $N^2_D, W, B$ as

$$D \leq \left[\log_2 d\right] + 3, W \leq O(4d \left[n^{-d/2} - 4\right]^d).$$

Let $B$ be the constraint on the output of function value and gradient norm of $u \in N^2$ according to (3.5). Then,

$$\mathbb{E}_X \mathbb{Y} [\|\hat{u}_\phi - u^*\|_{H^1(\Omega)}^2] \leq C_{B, c_1, c_2, c_3, d} O(N^{-1/(d+2+v)} + M^{-1/(d+2+v)}),$$

where $v > 0$ but can be arbitrary small.

**Proof.** In order to apply Lemma 3.11, we need to bound the term

$$\frac{1}{\sqrt{n}} \int_\delta^B \sqrt{\log(2\mathcal{C}(\epsilon, N, n))} d\epsilon$$

$$\leq \frac{B}{\sqrt{n}} + \frac{1}{\sqrt{n}} \int_\delta^B \sqrt{\log \left(\frac{enB}{\epsilon \cdot \text{Pdim}(N)}\right)^{\text{Pdim}(N)}} d\epsilon$$

$$\leq \frac{B}{\sqrt{n}} + \left(\frac{\text{Pdim}(N)}{n}\right)^{1/2} \int_\delta^B \sqrt{\log \left(\frac{enB}{\epsilon \cdot \text{Pdim}(N)}\right)} d\epsilon,$$
where in the first inequality we use Proposition 3.18. Now we calculate the integral. Set

\[ t = \sqrt{\log \left( \frac{e B}{\epsilon \cdot \text{Pdim}(\mathcal{N})} \right)} \]

then \( \epsilon = \frac{e B}{\text{Pdim}(\mathcal{N})} \cdot e^{-t^2} \). Denote \( t_1 = \sqrt{\log \left( \frac{e B}{\text{Pdim}(\mathcal{N})} \right)} \), \( t_2 = \sqrt{\log \left( \frac{e B}{\delta \cdot \text{Pdim}(\mathcal{N})} \right)} \). And

\[
\begin{align*}
\int_\delta^B \sqrt{\log \left( \frac{e B}{\epsilon \cdot \text{Pdim}(\mathcal{N})} \right)} d\epsilon &= \frac{2eB}{\text{Pdim}(\mathcal{N})} \int_{t_1}^{t_2} t^2 e^{-t^2} dt \\
&= \frac{2eB}{\text{Pdim}(\mathcal{N})} \int_{t_1}^{t_2} t \left(-e^{-t^2}\right)' dt \\
&= \frac{eB}{\text{Pdim}(\mathcal{N})} \left[ t_1 e^{-t_1^2} - t_2 e^{-t_2^2} + \int_{t_1}^{t_2} e^{-t^2} dt \right] \\
&\leq \frac{eB}{\text{Pdim}(\mathcal{N})} \left[ t_1 e^{-t_1^2} - t_2 e^{-t_2^2} + (t_2 - t_1)e^{-t_1^2} \right] \\
&\leq \frac{eB}{\text{Pdim}(\mathcal{N})} \cdot t_2 e^{-t_1^2} = B \sqrt{\log \left( \frac{e B}{\delta \cdot \text{Pdim}(\mathcal{N})} \right)}.
\end{align*}
\]

Choosing \( \delta = B \left( \frac{\text{Pdim}(\mathcal{N})}{n} \right)^{1/2} \leq B \), by Lemma 3.11 and the above display, we get for both \( \mathcal{N} = \mathcal{N}^2 \), and \( \mathcal{N} = \mathcal{N}^{1,2} \) there holds

\[ \Re(\mathcal{N}) \]

\[
\begin{align*}
&\leq 4\delta + \frac{12}{\sqrt{n}} \int_\delta^B \sqrt{\log(2\mathcal{C}(\epsilon, \mathcal{N}, n))} d\epsilon \\
&\leq 4\delta + \frac{12B}{\sqrt{n}} + 12B \left( \frac{\text{Pdim}(\mathcal{N})}{n} \right)^{1/2} \sqrt{\log \left( \frac{eB}{\delta \cdot \text{Pdim}(\mathcal{N})} \right)} \\
&\leq 28 \sqrt{\frac{3}{2} \frac{B}{n} \left( \frac{\text{Pdim}(\mathcal{N})}{n} \right)^{1/2} \sqrt{\log \left( \frac{eB}{\text{Pdim}(\mathcal{N})} \right)}}. \tag{3.12}
\end{align*}
\]

Then by Lemma 3.4, 3.7, 3.8, 3.11, 3.13 and equation (3.12), we have

\[ E_{stu} = 2 \sup_{u \in \mathcal{N}^2} |\mathcal{L}(u) - \widehat{\mathcal{L}}(u)| \]

\[
\begin{align*}
&\leq 2(c_3 + c_3 + c_3^2) \Re(\mathcal{N}^2) + 2\Re(\mathcal{N}^{1,2}) \\
&\leq 56 \sqrt{\frac{3}{2} \frac{B}{n} \left( \frac{\text{Pdim}(\mathcal{N}^2)}{n} \right)^{1/2} \sqrt{\log \left( \frac{eB}{\text{Pdim}(\mathcal{N}^2)} \right)}} + 56 \sqrt{\frac{3}{2} \frac{B}{n} \left( \frac{\text{Pdim}(\mathcal{N}^{1,2})}{n} \right)^{1/2} \sqrt{\log \left( \frac{eB}{\text{Pdim}(\mathcal{N}^{1,2})} \right)}}.
\end{align*}
\]
Since $\text{Pdim}(N^2) \leq \text{Pdim}(N^{1,2})$, plugging the upper bound of $\text{Pdim}(N^{1,2})$ derived in Theorem 3.2 into the above display and using the relationship of depth and width between $N^2$ and $N^{1,2}$ proved in Lemma 3.13, we get

$$E_{\text{sta}} \leq CB_{c,2} \left[ d(D + 3)(D + 2)W \sqrt{\frac{D + 3 + \log(d(D + 2)W)}{n}} \right]^{1-\nu}, \quad (3.13)$$

where $\nu > 0$ can be arbitrarily small. Combing (3.13) with the approximation error in Theorem 3.1 and taking $\epsilon^2 = CB_{c,1,2,3} C (\frac{1}{n})^{1+\frac{\nu}{d+2}}$, we get that

$$\mathbb{E}_{X,Y}[\|\hat{u}_\phi - u^*\|^2_{H^1(\Omega)}] \leq \frac{2}{c_1} \left[ CB_{c,2} \left( d(D + 3)(D + 2)W \sqrt{\frac{D + 3 + \log(d(D + 2)W)}{n}} \right)^{1-\nu} + \frac{c_3 + 1}{2} \epsilon^2 \right].$$

**Remark 3.1.** Deep Ritz method is actually a kinds of deep nonparametric estimation method where we estimate functions from random samples. The benefit is that we can use DRM to handle PDEs in high dimension since only a small bach of samples is needed during SGD training. In contrast, we have form the loading matrix and vector explicitly in FEM. However, what we have to pay is that the convergence rate as illustrate as follows. Let $N = N = O\left(\frac{1}{h^d}\right)$, where $h$ is the size of the mesh in FEM. From Theorem 3.3, we get

$$\mathbb{E}_{X,Y}[\|\hat{u}_\phi - u^*\|^2_{H^1(\Omega)}] = O\left(h^{-\frac{d}{d+2}}\right).$$

In the case $d = 2$, by Markov’s inequality and the above display, we get with high probability,

$$\|\hat{u}_\phi - u^*\|_{H^1(\Omega)} \leq O(h^{-\frac{2}{3}}).$$

Comparing the well known results of FEM where the convergence rate in $H^1$ norm is $O(h)$, the rate proved here for the DRM method is far from satisfactory.

In the literature of nonparametric regression, where functions living in certain function classes are estimated from $n$ paired random samples, the best convergence rate in $H^1$ norm for estimating functions in $H^2$ is $O(n^{-\frac{d}{d+2}})$ [41]. Obviously, the nonparametric learning task in DRM is not easier than nonparametric regressions. What we proved here is $O(n^{-\frac{d}{d+2+\nu}})$ in $H^1$ norm with $\nu$ can be arbitrary small. We conjecture that the best convergence rate in $H^1$ norm of DRM with both $n$ training samples in both the domain and boundary is
also $O(n^{-1/4})$. Upon this conjecture, we get the optimal convergence rate of DRM for $d = 2$ is $\|\tilde{u}_0 - u^*\|_{H^1(\Omega)} \leq O(h^{1/2})$ with high probability.

The convergence rate of DRM proved here suffers the curse of dimensionality. One possible direction to improve the convergence rate and reduce the curse of dimensionality is considering solutions of PDE (2.1) with higher regularity. For example, if we assume $f \in H^s(\Omega), s \geq 1$, deep Ritz method using deep neural network with ReLU$^{s+2}$ activation functions will reduce the curse since the higher regularity assumption can improve both the approximation and statistical error. We leave the detail of this idea in a following up work.

4. Conclusion and extension. In this paper, we provide a rigorous numerical analysis on deep Ritz method (DRM) [48] for second order elliptic equations with Neumann boundary conditions. We establish the first nonasymptotic convergence rate in $H^1$ norm on DRM for general deep networks with ReLU$^2$ activation functions. In addition to provide theoretical justification of DRM, our study also provides guidance on how to set the hyper-parameter of depth and width to achieve the desired convergence rate in terms of number of training samples. Technically, we derive bounds on the approximation error of deep ReLU$^2$ network in $H^1$ norm and on the Rademacher complexity of the non-Lipschitz composition of gradient norm and ReLU$^2$ network.

There are several directions for our future exploration. First, it is easy to extend the current analysis for general second order elliptic equations with variational form under Dirichlet or Robin boundary conditions. Second, the approximation and statistical error bounds deriving here can be used for studying the nonasymptotic convergence rate for residual based method (PINNS). Studying deep DGM by combing current analysis with the tools for analyzing GAN [14] is also of immense interest.

Acknowledgements. Y. Jiao was supported in part by the National Science Foundation of China under Grant 11871474 and by the research fund of KLATASDSMOE. X. Lu is partially supported by the National Science Foundation of China (No. 11871385), the National Key Research and Development Program of China (No.2018YFC1314600) and the Natural Science Foundation of Hubei Province (No. 2019CFA007). J. Yang is supported by National Science Foundation of China (No. 12071362 and 11671312), the National Key Research and Development Program of China (No. 2020YFA0714200), the Natural Science Foundation of Hubei Province (No. 2019CFA007). The numerical simulations in this work have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

REFERENCES

[1] Martin Anthony and Peter L. Bartlett, Neural network learning: Theoretical foundations, Cambridge University Press, 2009.

[2] Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian, Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks, J. Mach. Learn. Res., 20 (2019), pp. 1–17.

[3] Christian Beck, Weinan E, and Arnulf Jentzen, Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations, Journal of Nonlinear Science, 29 (2019), pp. 1563–1619.
[29] G. Pang, M. D'Elia, M. Parks, and G.E. Karniadakis, npinns: Nonlocal physics-informed neural networks for a parametrized nonlocal universal laplacian operator: algorithms and applications, Journal of Computational Physics, 422 (2020), p. 109760.

[30] Guofei Pang, Lu Lu, and George Em Karniadakis, fpinns: Fractional physics-informed neural networks, SIAM Journal on Scientific Computing, 41 (2019), pp. A2603–A2626.

[31] 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 (2019), pp. 686–707.

[32] Deep Ray and Jan S. Hesthaven, An artificial neural network as a troubled-cell indicator, Journal of Computational Physics, 367 (2018), pp. 166–191.

[33] Johannes Schmidt-Hieber, Deep relu network approximation of functions on a manifold, arXiv preprint arXiv:1908.00695, (2019).

[34] Martin H Schultz, Approximation theory of multivariate spline functions in sobolev spaces, SIAM Journal on Numerical Analysis, 6 (1969), pp. 570–582.

[35] Larry Schumaker, Spline functions: basic theory, Cambridge University Press, 2007.

[36] Zuowei Shen, Hanzhao Yang, and Shijun Zhang, Deep network approximation characterized by number of neurons, arXiv preprint arXiv:1906.05497, (2019).

[37] Yeonjong Shin, Jerome Darbon, and George Em Karniadakis, On the convergence and generalization of physics informed neural networks, arXiv preprint arXiv:2004.01806, (2020).

[38] 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).

[39] Jonathan W Siegel and Jinchao Xu, Approximation rates for neural networks with general activation functions, Neural Networks, 128 (2020), pp. 313–321.

[40] Justin A. Sirignano and K. Spiliopoulos, Dgm: A deep learning algorithm for solving partial differential equations, Journal of Computational Physics, 375 (2018), pp. 1339–1364.

[41] Charles J Stone, Optimal global rates of convergence for nonparametric regression, The annals of statistics, (1982), pp. 1040–1053.

[42] Zheng Sun, Shuyi Wang, Lo-Bin Chang, Yulong Xing, and Dongbin Xiu, Convolution neural network shock detector for numerical solution of conservation laws, Communications in Computational Physics, 28 (2020), pp. 2075–2108.

[43] 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).

[44] J.W. Thomas, Numerical Partial Differential Equations: Finite Difference Methods, vol. 22, Springer Science & Business Media, 2013.

[45] Aad W van Der Vaart and Jon A Wellner, Weak convergence, in Weak convergence and empirical processes, Springer, 1996.

[46] Vladimir N Vapnik and A Ya Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, in Measures of complexity, Springer, 2015, pp. 11–30.

[47] Yufei Wang, Ziju Shen, Zichao Long, and Bin Dong, Learning to discretize: Solving 1d scalar conservation laws via deep reinforcement learning, Communications in Computational Physics, 28 (2020), pp. 2158–2179.

[48] E. Weinan and Teng Yu, The deep ritz method: A deep learning-based numerical algorithm for solving variational problems, Communications in Mathematics and Statistics, 6 (2017), pp. 1–12.

[49] Jinchao Xu, The finite neuron method and convergence analysis, arXiv preprint arXiv:2010.01458, (2020).

[50] Dmitry Yarotsky, Error bounds for approximations with deep relu networks, Neural Networks, 94 (2017), pp. 103–114.

[51] Yaohua Zang, Gang Bao, Xiaoming Ye, and Haomin Zhou, Weak adversarial networks for high-dimensional partial differential equations, Journal of Computational Physics, 411 (2020), p. 109409.