Semi-Supervised Deep Sobolev Regression: Estimation and Variable Selection by ReQU Neural Network

Zhao Ding, Chenguang Duan, Yuling Jiao, and Jerry Zhijian Yang

***Abstract*—We propose SDORE, a semi-supervised deep Sobolev regressor, for the nonparametric estimation of the underlying regression function and its gradient. SDORE employs deep ReQU neural networks to minimize the empirical risk with gradient norm regularization, allowing the approximation of the regularization term by unlabeled data. Our study includes a thorough analysis of the convergence rates of SDORE in L^2-norm, achieving the minimax optimality. Further, we establish a convergence rate for the associated plug-in gradient estimator, even in the presence of significant domain shift. These theoretical findings offer valuable insights for selecting regularization parameters and determining the size of the neural network, while showcasing the provable advantage of leveraging unlabeled data in semi-supervised learning. To the best of our knowledge, SDORE is the first provable neural network-based approach that simultaneously estimates the regression function and its gradient, with diverse applications such as nonparametric variable selection. The effectiveness of SDORE is validated through an extensive range of numerical simulations.**

***Index Terms*—Nonparametric regression, gradient estimation, variable selection, convergence rate, gradient penalty, deep neural network.**

I. Introduction

NONPARAMETRIC regression plays a pivotal role in both statistics and machine learning, possessing an illustrious research history as well as a vast compendium of related literature [1], [2], [3]. Let $\Omega \subseteq \mathbb{R}^d$, $d \geq 1$, be a bounded and connected domain with sufficiently smooth boundary $\partial\Omega$. Consider the following nonparametric regression model

$$Y = f_0(X) + \xi, \tag{1}$$

where $Y \in \mathbb{R}$ is the response associated with the covariate $X \in \Omega$, and f_0 is the unknown regression function. Here ξ represents a random noise term satisfying $\mathbb{E}[\xi|X] = 0$ and $\mathbb{E}[\xi^2|X] < \infty$. The primary task of nonparametric regression involves estimating the conditional expectation $f_0(x)$ of the response Y, given a covariate $X = x$. This estimation is typically achieved through empirical least-squares risk minimization:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2,$$

where $\{(X_i, Y_i)\}_{i=1}^n$ is a set of independently and identically distributed random copies of (X, Y), and $\mathcal{F}$ is a pre-specific hypothesis class, such as deep ReQU neural network class in this paper. While empirical least-squares risk minimization is straightforward to implement and comes with solid theoretical guarantees, it does not fully meet all desired criteria. One major drawback is that the method places no constraints on the gradient of the estimator, allowing for the possibility of an arbitrarily large gradient norm. This can make the least-squares estimator highly sensitive to the input perturbations. Furthermore, while the least-squares estimator ensures convergence in terms of function values, the convergence in terms of derivatives can not be guaranteed.

To address these challenges, Sobolev regularization, also known as gradient penalty, was introduced in deep learning by [4] and [5]:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 + \frac{\lambda}{n}\sum_{i=1}^{n}\sum_{k=1}^{d}|D_k f(X_i)|^2, \tag{2}$$

where $\lambda > 0$ is the regularization parameter, and $D_k f$ denotes the partial derivative of f with respect to the k-th input variable. Substantial numerical experiments have consistently demonstrated that the imposition of a gradient penalty contributes to the enhancement of the stability and generalization of deep learning models. The strategy surrounding gradient penalty was adopted by [6] as a technique to learn robust features using auto-encoders. This method was further utilized to augment the stability of deep generative models as highlighted in the work of [7], [8], [9], and [10]. Significantly, the gradient norm,

Received 21 March 2024; revised 17 October 2024; accepted 26 January 2025. Date of publication 31 January 2025; date of current version 19 March 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFA1014202; in part by the National Natural Science Foundation of China under Grant 12125103, Grant U24A2002, and Grant 12371441; and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Jerry Zhijian Yang.)*

Zhao Ding and Chenguang Duan are with the School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China (e-mail: zd1998@whu.edu.cn; cgduan.math@whu.edu.cn).

Yuling Jiao is with the School of Artificial Intelligence, the National Center for Applied Mathematics in Hubei, Hubei Key Laboratory of Computational Science, and the School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China (e-mail: yulingjiaomath@whu.edu.cn).

Jerry Zhijian Yang is with the National Center for Applied Mathematics in Hubei, Wuhan Institute for Math & AI, the School of Mathematics and Statistics, and Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan, Hubei 430072, China (e-mail: zjyang.math@whu.edu.cn).

Communicated by R. Venkataramanan, Associate Editor for Machine Learning.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2025.3537594.

Digital Object Identifier 10.1109/TIT.2025.3537594

being a local measure of sensitivity to input perturbations, has seen a plethora of research focusing on its use for adversarial robust learning. This is reflected in studies conducted by [11], [12], [13], [14], [15], and [16].

Simultaneously estimation the regression function and its of gradient (derivatives) carries a wide span of applications across various fields, including the factor demand and cost estimation in economics [17], trend analysis for time series data [18], the analysis of human growth data [19], and the modeling of spatial process [20]. Furthermore, estimating gradient plays a pivotal role in the modeling of functional data [21], [22], variable selection in nonparametric regression [23], [24], [25], and inverse problems [26]. There are four classical approaches to nonparametric gradient estimation: local polynomial regression [27], smoothing splines [28], kernel ridge regression [29], and difference quotients [30]. However, local polynomial regression and smoothing spline regression are only applicable to fixed-design setting and low-dimensional problems. The generalization of these methodologies to address high-dimensional problems is met with a significant challenge popularly known as the computational curse of dimensionality [2], [31], [32]. This phenomenon refers to the fact that the computational complexity can increase exponentially with dimension. In contrast, deep neural network-based methods, which are mesh-free, exhibit direct applicability to high-dimensional problems, providing a solution to mitigate this inherent challenge. The plug-in kernel ridge regression estimators have demonstrated applicability for estimating derivatives across both univariate and multivariate regressions within a random-design setting [29], [33]. However, these estimators present certain inherent limitations compared to deep neural networks. From a computational complexity standpoint, the scale of the kernel grows quadratically or even cubically with the number of samples. In contrast, deep neural networks exhibit the ability to handle larger datasets, especially when deployed on modern hardware architectures.

Recently, there has been a substantial literature outlining the convergence rates of deep nonparametric regression [34], [35], [36], [37], [38], [39], [40]. However, the theoretical foundation of Sobolev regularized least-squares using deep neural networks remains relatively underdeveloped. Consequently, two fundamental questions need to be addressed:

> *What accounts for the enhanced stability and superior generalization capacity of the Sobolev penalized estimator compared to the standard least-squares estimator? Furthermore, does the plug-in gradient estimator of the Sobolev penalized regressor close to the true gradient of the regression function, and if so, what is the corresponding convergence rate?*

In this paper, we introduce SDORE, a <u>**s**</u>emi-supervised <u>**d**</u>eep S<u>**o**</u>bolev <u>**re**</u>gressor, for simultaneously estimation of both the regression function and its gradient. SDORE leverages deep neural networks to minimize an empirical risk, augmented with unlabeled-data-driven Sobolev regularization:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 + \frac{\lambda}{m}\sum_{i=1}^{m}\sum_{k=1}^{d}|D_k f(Z_i)|^2, \tag{3}$$

where $\{Z_i\}_{i=1}^m$ is a set of unlabeled data independently and identically drawn from a distribution on Ω. Notably, our methodology does not necessitate alignment of the unlabeled data distribution with the marginal distribution of the labeled data, remaining effective even under significant domain shifts. In the context of semi-supervised learning, data typically consists of a modestly sized labeled dataset supplemented with vast amounts of unlabeled data. As a result, the empirical semi-supervised deep Sobolev regression risk aligns tightly with the following deep Sobolev regression problem:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 + \lambda\|\nabla f\|_{L^2(\Omega)}^2,$$

plays a pivotal role in nonparametric regression and has been investigated by [1], [41], [42], and [43]. We establish non-asymptotic convergence rates for the deep Sobolev regressor and demonstrate that the norm of its gradient is uniformly bounded, shedding light on the considerable stability and favorable generalization properties of the estimator. Furthermore, under certain mild conditions, we derive non-asymptotic convergence rates for the plug-in derivative estimator based on SDORE. This illustrates how abundant unlabeled data used in SDORE (3) improves the performance of the standard gradient penalized regressor (2). We subsequently apply SDORE to nonparametric variable selection. The efficacy of this method is substantiated through numerous numerical examples.

A. Contributions

Our contributions can be summarized in four folds:

(i) We introduce a novel semi-supervised deep estimator within the framework of Sobolev penalized regression. A large amount of unlabeled data is employed to estimate the Sobolev penalty term. We demonstrate that this deep ReQU neural network-based estimator achieves the minimax optimal rate (Theorem 1). Meanwhile, with the appropriate selection of the regularization parameter, the norm of the estimator's gradient can be uniformly bounded, thereby illustrating its remarkable stability and generalization capacities from a theoretical standpoint.

(ii) Under certain mild conditions, we establish an oracle inequality for gradient estimation using the plug-in deep Sobolev regressor (Lemma 5). Notably, this oracle inequality is applicable to any convex hypothesis class. This represents a significant theoretical advancement beyond existing nonparametric plug-in gradient estimators, which are based on linear approximation [29], [44], by extending the framework to handle more complex hypothesis classes involved in nonlinear approximation [45]. Furthermore, we derive a convergence rate for the gradient of the deep ReQU neural network-based estimator, providing valuable a priori guidance for selecting regularization parameters and choosing the size of the neural network (Theorem 2).

(iii) We derive a convergence rate for semi-supervised estimator (Theorem 3), which sheds light on the quantifiable advantages of incorporating unlabeled data into the supervised learning. This improvement is actualized

TABLE I
CONVERGENCE RATES FOR SOBOLEV PENALIZED ESTIMATORS

Estimator	Reg. Param.	Convergence Rates		Minimax Opt.	
DORE Def. (9)	$\lambda = \mathcal{O}(n^{-\frac{2s}{d+2s}} \log^3 n)$	$\mathbb{E}\Vert \hat{f}^{\lambda}_{\mathcal{D}} - f_0 \Vert^2$	$\mathcal{O}(n^{-\frac{2s}{d+2s}} \log^3 n)$	✔	Theorem IV.3
		$\mathbb{E}\Vert \nabla \hat{f}^{\lambda}_{\mathcal{D}} \Vert^2$	$\mathcal{O}(1)$		
DORE Def. (12)	$\lambda = \mathcal{O}(n^{-\frac{s}{d+4s}} \log^2 n)$	$\mathbb{E}\Vert \hat{f}^{\lambda}_{\mathcal{D}} - f_0 \Vert^2$	$\mathcal{O}(n^{-\frac{2s}{d+4s}} \log^4 n)$	✘	Theorem V.4
		$\mathbb{E}\Vert \nabla(\hat{f}^{\lambda}_{\mathcal{D}} - f_0) \Vert^2$	$\mathcal{O}(n^{-\frac{s}{d+4s}} \log^2 n)$	✘	
SDORE Def. (12)	$\lambda = \mathcal{O}(n^{-\frac{s}{d+4s}} \log^2 n)$	$\mathbb{E}\Vert \hat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0 \Vert^2$	$\mathcal{O}(n^{-\frac{2s}{d+4s}} \log^4 n) + \mathcal{O}(n^{\frac{d}{d+4s}} \log^4 nm^{-1})$	✘	Theorem V.6
		$\mathbb{E}\Vert \nabla(\hat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0) \Vert^2$	$\mathcal{O}(n^{-\frac{s}{d+4s}} \log^2 n) + \mathcal{O}(n^{\frac{d+s}{d+4s}} \log^2 nm^{-1})$	✘	

under the condition that density ratio between the marginal distribution of the labeled data and the distribution of the unlabeled data remains uniformly bounded. This novel finding promises to enrich our theoretical comprehension of semi-supervised learning, particularly in the context of deep neural networks.

(iv) The gradient estimator introduces a novel tool with potential applications in areas such as nonparametric variable selection. In the case where the regression function exhibits sparsity structure (Assumption 7), we prove that the convergence rate depends only on the number of relevant variables, rather than the data dimension (Corollary 1). Moreover, we establish the selection consistency of the deep Sobolev regressor (Corollary 2), showing that, with a sufficiently large number of labeled data pairs, the estimated relevant set is highly likely to match the ground truth relevant set. To validate our approach, we conduct a series of numerical experiments, which confirm the effectiveness and reliability of our proposed methodology.

B. Main Results Overview

In this work, we focus on two estimators in the setting of nonparametric regression (1). The **D**eep S**O**bolev **RE**gressor (DORE) is derived from the regularized empirical risk minimization:

$$\widehat{f}^{\lambda}_{\mathcal{D}} \in \arg\min_{f\in\mathcal{F}} \widehat{L}^{\lambda}_{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 + \lambda\Vert\nabla f\Vert^2_{L^2(\nu_X)}, \qquad \text{DORE}$$

where $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{n}$ is a set of independent copies of (X, Y), $\lambda > 0$ is the regularization parameter, and $\mathcal{F}$ is a class of deep ReQU neural networks. In some application scenarios, the regularization term in (DORE) is intractable analytically. To address this issue, we approximate the regularization term by its data-driven counterpart, yielding the following semi-supervised empirical risk minimizer

$$\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} \in \arg\min_{f\in\mathcal{F}} \widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2 + \frac{\lambda}{m}\sum_{i=1}^{m}\sum_{k=1}^{d}|D_k f(Z_i)|^2, \qquad \text{SDORE}$$

where $\mathcal{S} = \{Z_i\}_{i=1}^{m}$ is a set of independently and identically random variables drawn from ν_X.

The main theoretical results derived in this paper are summarized in Table I. As shown in Theorem 1, the convergence rate of the deep Sobolev regressor in L^2-norm achieves the minimax optimality. However, Theorems 3 and 2 demonstrate that the convergence rates in L^2-norm and H^1-semi-norm is sub-optimal.

We utilize the deep Sobolev regressor to tackle an application scenarios: nonparametric variable selection. We present the theoretical findings related to nonparametric variable selection in Table II, including the convergence rate and selection consistency.

C. Preliminaries and Notations

Before proceeding, we introduce some notation and definitions. Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain, and let μ_X and ν_X be two probability measures on Ω with densities $p(x)$ and $q(x)$, respectively. The $L^2(\mu_X)$ inner-product and norm are given, respectively, by

$$(u, v)_{L^2(\mu_X)} = \int_{\Omega} uv d\mu_X,$$
$$\Vert u\Vert^2_{L^2(\mu_X)} = (u, u)_{L^2(\mu_X)}.$$

Similarly, one can define the $L^2(\nu_X)$ inner-product and norm. Furthermore, define the density ratio between ν_X and μ_X by $r(x) = q(x)/p(x)$. Suppose the density ratio is uniformly upper- and lower-bounded, that is, $\kappa := \sup_{x\in\Omega}|r(x)| < \infty$ and $\zeta := \inf_{x\in\Omega}|r(x)| > 0$. Then it is straightforward to verify that

$$\zeta\Vert u\Vert^2_{L^2(\mu_X)} \leq \Vert u\Vert^2_{L^2(\nu_X)} \leq \kappa\Vert u\Vert^2_{L^2(\mu_X)}.$$

For two functions $u, v \in H^1(\nu_X)$, the inner products between their gradients is defined as

$$(\nabla u, \nabla v)_{L^2(\nu_X)} = \int_{\Omega}\sum_{k=1}^{d} D_k u D_k v d\nu_X.$$

Definition 1 (Continuous functions space): Let Ω be a bounded domain in $\mathbb{R}^d$ and $s \in \mathbb{N}$. Let $C^s(\Omega)$ denote the vector space consisting of all functions f which, together with all their partial derivatives $D^{\alpha}f$ of orders $\Vert\alpha\Vert_1 \leq s$, are continuous on Ω. The Banach space $C^s(\Omega)$ is equipped with the norm

$$\Vert f\Vert_{C^s(\Omega)} := \max_{\Vert\alpha\Vert_1\leq s}\sup_{x\in\Omega}|D^{\alpha}f(x)|,$$

where $D^{\alpha} = D_1^{\alpha_1}\cdots D_d^{\alpha_d}$ with $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}^d$.

Next, we introduce the concept of a deep neural network. While deep ReLU neural networks have shown empirical

TABLE II
THEORETICAL RESULTS FOR APPLICATIONS

Nonparametric Variable Selection			
Reg. Param.		Convergence Rates	
$\lambda = \mathcal{O}(n^{-\frac{s}{d^*+4s}} \log^2 n)$	$\mathbb{E}\|\hat{f}_{\mathcal{D}}^{\lambda} - f_0\|^2$ $\mathbb{E}\|\nabla(\hat{f}_{\mathcal{D}}^{\lambda} - f_0)\|^2$	$\mathcal{O}(n^{-\frac{2s}{d^*+4s}} \log^4 n)$ $\mathcal{O}(n^{-\frac{s}{d^*+4s}} \log^2 n)$	Corollary VI.1
Selection consistency	Rel. Set $\mathcal{I}(f_0)$ Esti. Rel. Set $\mathcal{I}(\hat{f}_{\mathcal{D}}^{\lambda})$	$\lim_{n\to\infty} \Pr\{\mathcal{I}(f_0) = \mathcal{I}(\hat{f}_{\mathcal{D}}^{\lambda})\} = 1$	Corollary VI.2

success in nonparametric regression tasks, they are not suitable for scenarios where derivatives of the network are required in the objective function [46]. This limitation arises from the piecewise linear nature of the ReLU activation function, which results in a lack of continuous derivatives. In contrast, the Rectified Quadratic Unit (ReQU) activation function, defined as the square of the ReLU function, possesses a continuous first derivative. This characteristic allows us to incorporate the deep ReQU neural network in the SDORE framework, thereby expanding the possibilities for the simultaneous estimation of regression values and their derivatives.

Definition 2 (Deep ReQU Neural Network): A neural network $\psi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ is a function defined by

$$\psi(x) = T_L(\varrho(T_{L-1}(\cdots \varrho(T_0(x))\cdots))), \tag{4}$$

where the ReQU activation function $\varrho(x) = (\max\{x, 0\})^2$ is applied component-wisely and $T_\ell(x) := A_\ell x + b_\ell$ is an affine transformation with $A_\ell \in \mathbb{R}^{N_{\ell+1}\times N_\ell}$ and $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell = 0, \ldots, L$. In this paper, we consider the case $N_0 = d$ and $N_{L+1} = 1$. The number L is called the depth of the neural network, and the number $\max_{1\leq\ell\leq L} N_\ell$ is called the width of the neural network.Additionally, $\sum_{\ell=0}^{L}(\|A_\ell\|_0 + \|b_\ell\|_0)$ represents the total number of non-zero weights within the neural network. The space of deep ReQU neural networks with given network architecture is defined as

$$\mathcal{N}(L, W, S) := \Big\{\psi \text{ is of the form (4)} : \max_{1\leq\ell\leq L} N_\ell \leq W,\ \sum_{\ell=0}^{L}(\|A_\ell\|_0 + \|b_\ell\|_0) \leq S\Big\}.$$

To measure the complexity of a function class, we next introduce the empirical covering number.

Definition 3 (Empirical Covering Number): Let $\mathcal{F}$ be a class of functions from Ω to $\mathbb{R}$ and $\mathcal{D} = \{X_i\}_{i=1}^n \subseteq \Omega$. Define the $L^p(\mathcal{D})$-norm of the function $f \in \mathcal{F}$ as

$$\|f\|_{L^p(\mathcal{D})} = \left(\frac{1}{n}\sum_{i=1}^{n}|f(X_i)|^p\right)^{1/p}, \quad 1 \leq p < \infty.$$

For $p = \infty$, define $\|f\|_{L^\infty(\mathcal{D})} = \max_{1\leq i\leq n}|f(X_i)|$. A function set $\mathcal{F}_\delta$ is called an $L^p(\mathcal{D})$ δ-cover of $\mathcal{F}$ if for each $f \in \mathcal{F}$, there exits $f_\delta \in \mathcal{F}_\delta$ such that $\|f - f_\delta\|_{L^p(\mathcal{D})} \leq \delta$. Furthermore,

$$N(\delta, \mathcal{F}, L^p(\mathcal{D})) = \inf\Big\{|\mathcal{F}_\delta| : \mathcal{F}_\delta \text{ is a } L^p(\mathcal{D})\delta\text{-cover of } \mathcal{F}\Big\}$$

is called the $L^p(\mathcal{D})$ δ-covering number of $\mathcal{F}$.

We now introduce some basic notations. The set of positive integers is denoted by $\mathbb{N}_+ = \{1, 2, \ldots\}$. Denote $\mathbb{N} = \{0\} \cup \mathbb{N}_+$ for convenience. For a positive integer $m \in \mathbb{N}_+$, let $[m]$ denote the set $\{1, \ldots, m\}$. We employ the notations $A \lesssim B$ and $B \gtrsim A$ to signify that there exists an absolute constant $c > 0$ such that $A \leq cB$.

D. Organization

The remainder of the article is organized as follows. We commence with a review of related work in Section II. Subsequently, we outline the deep Sobolev penalized regression and propose the semi-supervised estimator in Section III. We present the convergence rate analysis for the regression in Section IV and for the derivative estimation in Section V. In Section VI, we apply our method to nonparametric variable selection, and provide an abundance of numerical studies. The article concludes with a few summarizing remarks in Section VII. All technical proofs are relegated to the supplementary material.

II. Related Work

In this section, we review the topics and literature related to this work, including derivative estimation, regression using deep neural network, nonparametric variable selection and semi-supervised learning.

A. Nonparametric Derivative Estimation

As previously indicated, the necessity to estimate derivatives arises in various application contexts. Among the simplest and most forthright methods for derivative estimation is the direct measurement of derivatives. For example, in the field of economics, estimating cost functions [17] frequently involves data on a function and its corresponding set of derivatives. A substantial volume of literature [47], [48], [49] considers this scenario by reverting to a corresponding regression model:

$$Y^\alpha = D^\alpha f_0(X) + \xi^\alpha,$$

where $\alpha \in \mathbb{N}^d$ is a multi-index, D^α is the α-th derivative operator, and ξ^α are random noise. The theoretical framework underpinning this method can be seamlessly generalized from that of classical nonparametric regression. However, it may be worth noting that in some practical application settings, measurements of derivatives are often not readily available.

To estimate derivatives with noisy measurements only on function values, researchers have put forward nonparametric derivative estimators [50]. Nonparametric derivative estimation encompasses four primary approaches: local polynomial regression [27], smoothing splines [28], kernel ridge regression [29], and difference quotients [30], [51], [52]. Among these approaches, the first three are categorized as plug-in derivative estimators. In this article, we present a review of these plug-in approaches using the one-dimensional case as an illustrative example.

1) Local Polynomial Regression: In standard polynomial regression, a single polynomial function is used to fit the data. One of the main challenges with this method is the need to use high-order polynomials to achieve a more accurate approximation. However, high-order polynomials may be oscillative in some regions, which is known as Runge phenomenon [53]. To repair the drawbacks of the polynomial regression, a natural way is to employ the low-degree polynomial regression locally, which is called local polynomial regression [54]. Derivative estimation using local polynomial regression was first proposed by [27]. Let K be a kernel function and h be the bandwidth controlling the smoothness. We assign a weight $K((X_i - x)/h)$ to the point (X_i, Y_i), leading to the following weighted least-squaress problem:

$$\min_{\{\beta_\ell(x)\}_{\ell=0}^p} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)\left(Y_i - \sum_{\ell=0}^p \beta_\ell(x)(X_i - x)^\ell\right)^2. \tag{5}$$

Herr the kernel K should decay fast enough to eliminate the impact of a remote data point. Denote by $\{\widehat{\beta}_\ell\}_{\ell=0}^p$ the estimator obtained by (5). The estimated regression curve at point x is given by $\widehat{f}(x) = \sum_{\ell=0}^p \widehat{\beta}_\ell(x)(X_i - x)^\ell$. Further, according to Taylor's theorem, the estimator of the first order derivative f_0' at point x is given by $\widehat{f}'(x) = \widehat{\beta}_1(x)$. References [55] and [56] established the uniform strong consistency and the convergence rates for the regression function and its partial derivatives. Derivative estimation using local polynomial regression in multivariate data has been discussed in [57].

2) Smoothing Splines: Extensive research has been conducted on the use of smoothing splines in nonparametric regression [2], [41], [58], [59]. This method starts from the minimization of a penalized least-squaress risk

$$\min_{f \in H^2([0,1])} \frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2 + \gamma \int_0^1 (f''(x))^2 dx, \tag{6}$$

where the first term encourages the fitting of estimator to data, the second term penalizes the roughness of the estimator, and the smoothing parameter $\gamma > 0$ controls the trade-off between the two conflicting goals. The minimizer $\widehat{f}$ of (6) is an estimator of the regression function f_0, which is called cubic smoothing spline. The plug-in derivative estimator $\widehat{f}'$ is a direct estimate of the derivative f_0' of the regression function. This idea has been pursued by [28] and [44]. In the perspective of theoretical analysis, [44] shows that spline derivative estimators can achieve the optimal rate of convergence, and [60] studies local asymptotic properties of derivative estimators.

3) Kernel Ridge Regression: Kernel ridge regression is a technique extensively employed in the domain of nonparametric regression. Reference [29] introduced a plug-in kernel ridge regression estimator for derivatives of the regression function, establishing a nearly minimax convergence rate for univariate function classes within a random-design setting. Further expanding upon this method, [33] applied it to multivariate regressions under the smoothing spline ANOVA model and established minimax optimal rates. Additionally, [33] put forth a hypothesis testing procedure intended to determine whether a derivative is zero.

B. Nonparametric Regression Using Deep Neural Network

In comparison to the nonparametric methods mentioned above, deep neural networks [32] also stand out as a formidable technique employed within machine learning and nonparametric statistics. Rigorous of the convergence rate analysis have been established for deep nonparametric regression [34], [35], [36], [37], [38], [39], [40], but derivative estimation using deep neural networks remained an open problem prior to this paper, even though the derivative of the regression estimate is of great importance as well.

Unfortunately, estimating derivatives is not always a by-product of function estimation. Indeed, the basic mathematical analysis [61, Section 3.7] shows that, even if estimators $\{f_n\}_{n\geq 1}$ converge to the regression function f_0, the convergence of plug-in derivative estimators $\{\nabla f_n\}_{n\geq 1}$ is typically not guaranteed. To give a counterexample, we consider the functions $f_n : I \to \mathbb{R}, x \mapsto n^{-1}\sin(nx)$, and let $f_0 : I \to \mathbb{R}$ be the zero function $f_0(x) = 0$, where $I := [0, 2\pi]$. Then $\|f_n - f_0\|_{L^p(I)} \to 0$ as $n \to 0$, but $\lim_{n\to\infty} \|f_n' - f_0'\|_{L^p(I)} \neq 0$ for each $1 \leq p \leq \infty$.

Roughly speaking, the success of classical approaches for derivative estimation can be attributed to their smoothing techniques, such as the kernel function incorporated in local polynomial regression, or the regularization in smoothing spline and kernel ridge regression. Thus, to guarantee the convergence of the plug-in derivative estimator, the incorporation of a Sobolev regularization term is imperative within the loss function, akin to the methodology applied in smoothing spline.

C. Nonparametric Vairable Selection

Data collected in real-world applications tend to be high-dimensional, although only a subset of the variables within the covariate vector may genuinely exert influence. Consequently, variable selection becomes critical in statistics and machine learning as it both mitigates computational complexity and enhances the interpretability of the model. However, traditional methods for variable selection have been primarily focused on linear or additive models and do not readily extend to nonlinear problems. One inclusive measure of the importance of each variable in a nonlinear model is its corresponding partial derivatives. Building on this concept, a series of works [23], [24], [25] introduced sparse regularization to kernel ridge regression for variable selection. They have consequently devised a feasible computational learning scheme and developed consistency properties of the estimator. However, the theoretical analysis is limited to reproducing kernel Hilbert space and cannot be generalized to deep neural network-based methods.

D. Semi-Supervised Learning

Semi-supervised learning has recently gained significant attention in statistics and machine learning [62], [63]. The basic setting of semi-supervised learning is common in many practical applications where the label is often more difficult or costly to collect than the covariate vector. Therefore, the fundamental question is how to design appropriate learning algorithms to fully exploit the value of unlabeled data. In the past years, significant effort has been devoted to studying the algorithms and theory of semi-supervised learning [64], [65], [66], [67], [68], [69], [70]. The most related work is [65], whose main idea is to introduce an unlabeled-data-driven regularization term to the loss function. Specifically, [65] employ a manifold regularization to incorporate additional information about the geometric structure of the marginal distribution, where the regularization term is estimated on the basis of unlabeled data. In addition, our method does not require the distribution of the unlabeled data to be aligned with the marginal distribution of the labeled data exactly, which expands the applicability scenarios.

III. Deep Sobolev Regression

In this section, we present an in-depth examination of Sobolev penalized least-squares regression as implemented through deep neural networks. Initially, we incorporate the H^1-semi-norm penalty into the least-squares risk. Subsequently, we delineate the deep Sobolev regressor as referenced in Section III-A, followed by an introduction to the semi-supervised Sobolev regressor elaborated in Section III-B.

We focus on the following $H^1(\nu_X)$-semi-norm penalized least-squares risk:

$$\min_{f\in\mathcal{A}} L^\lambda(f) = \mathbb{E}_{(X,Y)\sim\mu}\big[(f(X)-Y)^2\big] + \lambda\|\nabla f\|^2_{L^2(\nu_X)}, \tag{7}$$

where μ is a probability measure on $\Omega\times\mathbb{R}$ associated to the regression model (1), and ν_X is another probability measure on Ω. The admissible set $\mathcal{A}$ defined as

$$\mathcal{A} = \Big\{f\in L^2(\mu_X) : D_k f\in L^2(\nu_X),\ 1\le k\le d\Big\}.$$

Here the regularization parameter $\lambda>0$ governs the delicate equilibrium between conflicting objectives: data fitting and smoothness. Specifically, when λ nearly or entirely vanishes, (7) aligns with the standard population least-squares risk. Conversely, as λ approaches infinity, the minimizer of (7) tends towards a constant estimator. For the joint distribution μ of (X,Y), let μ_X denote the margin distribution of X. According to (1), one obtains easily

$$L^\lambda(f) = \|f-f_0\|^2_{L^2(\mu_X)} + \lambda\|\nabla f\|^2_{L^2(\nu_X)} + \mathbb{E}[\xi^2], \tag{8}$$

where the $L^2(\mu_X)$-risk may be respect to a different measure μ_X than that ν_X associated with Sobolev penalty. Throughout this paper, we assume that the distributions μ_X and ν_X have density function p and q, respectively. Furthermore, the density ratio $r(x) := q(x)/p(x)$ satisfies the following condition, which may encourage significant domain shift.

Assumption 1 (Uniformly bounded density ratio): The density ratio between ν_X and μ_X has a uniform upper-bound and a positive lower-bound, that is,

$$\kappa := \sup_{x\in\Omega}|r(x)| < \infty \quad \text{and} \quad \zeta := \inf_{x\in\Omega}|r(x)| > 0.$$

Sobolev penalized regression can be interpreted as a PDE-based smoother of the regression function f_0. Let f^λ denote a solution to the quadratic optimization problem (7). Some standard calculus of variations [71], [72] show that, if the minimizer f^λ has square integrable second derivatives, then f^λ solves the following second-order linear elliptic equation with homogeneous Neumann boundary condition:

$$\begin{cases} -\lambda\Delta f^\lambda + f^\lambda &= f_0, \text{in } \Omega, \\ \nabla f^\lambda\cdot\mathbf{n} &= 0, \text{on } \partial\Omega. \end{cases}$$

In the context of partial differential equations (PDE), the variational problem (7) is called Ritz method [71, Remark 2.5.11]. The following lemma shows the uniqueness of solution to the above PDE.

Lemma 1 (Existence and Uniqueness of Population Risk Minimizer): Suppose Assumption 1 holds and $f_0\in L^2(\mu_X)$. Then (7) has a unique minimizer in $H^1(\nu_X)$. Furthermore, the minimizer f^λ satisfies $f^\lambda\in H^2(\nu_X)$.

In practical applications, the data distribution μ in (7) remains unknown, making the minimization of population risk (7) unattainable. The goal of regression is to estimate the function f_0 from a finite set of data pairs $\mathcal{D}=\{(X_i,Y_i)\}_{i=1}^n$ which are independently and identically drawn from μ, that is,

$$Y_i = f_0(X_i) + \xi_i, \quad i=1,\ldots,n.$$

We introduce two Sobolev regressor based on the random sample $\mathcal{D}$ in the following two subsections, respectively.

A. Deep Sobolev Regressor

Suppose that the probability measure ν_X is either provided or selected by the user. Then the regularization term can be estimated with an arbitrarily small error. Hence, without loss of generality, this error is omitted in this discussion. In this setting, the deep Sobolev regressor is derived from the regularized empirical risk minimization:

$$\widehat{f}^\lambda_{\mathcal{D}} \in \arg\min_{f\in\mathcal{F}} \widehat{L}^\lambda_{\mathcal{D}}(f) = \frac{1}{n}\sum_{i=1}^n (f(X_i)-Y_i)^2 + \lambda\|\nabla f\|^2_{L^2(\nu_X)}, \tag{9}$$

where $\mathcal{F}\subseteq\mathcal{A}$ is a class of deep neural networks.

The objective functional in (9) has been investigated previously within the literature of splines, according to research by [1], [41], [42] and [43]. However, in these studies, minimization was undertaken within the Sobolev space $H^1(\Omega)$ or the continuous function space $C^1(\Omega)$ as opposed to within a class of deep neural networks.

B. Semi-Supervised Deep Sobolev Regressor

In numerous application scenarios, the probability measure ν_X remains unknown and cannot be provided by the user. Nevertheless, a substantial quantity of samples drawn from ν_X

can be obtained at a very low cost. This is a semi-supervised setting that provides access to labeled data and a relatively large amount of unlabeled data.

Let $\mathcal{S} = \{Z_i\}_{i=1}^m$ be a random sample with $\{Z_i\}_{i=1}^m$ independently and identically drawn from ν_X. Then replacing the population regularization term in (9) by its data-driven counterpart, we obtain the following semi-supervised empirical risk minimizer

$$\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 + \frac{\lambda}{m}\sum_{i=1}^{m}\sum_{k=1}^{d}|D_k f(Z_i)|^2, \tag{10}$$

where the deep neural network class $\mathcal{F}$ satisfies $\mathcal{F} \subseteq W^{1,\infty}(\Omega)$. A similar idea was mentioned by [65] in the context of manifold learning.

The estimator presented in (10), which incorporates unlabeled data into a supervised learning framework, is commonly referred to as a semi-supervised estimator. The availability of labeled data is often limited due to its high cost, but in many cases, there is an abundance of unlabeled data that remains underutilized. Given that there are no strict constraints on the measure ν_X in our method, it is possible to generate a substantial amount of unsupervised data from supervised data through data augmentation, even without a large quantity of unlabeled data. Hence, this semi-supervised learning framework exhibits a broad range of applicability across various scenarios.

It is worth highlighting that when the measure ν_X is equal to μ_X, the formulation (10) is reduced to

$$\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} \in \operatorname*{arg\,min}_{f\in\mathcal{F}} \widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(f) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i)-Y_i)^2 + \frac{\lambda}{n+m}\sum_{i=1}^{n+m}\sum_{k=1}^{d}|D_k f(X_i)|^2, \tag{11}$$

where $X_{n+i} = Z_i$ for $1 \leq i \leq m$. The semi-supervised Sobolev regressor, deployed in (10) or (11), imparts meaningful insights on how to leverage unlabeled data to enhance the efficacy of original supervised learning approach.

IV. Deep Sobolev Regressor With Gradient-Norm Constraint

In this section, we provide a theoretical analysis for the deep Sobolev regressor (9). The first result, given in Lemma 2, is an oracle-type inequality, which provides an upper-bound for the $L^2(\mu_X)$-error of the deep Sobolev regressor along with an upper-bound for the $L^2(\nu_X)$-norm of its gradient. Further, we show that (9) attains the minimax optimal convergence rate, given that the regularization parameter are chosen appropriately. We also confirm that the gradient norm of the deep Sobolev regressor can be uniformly bounded by a constant.

Assumption 2 (Sub-Gaussian noise): The noise ξ in (1) is sub-Gaussian with mean 0 and finite variance proxy σ^2 conditioning on $X = x$ for each $x \in \Omega$, that is, its conditional moment generating function satisfies

$$\mathbb{E}[\exp(t\xi)|X=x] \leq \exp\left(\frac{\sigma^2 t^2}{2}\right), \quad \forall\, t \in \mathbb{R},\ x \in \Omega.$$

Assumption 3 (Bounded hypothesis): There exists an absolute positive constant B_0, such that $\sup_{x\in\Omega}|f_0(x)| \leq B_0$. Further, functions in hypothesis class $\mathcal{F}$ are also bounded, that is, $\sup_{x\in\Omega}|f(x)| \leq B_0$.

Assumptions 2 and 3 are standard and very mild conditions in nonparametric regression, as extensively discussed in the literature [1], [3], [35], [36], [38], [40]. It is worth noting that the upper-bound B_0 of hypothesis may be arbitrarily large and does not vary with the sample size n. In fact, this assumption can be removed through the technique of truncation, without affecting the subsequent proof, which can be found in [34], [37], and [39] for details.

The convergence rate relies on an oracle-type inequality as follows.

Lemma 2 (Oracle inequality): Suppose Assumptions 1 to 3 hold. Let $\widehat{f}^{\lambda}_{\mathcal{D}}$ be the deep Sobolev regressor defined as (9) with regularization parameter $\lambda > 0$. Then it follows that for each $n \geq \log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\widehat{f}^{\lambda}_{\mathcal{D}} - f_0\|^2_{L^2(\mu_X)}\Big] \\
&\lesssim \inf_{f\in\mathcal{F}}\Big\{\|f-f_0\|^2_{L^2(\mu_X)} + \lambda\|\nabla f\|^2_{L^2(\nu_X)}\Big\} \\
&\quad + (B_0^2+\sigma^2)\inf_{\delta>0}\left\{\frac{\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\}, \\
&\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\nabla\widehat{f}^{\lambda}_{\mathcal{D}}\|^2_{L^2(\nu_X)}\Big] \\
&\lesssim \inf_{f\in\mathcal{F}}\left\{\frac{1}{\lambda}\|f-f_0\|^2_{L^2(\mu_X)} + \|\nabla f\|^2_{L^2(\nu_X)}\right\} \\
&\quad + \frac{B_0^2+\sigma^2}{\lambda}\inf_{\delta>0}\left\{\frac{\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\}.
\end{aligned}$$

Roughly speaking, the first inequality of Lemma 2 decomposes the $L^2(\mu_X)$-error of the deep Sobolev regressor into three terms, namely: the approximation error, the regularization term, and the generalization error. Intriguingly, from the perspective of the first two terms, we need to find a deep neural network in $\mathcal{F}$ that not only has an sufficiently small $L^2(\mu_X)$-distance from the regression function f_0, but also has an $H^1(\nu_X)$-semi-norm as small as possible.

The literature on deep learning theory has extensively investigated the approximation properties of deep neural networks [40], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82]. However, there is limited research on the approximation error analysis for neural networks with gradient norm constraints [83], [84]. The following lemma illustrates the approximation power of deep ReQU neural networks with gradient norm constraints.

Lemma 3 (Approximation With Gradient Constraints): Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Set the hypothesis class as a deep ReQU neural network class $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then for each $\phi \in C^s(K)$ with $s \in \mathbb{N}_{\geq 1}$, there exists a neural network $f \in \mathcal{F}$ such that

$$\begin{aligned}
\|f-\phi\|_{L^2(\mu_X)} &\leq CN^{-s}\|\phi\|_{C^s(K)}, \\
\|\nabla f\|_{L^2(\nu_X)} &\leq \|\nabla\phi\|_{L^2(\nu_X)} + C\|\phi\|_{C^s(K)},
\end{aligned}$$

where C is a constant independent of N.

This lemma provides a novel approximation error bound of deep ReQU networks with gradient norm constraint. This highlights a fundamental difference between deep ReLU and ReQU neural networks. As presented by [83] and [84], the gradient norm of deep ReLU networks goes to infinity when the approximation error diminishes. In contrast, Lemma 3 demonstrates that deep ReQU neural networks, under a gradient norm constraint, can approximate the target function with an arbitrarily small error.

With the aid of the preceding lemmas, we can now establish the following convergence rates for the regularized estimator.

Theorem 1 (Convergence rates): Suppose Assumptions 1 to 3 hold. Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Assume that $f_0 \in C^s(K)$ with $s \in \mathbb{N}_{\geq 1}$. Set the hypothesis class as a deep ReQU neural network class $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log n)$ and $S = \mathcal{O}\left(n^{\frac{d}{d+2s}}\right)$. Let $\widehat{f}_{\mathcal{D}}^{\lambda}$ be the deep Sobolev regressor defined as (9) for each $\lambda > 0$. Then it follows that

$$\mathbb{E}_{\mathcal{D}\sim\mu^n}\left[\|\widehat{f}_{\mathcal{D}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\right] \leq \mathcal{O}(\lambda) + \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^3 n\right),$$

$$\mathbb{E}_{\mathcal{D}\sim\mu^n}\left[\|\nabla\widehat{f}_{\mathcal{D}}^{\lambda}\|_{L^2(\nu_X)}^2\right] \leq \mathcal{O}(1) + \mathcal{O}\left(\lambda^{-1}n^{-\frac{2s}{d+2s}}\log^3 n\right).$$

Further, setting $\lambda = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^3 n\right)$ implies

$$\mathbb{E}_{\mathcal{D}\sim\mu^n}\left[\|\widehat{f}_{\mathcal{D}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\right] \leq \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^2 n\right),$$

$$\mathbb{E}_{\mathcal{D}\sim\mu^n}\left[\|\nabla\widehat{f}_{\mathcal{D}}^{\lambda}\|_{L^2(\nu_X)}^2\right] \leq \mathcal{O}(1).$$

Here the constant behind the big $\mathcal{O}$ notation is independent of n.

Theorem 1 quantifies how the regularization parameter λ balances two completing goals: data fitting and the gradient norm of the estimator, and thus provides an a priori guidance for the selection of the regularization term. When one chooses $\lambda = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^3 n\right)$, the rate of the deep Sobolev regressor $\mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^3 n\right)$ aligns with the minimax optimal rate up to a log-factor, as established in [1], [3], [85], and [86]. Additionally, our theoretical findings correspond to those in nonparametric regression using deep neural networks [34], [35], [36], [37], [38], [39], [40]. In contrast to standard empirical risk minimizers, the deep Sobolev regressor imposes a constraint on the gradient norm while simultaneously ensuring the minimax optimal convergence rate. Consequently, Sobolev regularization improves the stability and enhances the generalization abilities of deep neural networks.

A similar problem has been explored by researchers within the context of splines [1], [42], [43], where the objective functional aligns with that of the deep Sobolev regressor (9). However, in these studies, minimization was token over the Sobolev space $H^1(\Omega)$ or the continuous function space $C^1(\Omega)$ instead of a deep neural network class. The consistency in this setting was studied by [42], and the convergence rate was proven to be minimax optimal by [43] or [1, Theorem 21.2]. It is worth noting that the rate analysis in these studies relies heavily on the theoretical properties of the spline space and cannot be generalized to our setting.

V. Simultaneous Estimation of Regression Function and Its Derivative

In this section, we demonstrate that under certain mild conditions, deep Sobolev regressors converge to the regression function in both the $L^2(\mu_X)$-norm and the $H^1(\nu_X)$-semi-norm. We establish rigorous convergence rates for both the deep Sobolev regressor and its semi-supervised counterpart. Additionally, we provide a priori guidance for selecting the regularization parameter and determining the appropriate size of neural networks.

To begin with, we define the convex-hull of the neural network class $\mathcal{F}$, denoted as conv$(\mathcal{F})$. Subsequently, we proceed to redefine both the deep Sobolev regressor (9) and its semi-supervised counterpart (10) as

$$\widehat{f}_{\mathcal{D}}^{\lambda} \in \operatorname*{arg\,min}_{f\in\mathrm{conv}(\mathcal{F})} \widehat{L}_{\mathcal{D}}^{\lambda}(f), \quad \widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda} \in \operatorname*{arg\,min}_{f\in\mathrm{conv}(\mathcal{F})} \widehat{L}_{\mathcal{D},\mathcal{S}}^{\lambda}(f). \tag{12}$$

Notice that the functions within the convex-hull conv$(\mathcal{F})$ are also deep neural networks, which can be implemented by the parallelization of neural networks [87], [88]. Therefore, in the algorithmic implementation, solving (12) will only result in mere changes compared to solving in the original problem (9) or (10).

Throughout this section, suppose the following assumptions are fulfilled.

Assumption 4 (Regularity of Regression Function): The regression function in (1) satisfies $\Delta f_0 \in L^2(\nu_X)$ and $\nabla f_0 \cdot \mathbf{n} = 0$ a.e. on $\partial\Omega$, where $\mathbf{n}$ is the unit normal to the boundary.

Since there are no measurements available on the boundary $\partial\Omega$ or out of the domain Ω, it is not possible to estimate the derivatives on the boundary accurately. Hence, to simplify the problem without loss of generality, we assume that the underlying regression f_0 has zero normal derivative on the boundary, as stated in Assumption 4. This assumption corresponds to the homogeneous Neumann boundary condition in the context of partial differential equations [72].

We also make the following assumption regarding the regularity of the density function.

Assumption 5 (Bounded Score Function): The score function of the probability measure ν_X is bounded in $L^2(\nu_X)$-norm, that is, $\|\nabla(\log q)\|_{L^2(\nu_X)} < \infty$.

A sufficient condition for Assumption 5 is that ∇q is uniformly upper bounded and q has a uniform positive lower bound. In fact, this stronger assumption is mild and standard for a distribution ν_X.

In the following lemma, we show that the population Sobolev penalized risk minimizer f^{λ} converges to the regression function f_0 in $L^2(\mu_X)$-norm with rate $\mathcal{O}(\lambda^2)$. Additionally, the $L^2(\nu_X)$-rate of its derivatives is $\mathcal{O}(\lambda)$.

Lemma 4: Suppose Assumptions 1, 4 and 5 hold. Let f^{λ} be the unique minimizer of the population risk (7). Then it follows that for each $\lambda > 0$

$$\begin{aligned}&\|f^{\lambda} - f_0\|_{L^2(\mu_X)}^2\\ &\lesssim \lambda^2\kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\},\\ &\|\nabla(f^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\\ &\lesssim \lambda\kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\}.\end{aligned}$$

Up to now, we have shown the convergence of the population Sobolev penalized risk minimizer. However, researchers are primarily concerned with convergence rates of the empirical estimators obtained via a finite number of labeled data pairs $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$. In the remaining part of this section, we mainly focus on the convergence rate analysis for the deep Sobolev regressor and its semi-supervised counterpart in (12).

A. Analysis for Deep Sobolev Regressor

The theoretical foundation for simultaneous estimation of the regression function and its gradient is the following oracle-type inequality.

Lemma 5 (Oracle Inequality): Suppose Assumptions 1 to 5 hold. Let $\widehat{f}_{\mathcal{D}}^{\lambda}$ be the deep Sobolev regressor defined as (12). Then it follows that for each $\lambda > 0$ and each $n \geq \log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))$,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\widehat{f}_{\mathcal{D}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\Big]\\&\lesssim \beta\lambda^2 + \varepsilon_{\mathrm{app}}(\mathcal{F}, \lambda) + \varepsilon_{\mathrm{gen}}(\mathcal{F}, n),\\&\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\Big]\\&\lesssim \beta\lambda + \lambda^{-1}\varepsilon_{\mathrm{app}}(\mathcal{F}, \lambda) + \lambda^{-1}\varepsilon_{\mathrm{gen}}(\mathcal{F}, n),\end{aligned}$$

where β is a positive constant defined as

$$\beta = \kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\},$$

the approximation error $\varepsilon_{\mathrm{app}}(\mathcal{F}, \lambda)$ and the generalization error $\varepsilon_{\mathrm{gen}}(\mathcal{F}, n)$ are defined, respectively, as

$$\begin{aligned}&\varepsilon_{\mathrm{app}}(\mathcal{F}, \lambda)\\&= \inf_{f\in\mathcal{F}}\Big\{\|f - f_0\|_{L^2(\mu_X)}^2 + \lambda\|\nabla(f - f_0)\|_{L^2(\nu_X)}^2\Big\},\\&\varepsilon_{\mathrm{gen}}(\mathcal{F}, n)\\&= \frac{B_0^2 + \sigma^2}{\log^{-1} n}\inf_{\delta>0}\left\{\left(\frac{2\log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))}{n}\right)^{\frac{1}{2}} + \delta\right\}.\end{aligned}$$

As discussed in Section IV, [1, Chapter 21] has investigated an optimization problem similar to the deep Sobolev regressor. However, to the best of our knowledge, we are the first to demonstrate the oracle inequality for the gradient of estimator. The proof employs a similar technique as that of Lemma 4. Specifically, the deep Sobolev regressor acts as the minimizer of (12), which implies that it satisfies a variational inequality derived from the first-order optimality condition [89], [90]. By utilizing standard techniques from statistical learning theory, we are able to derive the desired oracle inequality.

In simple terms, if we select an appropriate neural network class and have a sufficiently large number of labeled data pairs, we can make the approximation error and generalization error arbitrarily small. Consequently, the overall error is primarily determined by the regularization parameter λ. At this point, the error bound aligns with rates in Lemma 4.

Recall the oracle inequality derived in Lemma 2, which requires the neural network to approximate the regression function while restricting its gradient norm. In contrast, the approximation term in Lemma 5 necessitates the neural network to approximate both the regression function and its derivatives simultaneously. Thus, we now introduce the following approximation error bound in H^1-norm.

Lemma 6 (Approximation in H^1-norm): Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Set the hypothesis class as a deep ReQU neural network $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then for each $\phi \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$, there exists $f \in \mathcal{F}$ such that

$$\begin{aligned}\|f - \phi\|_{L^2(\mu_X)} &\leq CN^{-s}\|\phi\|_{C^s(K)},\\ \|\nabla(f - \phi)\|_{L^2(\nu_X)} &\leq CN^{-(s-1)}\|\phi\|_{C^s(K)},\end{aligned}$$

where C is a constant independent of N.

With the aid of previously prepared lemmas, we have following convergence rates for the deep Sobolev regressor.

Theorem 2 (Convergence Rates): Suppose Assumptions 1 to 5 hold. Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Assume that $f_0 \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$. Set the hypothesis class as a deep ReQU neural network class $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log n)$ and $S = \mathcal{O}\left(n^{\frac{d}{d+4s}}\right)$. Let $\widehat{f}_{\mathcal{D}}^{\lambda}$ be the deep Sobolev regressor defined in (12) with regularization parameter $\lambda > 0$. Then it follows that

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\widehat{f}_{\mathcal{D}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\Big]\\&\leq \mathcal{O}(\lambda^2) + \mathcal{O}\left(n^{-\frac{2s}{d+4s}}\log^4 n\right),\\&\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\Big]\\&\leq \mathcal{O}(\lambda) + \mathcal{O}\left(\lambda^{-1}n^{-\frac{2s}{d+4s}}\log^4 n\right).\end{aligned}$$

Further, setting $\lambda = \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right)$ implies

$$\begin{aligned}\mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\widehat{f}_{\mathcal{D}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\Big] &\leq \mathcal{O}\left(n^{-\frac{2s}{d+4s}}\log^4 n\right),\\ \mathbb{E}_{\mathcal{D}\sim\mu^n}\Big[\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\Big] &\leq \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right).\end{aligned}$$

Here the constant behind the big $\mathcal{O}$ notation is independent of n.

Theorem 2 provides theoretical guidance for the selection of the size of neural networks and the choice of regularization parameters. In comparison to the regularization parameter $\lambda = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^3 n\right)$ employed in Theorem 1, $\lambda = \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right)$ utilized in Theorem 2 is much larger. The $L^2(\mu_X)$-rate $\mathcal{O}\left(n^{-\frac{2s}{d+4s}}\right)$ of the deep Sobolev regressor does not attain the minimax optimality. Furthermore, the convergence rate $\mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^4 n\right)$ for the derivatives is also slower than the minimax optimal rate $\mathcal{O}\left(n^{-\frac{2(s-1)}{d+2s}}\right)$ derived in [85].

B. Analysis for Semi-Supervised Deep Sobolev Regressor

In scenarios where the distribution ν_X is unknown, estimating the Sobolev penalty using the unlabeled data becomes crucial. In qualitative terms, having a sufficiently large number of unlabeled data points allows us to estimate the regularization term with an arbitrarily small error. However, the following questions are not answered quantitatively:

How does the error of the semi-supervised estimator depend on the number of unlabeled data? How does the unlabeled data in semi-supervised learning improve the standard supervised estimators?

In this section, we provide a comprehensive and rigorous analysis for the semi-supervised deep Sobolev regressor. To begin with, we present the following oracle inequality.

Assumption 6 (Bounded Derivatives of Hypothesis): There exists positive constants $\{B_{1,k}\}_{k=1}^{d}$, such that $\sup_{x\in\Omega}|D_k f_0(x)| \leq B_{1,k}$ for $1 \leq k \leq d$. Further, the first-order partial derivatives of functions in hypothesis class $\mathcal{F}$ are also bounded, i.e., $\sup_{x\in\Omega}|D_k f(x)| \leq B_{1,k}$ for each $1 \leq k \leq d$ and $f \in \mathcal{F}$. Denote by $B_1^2 := \sum_{k=1}^{d} B_{1,k}^2$

The inclusion of Assumption 6 is essential in the analysis of generalization error that involves derivatives, as it plays a similar role to Assumption 3 in the previous analysis.

Lemma 7 (Oracle Inequality): Suppose Assumptions 1 to 6 hold. Let $\widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda}$ be the semi-supervised deep Sobolev regressor defined in (12). For each $\lambda > 0$, $n \geq \log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))$ and $m \geq \max_{1\leq k\leq d} \log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))$,

$$\begin{aligned}
&\mathbb{E}_{(\mathcal{D},\mathcal{S})\sim\mu^n\times\nu_X^m}\Big[\|\widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\Big] \\
&\lesssim \tilde{\beta}\lambda^2 + \varepsilon_{\mathrm{app}}(\mathcal{F},\lambda) \\
&\quad + \varepsilon_{\mathrm{gen}}(\mathcal{F},n) + \varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},m), \\
&\mathbb{E}_{(\mathcal{D},\mathcal{S})\sim\mu^n\times\nu_X^m}\Big[\|\nabla(\widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\Big] \\
&\lesssim \tilde{\beta}\lambda + \lambda^{-1}\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda) \\
&\quad + \lambda^{-1}\varepsilon_{\mathrm{gen}}(\mathcal{F},n) + \lambda^{-1}\varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},m),
\end{aligned}$$

where $\tilde{\beta}$ is a positive constant defined as $\tilde{\beta} = \beta + B_1^2$, the approximation error $\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)$ and the generalization error $\varepsilon_{\mathrm{gen}}(\mathcal{F},n)$ are defined as those in Lemma 5. The generalization error $\varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},m)$ corresponding to the regularization term are defined as

$$\varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},m) = B_1^2 \inf_{\delta>0}\left\{\max_{1\leq k\leq d}\frac{N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m} + \delta\right\}.$$

In comparison to Lemma 5, the error bound has not undergone significant changes, and it has only been augmented by one additional generalization error associated with the regularization term. Further, this term vanishes as the number of unlabeled data increases.

In particular, we focus on the scenario where the distributions of covariates in both labeled and unlabeled data are identical, i.e., $\nu_X = \mu_X$. When only the labeled data pairs (e.g., (2)) are used, the generalization error corresponding to the regularization term is denoted as $\varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},n)$. In contrast, for the semi-supervised Sobolev regressor, the corresponding generalization term becomes:

$$\begin{aligned}
&\varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},m+n) \\
&= B_1^2 \inf_{\delta>0}\left\{\max_{1\leq k\leq d}\frac{\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m+n} + \delta\right\}.
\end{aligned}$$

It is worth noting that for every $m \in \mathbb{N}_{\geq 1}$, the inequality $\varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},m+n) < \varepsilon_{\mathrm{gen}}^{\mathrm{reg}}(\nabla\mathcal{F},n)$ holds. This demonstrates the provable advantages of incorporation of unlabeled data in the semi-supervised learning framework.

Finally, we derive convergence rates of the semi-supervised deep Sobolev regressor.

Theorem 3 (Convergence Rates): Suppose Assumptions 1 to 6 hold. Let $\Omega \subseteq K \subseteq \mathbb{R}^d$ be two bounded domain. Assume that $f_0 \in C^s(K)$ with $s \in \mathbb{N}_{\geq 2}$. Set the hypothesis class as a deep ReQU neural network class $\mathcal{F} = \mathcal{N}(L, W, S)$ with $L = \mathcal{O}(\log n)$ and $S = \mathcal{O}\left(n^{\frac{d}{d+4s}}\right)$. Let $\widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda}$ be the regularized empirical risk minimizer defined as (10) with regularization parameter $\lambda = \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right)$. Then it follows that

$$\begin{aligned}
&\mathbb{E}_{(\mathcal{D},\mathcal{S})\sim\mu^n\times\nu_X^m}\Big[\|\widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda} - f_0\|_{L^2(\mu_X)}^2\Big] \\
&\leq \mathcal{O}\left(n^{-\frac{2s}{d+4s}}\log^4 n\right) + \mathcal{O}\left(n^{\frac{d}{d+4s}}\log^4 nm^{-1}\right), \\
&\mathbb{E}_{(\mathcal{D},\mathcal{S})\sim\mu^n\times\nu_X^m}\Big[\|\nabla(\widehat{f}_{\mathcal{D},\mathcal{S}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\Big] \\
&\leq \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right) + \mathcal{O}\left(n^{\frac{d+s}{d+4s}}\log^2 nm^{-1}\right).
\end{aligned}$$

Here the constant behind the big $\mathcal{O}$ notation is independent of n.

For the number of unlabeled data m sufficiently large, the convergence rate of the semi-supervised deep Sobolev regressor tends to the rate derived in Theorem 2.

VI. Applications and Numerical Experiments

In this section, we demonstrate the effectiveness of our proposed SDORE in the context of derivative estimation, and nonparametric variable selection.

A. Derivative Estimation

In this section we give a one-dimensional example, and a detailed example in two dimensions is shown in Appendix F-A.

Example 1: Let the regression function be $f_0(x) = 1+36x^2-59x^3+21x^5+0.5\cos(\pi x)$. The labeled data pairs are generated from a regression model $Y = f_0(X) + \xi$, where X is sampled from the uniform distribution on $[0,1]$, and ξ is sampled from a Gaussian distribution $N(0,\sigma^2)$. Here the variance σ^2 is determined by a given signal-to-noise ratio $\frac{\mathbb{E}[f_0^2(X)]}{\sigma^2} = 30$. The unlabeled data are also drawn from the uniform distribution on $[0,1]$. The regularization parameter is set as $\lambda = 0.005$.

To demonstrate the effectiveness of SDORE in scenario where only few labeled sample is available, we conducted SDORE using 40 labeled data pairs and an additional 1000 unlabeled samples. The comparisons with the least-squares regression are presented in Figure 1, which includes point-wise comparisons of function values and derivatives. In the upper panel, the least-squares estimator generally matches the target function. However, the least-squares estimator fits the noise in the data rather than the underlying patterns near the left and right endpoints. Also, the lower panel shows that its estimated derivatives is inaccurate and unstable near the left and right endpoints. In comparison, our SDORE method successfully estimates the regression function and its derivatives simultaneously, and the regularization avoids the overfitting on the primitive function.

The errors in derivative estimates by SDORE are more pronounced near the interval boundary. This is primarily due to the lack of observations of function values outside the intervals, preventing accurate estimation of the boundary derivatives. From a theoretical perspective, the convergence of the derivative in L^2-norm is guaranteed by Theorem 1. However, this theorem does not provide guarantees for accuracy on the boundary. Estimating the boundary error requires the interior estimation of second-order derivatives, as outlined in the trace theorem [72, Theorem 1 in Section 5.5].

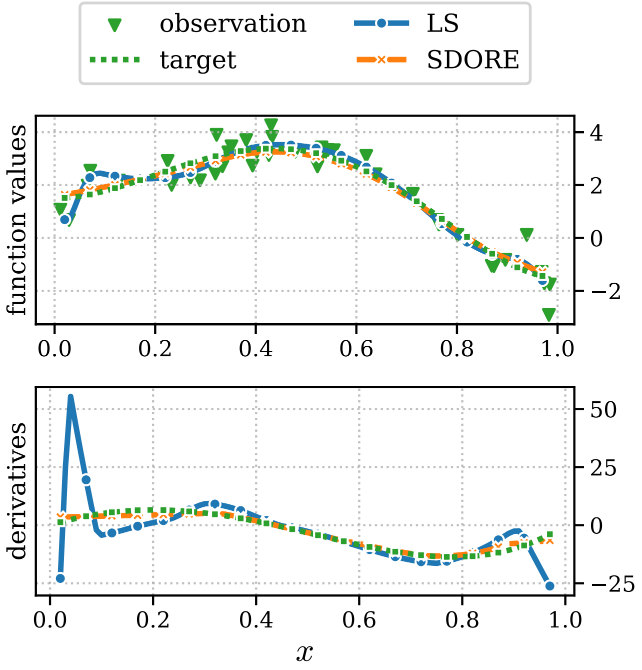

Fig. 1. Numerical results of Example 1. (left) Scatter plot of noisy observations (paired data used for supervised learning), line plot of the ground-truth regression function and its values predicted by least-squares (LS) regression and SDORE. (right) The ground truth derivative function and its estimated values by LS and SDORE.

B. Nonparametric Variable Selection

Deep neural network is a widely utilized tool in nonparametric statistics and machine learning. It effectively captures the nonlinear relationship between the covariate vector and the corresponding label. However, the interpretability of neural network estimators has faced significant criticism. This is primarily due to the inability to determine the relevance of variables in the covariate vector and quantify their impact on the neural network's output.

In this section, we propose a novel approach to address this issue by measuring the importance of a variable through its corresponding partial derivatives. Leveraging the deep Sobolev regressor, we introduce a nonparametric variable selection technique with deep neural networks. Remarkably, our method incorporates variable selection as a natural outcome of the regression process, eliminating the need for the design of a separate algorithm for this purpose.

Before proceeding, we impose additional sparsity structure on the underlying regression function, that is, there exists $f_0^*:\mathbb{R}^{d^*}\to\mathbb{R}$ $(1\le d^*\le d)$ such that

$$f_0(x_1,\dots,x_d)=f_0^*(x_{j_1},\dots,x_{j_{d^*}}),\qquad \{j_1,\dots,j_{d^*}\}\subseteq[d]. \tag{13}$$

This sparsity setting has garnered significant attention in the study of linear models and additive models, as extensively discussed in [91]. In the context of reproducing kernel Hilbert space, [23], [24], [25] introduced a nonparametric variable selection algorithm. Nevertheless, their approach and analysis heavily depend on the finite dimensional explicit representation of the estimator, making it unsuitable for generalizing to deep neural network estimators.

We introduce the definition of relevant set, which was proposed by [25, Definition 10]. The goal of the variable selection is to estimate the relevant set.

Definition 4 (Relevant Set): Let $f:\mathbb{R}^d\to\mathbb{R}$ be a differentiable function. A variable $k\in[d]$ is irrelevant for the function f with respect to the probability measure ν_X, if $D_kf(X)=0$ ν_X-almost surely, and relevant otherwise. The set of relevant variables is defined as

$$\mathcal{I}(f)=\{k\in[d]:\|D_kf\|_{L^2(\nu_X)}>0\}.$$

1) Convergence Rates and Selection Consistency:

Assumption 7 (Sparsity of the Regression Function): The number of relevant variables is less than the dimension d, that is, there exists a positive integer $d^*\le d$, such that $|\mathcal{I}(f_0)|=d^*$.

Under Assumption 7, our focus is solely on estimating the low-dimensional function f_0^* in (13) using deep neural networks. Consequently, the approximation and generalization error in Lemma 5 are reliant solely on the intrinsic dimension d^*. This implies an immediate result as follows.

Corollary 1: Suppose Assumptions 1 to 7 hold. Let $\Omega\subseteq K\subseteq\mathbb{R}^d$ be two bounded domain. Assume that $f_0\in C^s(K)$ with $s\in\mathbb{N}_{\ge2}$. Set the hypothesis class $\mathcal{F}$ as a ReQU neural network class $\mathcal{F}=\mathcal{N}(L,W,S)$ with $L=\mathcal{O}(\log n)$ and $S=\mathcal{O}\left(n^{\frac{d^*}{d^*+4s}}\right)$. Let $\widehat{f}_{\mathcal{D}}^{\lambda}$ be the regularized empirical risk minimizer defined as (10) with regularization parameter $\lambda=\mathcal{O}\left(n^{-\frac{s}{d^*+4s}}\log^2 n\right)$. Then the following inequality holds

$$\mathbb{E}_{\mathcal{D}\sim\mu^n}\left[\|\widehat{f}_{\mathcal{D}}^{\lambda}-f_0\|_{L^2(\mu_X)}\right]\le\mathcal{O}\left(n^{-\frac{s}{d^*+4s}}\log^4 n\right),$$
$$\mathbb{E}_{\mathcal{D}\sim\mu^n}\left[\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda}-f_0)\|_{L^2(\nu_X)}\right]\le\mathcal{O}\left(n^{-\frac{s}{2(d^*+4s)}}\log^2 n\right).$$

The convergence rate presented in Corollary 1 is solely determined by the intrinsic dimension d^* and remains unaffected by the data dimension d, which effectively mitigates the curse of dimensionality when d^* is significantly smaller than d.

Furthermore, we establish the selection properties of the deep Sobolev regressor, which directly follow from the convergence of derivatives.

Corollary 2 (Selection Consistency): Under the same conditions as Corollary 1. It follows that

$$\lim_{n\to\infty}\Pr\left\{\mathcal{I}(f_0)=\mathcal{I}(\widehat{f}_{\mathcal{D}}^{\lambda})\right\}=1,$$

where $\lambda=\mathcal{O}\left(n^{-\frac{s}{d^*+4s}}\log^2 n\right)$.

Corollary 2 demonstrates that, given a sufficiently large number of data pairs, the estimated relevant set $\mathcal{I}(\widehat{f}_{\mathcal{D}}^{\lambda})$ is equal to the ground truth relevant set $\mathcal{I}(f_0)$ with high probability. In comparison, [25, Theorem 11] only provided a one-side consistency

$$\lim_{n\to\infty}\Pr\left\{\mathcal{I}(f_0)\subseteq\mathcal{I}(\widehat{f}_{\mathcal{D}}^{\lambda})\right\}=1,$$

were unable to establish the converse inclusion.

2) Numerical Experiments: In this section, we present a high-dimensional example which has sparsity structure to verify the performance of SDORE in variable selection. The

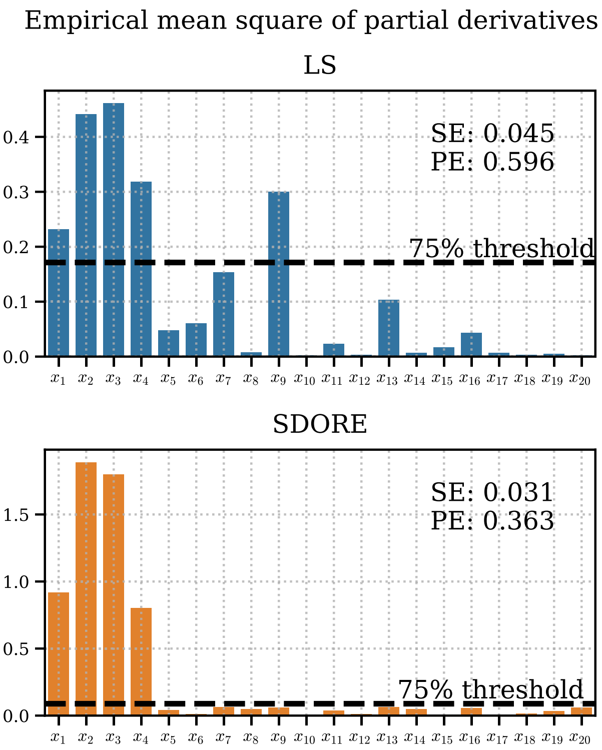

Fig. 2. Numerical results of Example 2. The empirical mean square of the partial derivatives of the regression function f_0 (which depends only on the x_1 to x_4), estimated by least-squares fitting (LS, left) and SDORE (right). The dashed line is the 75% quantile threshold for variable selection. We also report the mean selection error (SE) for the estimated derivative function and the root mean squared prediction error (PE) for the primitive function by each method.

additional experiments for variable selection are shown in Appendix F-B.

Example 2: Let the regression function be

$$f_0(x) = \sum_{i=1}^{3} \sum_{j=i+1}^{4} x_i x_j.$$

Suppose the covariate in both labeled and unlabeled data are sampled from the uniform distribution on $[0,1]^{20}$. The label Y is generated from the regression model $Y = f_0(X)+\xi$, where ξ is the noise term sampled from a Gaussian distribution with the signal-to-noise ratio to be 25, in the same way as Example 1. The regularization parameter is set as $\lambda = 1.0 \times 10^{-2}$.

In real-world applications, the process of labeling data can be prohibitively costly, resulting in a limited availability of labeled data. Conversely, there is an abundance of unlabeled data that is readily accessible. Hence, it becomes crucial to leverage few labeled data alongside a substantial amount of unlabeled data for the purpose of variable selection. Nevertheless, the task of variable selection with few labeled samples presents significant challenges. Due to the scarcity of data points, there is a restricted range of variability within the dataset, posing difficulties in accurately determining the variables that hold true significance in predicting the desired outcome.

To demonstrate the effectiveness of SDORE in this challenging scenario, we employ SDORE for the variable selection in this example, utilizing 50 labeled data pairs and an additional sample containing 100 unlabeled covariate vectors. Additionally, we use least-squares regression on the same data as a comparison. Figure 2 visually presents the empirical mean square (EMS) of estimated partial derivatives with respect to each variable on the test set, that is, for each $1 \leq k \leq d$,

$$\mathrm{EMS}_k = \frac{1}{n} \sum_{i=1}^{n} |D_k \widehat{f}(X_{i,k})|^2.$$

Here $\widehat{f}$ is an estimator, $\{X_i\}_{i=1}^n$ is a set of test data, and $X_{i,k}$ represents the k-th element of X_i. The results by SDORE reveals that the derivatives with respect to relevant variables x_1 to x_4 are significantly larger than those of the other variables, while least-squares regression wrongly regards x_9 as relevant variables, possibly due to the lack of paired training samples. This shows that our proposed method can estimate the derivatives accurately, which facilitates the variable selection. We select the variables by setting a 75% quantile threshold of the estimated partial derivatives. The partial derivatives greater than the threshold is considered relevant. Additionally, Figure 2 displays the mean selection error (SE), calculated as the mean of the false positive rate and false negative rate as defined by [25], as well as the root mean squared prediction error (PE) on the regression function. Notably, the results consistently demonstrate the superior performance of SDORE over least-squares regression, underscoring the advantages of incorporating unlabeled data.

Remark 1: Since in Figure 2, for SDORE, the estimated partial derivatives with respect to the first four features is significantly larger than others, we can choose the threshold directly. In other application scenarios, if we can not observe such a clear difference, we can employ the strategy such as cross-validation (CV) to determine the number of features. Specifically, the cross-validation process involves dividing the dataset into a training set and a validation set independently. The training set is used for Sobolev regression, and the model's performance is evaluated on the validation set. The mean square of partial derivatives, also known as the important score, is sorted from largest to smallest. The cross-validation process begins by selecting the feature with the largest important score, and adds the remaining most important features incrementally until the accuracy in the validation set no longer shows improvement.

VII. Conclusion

In this paper, we present a novel semi-supervised deep Sobolev regressor that allows for the simultaneous estimation of the underlying regression function and its gradient. We provide a thorough convergence rate analysis for this estimator, demonstrating the provable benefits of incorporating unlabeled data into the semi-supervised learning framework. To the best of our knowledge, these results are original contributions to the literature in the field of deep learning, thereby enhancing the theoretical understanding of semi-supervised learning and gradient penalty strategy. From an application standpoint, our approach introduces powerful new tools for nonparametric variable selection. Moreover, our method has demonstrated exceptional performance in various numerical examples, further validating its efficacy.

We would like to highlight the generality of our method and analysis, as it can be extended to various loss functions. In our upcoming research, we have extended the Sobolev penalized strategy to encompass a wide range of statistical and machine learning tasks, such as density estimation, deconvolution, classification, and quantile regression. Furthermore, we have discovered the significant role that the semi-supervised deep Sobolev regressor plays in addressing inverse problems related to partial differential equations. There still remains some challenges that need to be addressed. For example, in Theorem 2, the $L^2(\mu_X)$-rate $\mathcal{O}\left(n^{-\frac{2s}{d+4s}}\log^4 n\right)$ of the deep Sobolev regressor does not attain the minimax optimality. Additionally, the convergence rate $\mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right)$ for the derivatives is also slower than the minimax optimal rate $\mathcal{O}\left(n^{-\frac{2(s-1)}{d+2s}}\right)$ derived in [85]. Moreover, while Corollary 2 establishes selection consistency, it does not provide the rate of convergence. Furthermore, an interesting avenue for future research would be to investigate deep nonparametric regression with a sparse/group sparse penalty.

Appendix A
Supplemental Definitions and Lemmas

In this section, we present some definitions and lemmas for preparation. We first extend Green's formula in Lebesgue measure to general measures.

Lemma 8 (Green's Formula in General Measure): Let ν_X be a probability measure on Ω with density function $q(x) \in W^{1,\infty}(\Omega)$. Let $u \in H^1(\nu_X)$ and let $v \in H^2(\nu_X)$ satisfying $\nabla v \cdot \mathbf{n} = 0$ a.e. on $\partial\Omega$, where $\mathbf{n}$ is the unit normal to the boundary. Then it follows that

$$-(\nabla u, \nabla v)_{L^2(\nu_X)} = (\Delta v + \nabla v \cdot \nabla(\log q), u)_{L^2(\nu_X)}$$

Proof of Lemma 8 It is straightforward that

$$\begin{aligned}
&-(\nabla u, \nabla v)_{L^2(\nu_X)} \\
&= -\int_\Omega \nabla u \cdot \nabla v q dx \\
&= -\int_\Omega \nabla \cdot (\nabla v q u) dx + \int_\Omega \nabla \cdot (\nabla v q) u dx \\
&= -\int_{\partial\Omega} (\nabla v \cdot \mathbf{n}) u q ds + \int_\Omega \nabla \cdot (\nabla v q) u dx \\
&= \int_\Omega \Delta v u q dx + \int_\Omega \nabla v \cdot \nabla(\log q) u q dx \\
&= (\Delta v, u)_{L^2(\nu_X)} + (\nabla v \cdot \nabla(\log q), u)_{L^2(\nu_X)},
\end{aligned}$$

where the second equality holds from integration by parts, the third equality follows from the divergence theorem [72, Theorem 1 in Section C.2], and the forth one used the assumption $\nabla v \cdot \mathbf{n} = 0$ and the equality $\nabla(\log q) = \nabla q / q$. ■

We next present the maximal inequality for sub-Gaussian variables.

Lemma 9: Let ξ_j be σ^2-sub-Gaussian for each $1 \le j \le N$. Then

$$\mathbb{E}\Big[\max_{1\le j\le N} \xi_j^2\Big] \le 4\sigma^2(\log N + 1).$$

Proof of Lemma 9 By Jensen's inequality, it is straightforward that

$$\begin{aligned}
\exp\left(\frac{\lambda}{2\sigma^2}\mathbb{E}\Big[\max_{1\le j\le N}\xi_j^2\Big]\right) &\le \mathbb{E}\left[\max_{1\le j\le N}\exp\left(\frac{\lambda\xi_j^2}{2\sigma^2}\right)\right] \\
&\le N\mathbb{E}\left[\exp\left(\frac{\lambda\xi_1^2}{2\sigma^2}\right)\right] \le \frac{N}{\sqrt{1-\lambda}},
\end{aligned}$$

where the last inequality holds from [91, Theorem 2.6] for each $\lambda \in [0, 1)$. Letting $\lambda = 1/2$ yields the desired inequality. ■

To measure the complexity of a function class, we next introduce the Vapnik-Chervonenkis (VC) dimension and some associated lemmas.

Definition 5 (VC-Dimension): Let $\mathcal{F}$ be a class of functions from Ω to $\{\pm 1\}$. For any non-negative integer m, we define the growth function of $\mathcal{F}$ as

$$\Pi_{\mathcal{F}}(m) = \max_{\{x_i\}_{i=1}^m \subseteq \Omega} \left|\{(f(x_1), \dots, f(x_m)) : f \in \mathcal{F}\}\right|.$$

A set $\{x_i\}_{i=1}^m$ is said to be shattered by $\mathcal{F}$ when

$$|\{(f(x_1), \dots, f(x_m)) : f \in \mathcal{F}\}| = 2^m.$$

The VC-dimension of $\mathcal{F}$, denoted $\mathrm{VCdim}(\mathcal{F})$, is the size of the largest set that can be shattered by $\mathcal{F}$, that is, $\mathrm{VCdim}(\mathcal{F}) = \max\{m : \Pi_{\mathcal{F}}(m) = 2^m\}$. For a class $\mathcal{F}$ of real-valued functions, we define $\mathrm{VCdim}(\mathcal{F}) = \mathrm{VCdim}(\mathrm{sign}(\mathcal{F}))$.

The following lemma provides a VC-dimension bound for the empirical covering number.

Lemma 10 ([92, Theorem 12.2]): Let $\mathcal{F}$ be a set of real functions from Ω to the bounded interval $[-B, B]$. Let $\delta \in (0, 1)$ and $\mathcal{D} = \{X_i\}_{i=1}^n \subseteq \Omega$. Then for each $1 \le p \le \infty$ and $n \ge \mathrm{VCdim}(\mathcal{F})$, the following inequality holds

$$\log N(\delta, \mathcal{F}, L^p(\mathcal{D})) \le c\,\mathrm{VCdim}(\mathcal{F}) \log(nB\delta^{-1}),$$

where $c > 0$ is an absolute constant.

Lemma 10 demonstrates that the metric entropy of a function class is bounded by its VC-dimension. The following lemma provides a VC-dimension bound for a deep neural network classes with a piecewise-polynomial activation function, and with a fixed architecture, i.e., the positions of the nonzero parameters are fixed.

Lemma 11 ([93, Theorem 7]): Let $\mathcal{N}$ be a deep neural network architecture with L layers and S non-zero parameters. The activation function is piecewise-polynomial. Then $\mathrm{VCdim}(\mathcal{N}) \le cLS \log(S)$, where $c > 0$ is an absolute constant.

With the help of Lemmas 10 and 11, we can bound the metric entropy of the deep neural networks by its depth and number of nonzero parameters as the following lemma. The proof of this lemma is inspired by [36, Lemma 5].

Lemma 12: Let $\mathcal{N} \subseteq \mathcal{N}(L, W, S)$ be a set of deep neural networks from Ω to the bounded interval $[-B, B]$. The activation function is piecewise-polynomial. Let $\delta \in (0, 1)$ and $\mathcal{D} = \{X_i\}_{i=1}^n \subseteq \Omega$. Then

$$\log N(\delta, \mathcal{N}, L^2(\mathcal{D})) \le cLS \log(S) \log\left(\frac{nB}{\delta}\right),$$

where $c > 0$ is an absolute constant.

Proof of Lemma 12: Before proceeding, it follows from the technique of removal of inactive nodes [36, eq. (9)] that

$$\mathcal{N} \subseteq \mathcal{N}(L, W, S) = \mathcal{N}(L, W \wedge S, S). \tag{14}$$

For each deep neural network in $\mathcal{N}(L, W, S)$, the number of parameters T satisfies

$$\begin{aligned} T &:= \sum_{\ell=0}^{L} (N_\ell + 1) N_{\ell+1} \le (L+1) 2^{-L} \prod_{\ell=0}^{L} (N_\ell + 1) \\ &\le \prod_{\ell=0}^{L} (N_\ell + 1) \le (W+1)^{L+1} \le (S+1)^{L+1}, \end{aligned} \tag{15}$$

where the last inequality is due to (14). Then there exist $\binom{T}{s}$ combinations to pick s non-zero parameters from all T parameters, which yields a partition

$$\begin{aligned} \mathcal{N}^s &:= \left\{ \phi \in \mathcal{N} : \sum_{\ell=0}^{L} (\|A_\ell\|_0 + \|b_\ell\|_0) = s \right\} \\ &= \left\{ \mathcal{N}_1^s, \ldots, \mathcal{N}_m^s \right\}, \quad m_s = \binom{T}{s}, \end{aligned}$$

where the deep neural networks in the same subset have the same positions of the non-zeros parameters. Consequently,

$$\begin{aligned} &N(\delta, \mathcal{N}, L^2(\mathcal{D})) \\ &= \sum_{s=1}^{S} N(\delta, \mathcal{N}^s, L^2(\mathcal{D})) = \sum_{s=1}^{S} \sum_{i=1}^{m_s} N(\delta, \mathcal{N}_i^s, L^2(\mathcal{D})) \\ &\le \sum_{s=1}^{S} \sum_{i=1}^{m_s} \left(\frac{nB}{\delta} \right)^{\mathrm{VCdim}(\mathcal{N}_i^s)} \le \sum_{s=1}^{S} \binom{T}{s} \left(\frac{nB}{\delta} \right)^{cLs \log(s)} \\ &\le \sum_{s=1}^{S} (S+1)^{(L+1)s} \left(\frac{nB}{\delta} \right)^{cLs \log(s)} \\ &\le (S+1)^{(L+1)(S+1)} \left(\frac{nB}{\delta} \right)^{cL(S+1) \log(S)}, \end{aligned}$$

where the first inequality holds from Lemma 10, and the second inequality follows from Lemma 11. The third inequality used the inequality $\binom{T}{s} \le T^s$ and (15). Taking logarithm on both sides of the inequality yields the desired result. ■

By an argument similar to [81, Lemma 5.7], we derive the following lemma, which shows that the first-order derivative of a ReQU neural network can be represented by a ReQU-ReLU network. With the help of this lemma and Lemma 12, we can bound the metric entropy of the class of derivatives of ReQU networks.

Lemma 13: Let $f : \mathbb{R}^d \to \mathbb{R}$ be a ReQU neural network with depth no more that L and the number of non-zero weights no more than S. Then $D_k f$ can be implemented by a ReQU-ReLU neural network with depth no more that cL and the number of non-zero weights no more than $c'LS$, where c and c' are two positive absolute constants.

Proof of Lemma 13: We prove this lemma by induction. For simplicity of presentation, we omit the intercept terms in this proof. Denote by $\varrho_1 = \max\{0, x\}$ and $\varrho_2 = (\max\{0, x\})^2$. It is straightforward to verify that

$$\varrho_2'(z) = 2\varrho_1(z), \tag{16}$$

and

$$\begin{aligned} yz = \frac{1}{4} \Big(&\varrho_2(y+z) + \varrho_2(-y-z) \\ &- \varrho_2(y-z) - \varrho_2(z-y) \Big). \end{aligned} \tag{17}$$

■

For the two-layers ReQU sub-network, the p-th element can be defined as

$$f_p^{(2)}(x) := \sum_{j \in [N_2]} a_{pj}^{(2)} \varrho_2 \left(\sum_{i \in [N_1]} a_{ji}^{(1)} x_i \right).$$

The number of non-zero weights of $f_p^{(2)}$ is given by

$$S_{2,p} := \sum_{j \in [N_2] : a_{pj}^{(2)} \neq 0} \|(a_{ji}^{(1)})_{i=1}^{N_1}\|_0.$$

By some simple calculation, we have that for each $1 \le k \le d$,

$$\begin{aligned} D_k f_p^{(2)}(x) &= \sum_{j \in [N_2]} a_{pj}^{(2)} D_k \varrho_2 \left(\sum_{i \in [N_1]} a_{ji}^{(1)} x_i \right) \\ &= 2 \sum_{j \in [N_2]} a_{pj}^{(2)} \varrho_1 \left(\sum_{i \in [N_1]} a_{ji}^{(1)} x_i \right) a_{jk}^{(1)}, \end{aligned}$$

where the last equality holds from (16). Thus $D_k f_p^{(2)}$ can be implemented by a ReLU network with 2 layers and the number of non-zero weights is same to $f_p^{(2)}$, that is, $S_{2,p}^{\prime,k} = S_{2,p}$.

For the three layers ReQU sub-network, by a same argument, we have

$$f_p^{(3)}(x) := \sum_{j \in [N_3]} a_{pj}^{(3)} \varrho_2(f_j^{(2)}(x)),$$

the number of non-zeros weights of which is

$$S_{3,p} := \sum_{j \in [N_3] : a_{pj}^{(3)} \neq 0} S_{2,j}.$$

Then its derivatives are given by

$$\begin{aligned} D_k f_p^{(3)}(x) &= \sum_{j \in [N_3]} a_{pj}^{(3)} D_k \varrho_2(f_j^{(2)}(x)) \\ &= 2 \sum_{j \in [N_3]} a_{pj}^{(3)} \varrho_1(f_j^{(2)}(x)) D_k f_j^{(2)}(x) \\ &= \frac{1}{2} \sum_{j \in [N_3]} a_{pj}^{(3)} \Big\{ \varrho_2\Big(\varrho_1(f_j^{(2)}(x)) + D_k f_j^{(2)}(x) \Big) \\ &\quad + \varrho_2\Big(-\varrho_1(f_j^{(2)}(x)) - D_k f_j^{(2)}(x) \Big) \\ &\quad - \varrho_2\Big(\varrho_1(f_j^{(2)}(x)) - D_k f_j^{(2)}(x) \Big) \\ &\quad - \varrho_2\Big(-\varrho_1(f_j^{(2)}(x)) + D_k f_j^{(2)}(x) \Big) \Big\}, \end{aligned}$$

where the second equality holds from (16) and the last one is due to (17). This implies that $D_k f_p^{(3)}$ can be implemented by a ReQU-ReLU mixed network with 4 layers. Furthermore, the number of non-zero weights of $D_k f_p^{(3)}$ is given by

$$S_{3,p}^{\prime,k} := \sum_{j \in [N_3] : a_{pj}^{(3)} \neq 0} \left(S_{2,j} + S_{2,j}^{\prime,k} + 12 \right)$$

$$\begin{aligned}&\le \sum_{j\in[N_3]:a_{pj}^{(3)}\neq 0}\left(S_{2,j}+13S_{2,j}^{\prime,k}\right)\\&=14\sum_{j\in[N_3]:a_{pj}^{(3)}\neq 0}S_{2,j}=14S_{3,p}.\end{aligned}\tag{18}$$

We claim that the depth of $D_kf_p^{(\ell-1)}$ is no more than $2\ell-2$ and the number of non-zero weights satisfies

$$S_{\ell,p}^{\prime,k}\le 13\ell S_{\ell,p},\quad 3\le\ell\le L.\tag{19}$$

The case of $\ell=3$ has be shown in (18), and it remains to verify that this inequality also holds for ℓ, provided that (19) holds for $\ell-1$.

According to (19), suppose that $D_kf_j^{(\ell-1)}$ has $2(\ell-1)-2$ layers and no more than $13(\ell-1)S_{\ell-1,p}^{\prime,k}$ non-zero weights for $j\in[N_{\ell-1}]$. Notice the p-th element of the ℓ-th layer are given by

$$f_p^{(\ell)}(x):=\sum_{j\in[N_\ell]}a_{pj}^{(\ell)}\varrho_2(f_j^{(\ell-1)}(x)),$$

the number of non-zeros weights of which is

$$S_{\ell,p}:=\sum_{j\in[N_\ell]:a_{pj}^{(\ell)}\neq 0}S_{\ell-1,j}.$$

Then its derivatives are defined as

$$\begin{aligned}&D_kf_p^{(\ell)}(x)\\&=\sum_{j\in[N_\ell]}a_{pj}^{(\ell)}D_k\varrho_2(f_j^{(\ell-1)}(x))\\&=2\sum_{j\in[N_\ell]}a_{pj}^{(\ell)}\varrho_1(f_j^{(\ell-1)}(x))D_kf_j^{(\ell-1)}(x)\\&=\frac{1}{2}\sum_{j\in[N_\ell]}a_{pj}^{(\ell)}\Big\{\varrho_2\Big(\varrho_1(f_j^{(\ell-1)}(x))+D_kf_j^{(\ell-1)}(x)\Big)\\&\quad+\varrho_2\Big(-\varrho_1(f_j^{(\ell-1)}(x))-D_kf_j^{(\ell-1)}(x)\Big)\\&\quad-\varrho_2\Big(\varrho_1(f_j^{(\ell-1)}(x))-D_kf_j^{(\ell-1)}(x)\Big)\\&\quad-\varrho_2\Big(-\varrho_1(f_j^{(\ell-1)}(x))+D_kf_j^{(\ell-1)}(x)\Big)\Big\}.\end{aligned}$$

Hence $D_kf_p^{(\ell)}$ has $2(\ell-1)-2+2$ layers and the number of non-zero weights of $D_kf_p^{(\ell)}$ is given by

$$\begin{aligned}S_{\ell,p}^{\prime,k}&:=\sum_{j\in[N_\ell]:a_{pj}^{(\ell)}\neq 0}\left(S_{\ell-1,j}+S_{\ell-1,j}^{\prime,k}+12\right)\\&\le\sum_{j\in[N_\ell]:a_{pj}^{(\ell)}\neq 0}\left(S_{\ell-1,j}+13(\ell-1)S_{\ell-1,j}+12S_{\ell-1,j}\right)\\&=13S_{\ell,p},\end{aligned}$$

which deduces (19) for ℓ. Therefore, we complete the proof.

Remark 2: Notice that both ReLU and ReQU are piecewise-polynomial activation functions. By the proof of Lemma 11 in [93, Theorem 7], it is apparent that the VC-dimension bounds also hold for ReQU-ReLU neural networks, which are constructed in Lemma 13. In addition, see [81, Theorem 5.1] for a complete proof of the VC-dimension bound of ReQU-ReLU networks.

Combining Lemmas 12 and 13 yields the following results.

Lemma 14: Let $\mathcal{N}\subseteq\mathcal{N}(L,W,S)$ be a set of deep neural networks from Ω to the bounded interval $[-B,B]$. The activation function is piecewise-polynomial. Let $\delta\in(0,1)$ and $\mathcal{D}=\{X_i\}_{i=1}^n\subseteq\Omega$. Then

$$\log N(\delta,D_k\mathcal{N},L^2(\mathcal{D}))\le cL^2S\log(S)\log\left(\frac{nB}{\delta}\right),$$

where $c>0$ is an absolute constant.

We conclude this section by introducing an approximation error bound for deep ReQU neural networks.

Lemma 15 (Approximation Error): Let $\Omega\subseteq K\subseteq\mathbb{R}^d$ be two bounded domain. For each $\phi\in C^s(K)$ with $s\in\mathbb{N}_{\ge1}$, there exists a ReQU neural network f with the depth and the number of nonzero weights no more than $d\lfloor\log_2N\rfloor+d$ and $C'N^d$, respectively, such that $0\le k\le\min\{s,N\}$,

$$\inf_{f\in\mathcal{F}}\|f-\phi\|_{C^k(\Omega)}\le CN^{-(s-k)}\|\phi\|_{C^s(K)},$$

where C and C' are constants independent of N.

Proof of Lemma 15 We first approximate the target function $\phi\in C^s(K)$ by polynomials. According to [94, Theorem 2], for each $N\in\mathbb{N}$, these exists a polynomial p_N of degree at most N on $\mathbb{R}^d$ such that for $0\le|\gamma|\le\min\{s,N\}$,

$$\begin{aligned}&\sup_{x\in K}|D^\gamma(\phi(x)-p_N(x))|\\&\qquad\le\frac{C}{N^{s-|\gamma|}}\sum_{|\alpha|\le s}\sup_{x\in K}|D^\alpha\phi(x)|,\end{aligned}\tag{20}$$

where C is a positive constant depending only on d, s and K. Applying [79, Theorem 3.1], one obtains that there exists a ReQU neural network f with the depth $d\lfloor\log_2N\rfloor+d$ and nonzero weights no more than $C'N^d$, such that

$$f=p_N,\tag{21}$$

where C' is a constant independent of N. Combining (20) and 21 yields

$$\begin{aligned}\|f-\phi\|_{C^k(\Omega)}&\le\sup_{x\in K}|D^\gamma(f(x)-p_N(x))|\\&\le CN^{-(s-k)}\|\phi\|_{C^s(K)},\end{aligned}$$

for each $0\le k\le\min\{s,N\}$. This completes the proof. ■

Appendix B
Proofs of Results in Section III

Proofs of theoretical results in Section III are shown in this section.

Proof of Lemma 1 By (7) and the standard variational theory [72], it is sufficient to focus on the variational problem

$$\mathscr{B}(f^\lambda,g)=(f_0,g)_{L^2(\mu_X)},\quad\forall g\in H^1(\nu_X),\tag{22}$$

where the bilinear form $\mathscr{B}:H^1(\nu_X)\times H^1(\nu_X)\to\mathbb{R}$ is defined as

$$\mathscr{B}(f,g):=\lambda(\nabla f,\nabla g)_{L^2(\nu_X)}+(f,g)_{L^2(\mu_X)}.$$

It is straightforward to verify the boundedness and coercivity of the bilinear form from Assumption 1, that is,

$$|\mathscr{B}(f,g)|\le(\lambda\vee\zeta^{-1/2})\|f\|_{H^1(\nu_X)}\|g\|_{H^1(\nu_X)},$$

$$\mathscr{B}(f,f) \geq (\lambda \wedge \kappa^{-1/2})\|f\|_{H^1(\nu_X)},$$

for each $f, g \in H^1(\nu_X)$. Further, since that $f_0 \in L^2(\mu_X)$, the functional $F : H \to \mathbb{R}$, $g \mapsto (f_0, g)_{L^2(\mu_X)}$ is bounded and linear. Then according to Lax-Milgram theorem [72, Theorem 1 in Chapter 6.2], there exists a unique solution $f^\lambda \in H^1(\nu_X)$ to the varitional problem (22). This completes the proof of the uniqueness. See [95, Theorem 2.4.2.7] for the proof of the higher regularity of the solution. ■

Appendix C
Proofs of Results in Section IV

In this section, we demonstrate proofs of theoretical results in Section IV, including Lemma 2, Lemma 3 and Theorem 1. The proof of Lemma 2 uses the technique of offset Rademacher complexity, which has been investigated by [96].

Proof of Lemma 2 Recall the population excess risk $R(f)$ and the empirical excess risk $\widehat{R}_{\mathcal{D}}(f)$ defined in the proof of Lemma 5. We further define the regularized excess risk and regularized empirical risk as

$$R^\lambda(f) := R(f) + \lambda\|\nabla f\|^2_{L^2(\nu_X)},$$
$$\widehat{R}^\lambda_{\mathcal{D}}(f) := \widehat{R}_{\mathcal{D}}(f) + \lambda\|\nabla f\|^2_{L^2(\nu_X)}.$$

It suffices to shown that

$$\mathbb{E}_{\mathcal{D}}\Big[R^\lambda(\widehat{f}^\lambda_{\mathcal{D}})\Big] \lesssim \inf_{f\in\mathcal{F}} R^\lambda(f) + \frac{B_0^2+\sigma^2}{\log^{-1} n} \inf_{\delta>0}\left\{\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\}. \tag{23}$$

Before proceeding, we provide the proof sketch. Firstly, in *Step (I)*, we show that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[R^\lambda(\widehat{f}^\lambda_{\mathcal{D}}) - 2\widehat{R}^\lambda_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}})\Big] \\
&= \mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}^\lambda_{\mathcal{D}}) - 2\widehat{R}_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}})\Big] \\
&\leq cB_0^2 \inf_{\delta>0}\left\{\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\},
\end{aligned} \tag{24}$$

where c is an absolute positive constant. It remains to consider the regularized empirical risk. According to (1), we have

$$\begin{aligned}
&\widehat{L}^\lambda_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}}) \\
&= \widehat{R}^\lambda_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}}) - \frac{2}{n}\sum_{i=1}^n \xi_i(\widehat{f}^\lambda_{\mathcal{D}}(X_i) - f_0(X_i)) + \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i^2\right].
\end{aligned}$$

Taking expectation with respect to $\mathcal{D} \sim \mu^n$ on both sides of the equality yields that for each $f \in \mathcal{F}$,

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}^\lambda_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}})\Big] \\
&= \mathbb{E}_{\mathcal{D}}\Big[\widehat{L}^\lambda_{\mathcal{D}}(f)\Big] + 2\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^n \xi_i \widehat{f}^\lambda_{\mathcal{D}}(X_i)\right] - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i^2\right] \\
&\leq R^\lambda(f) + \frac{1}{2}\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}})\Big] \\
&\quad + c(B_0^2+\sigma^2)\inf_{\delta>0}\left\{\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\},
\end{aligned}$$

which implies

$$\begin{aligned}
&\widehat{R}^\lambda_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}}) \leq 2R^\lambda(f) \\
&\quad + 2c(B_0^2+\sigma^2)\inf_{\delta>0}\left\{\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\},
\end{aligned} \tag{25}$$

where c is an absolute positive constant. Here the inequality invokes

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^n \xi_i \widehat{f}^\lambda_{\mathcal{D}}(X_i)\right] \leq \frac{1}{4}\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}})\Big] \\
&\quad + \frac{8\sigma^2}{n}\inf_{\delta>0}\left\{\frac{\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n} + \delta\right\} \\
&\quad + 2(B_0^2+\sigma^2)\delta,
\end{aligned} \tag{26}$$

which is obtained in *Step (II)*. Combining (24) and (25) obtains (23).

Step (I). Given a ghost sample $\mathcal{D}' = \{(X_i', Y_i')\}_{i=1}^n$, where $\{X_i'\}_{i=1}^n$ are independently drawn from μ_X. Further, the ghost sample $\mathcal{D}'$ is independent of $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$. Let $\varepsilon = \{\varepsilon_i\}_{i=1}^n$ be a set of Rademacher variables and independent of $\mathcal{D}$ and $\mathcal{D}'$. Since that $\widehat{f}^\lambda_{\mathcal{D}} \in \mathcal{F}$, by the technique of symmetrization, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}^\lambda_{\mathcal{D}}) - 2\widehat{R}_{\mathcal{D}}(\widehat{f}^\lambda_{\mathcal{D}})\Big] \leq \mathbb{E}_{\mathcal{D}}\Big[\sup_{f\in\mathcal{F}} R(f) - 2\widehat{R}_{\mathcal{D}}(f)\Big] \\
&= \mathbb{E}_{\mathcal{D}}\Big[\sup_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{D}'}\left[\frac{1}{n}\sum_{i=1}^n (f(X_i') - f_0(X_i'))^2\right] \\
&\qquad - \frac{2}{n}\sum_{i=1}^n (f(X_i) - f_0(X_i))^2\Big] \\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(X_i') - f_0(X_i'))^2 \\
&\qquad - \frac{2}{n}\sum_{i=1}^n (f(X_i) - f_0(X_i))^2\Big] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\Big[\sup_{f\in\mathcal{F}} \frac{3}{2n}\sum_{i=1}^n (f(X_i') - f_0(X_i'))^2 \\
&\qquad - \frac{1}{2n}\sum_{i=1}^n (f(X_i') - f_0(X_i'))^2 \\
&\qquad - \frac{3}{2n}\sum_{i=1}^n (f(X_i) - f_0(X_i))^2 \\
&\qquad - \frac{1}{2n}\sum_{i=1}^n (f(X_i) - f_0(X_i))^2\Big] \\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\Big[\sup_{f\in\mathcal{F}} \frac{3}{2n}\sum_{i=1}^n (f(X_i') - f_0(X_i'))^2 \\
&\qquad - \frac{1}{8B_0^2 n}\sum_{i=1}^n (f(X_i') - f_0(X_i'))^4 \\
&\qquad - \frac{3}{2n}\sum_{i=1}^n (f(X_i) - f_0(X_i))^2 \\
&\qquad - \frac{1}{8B_0^2 n}\sum_{i=1}^n (f(X_i) - f_0(X_i))^4\Big] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\Big[\sup_{f\in\mathcal{F}} \frac{3}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f_0(X_i))^2
\end{aligned}$$

$$-\frac{1}{4B_0^2n}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\Big], \qquad (27)$$

where the second inequality follows from the convexity of supremum and Jensen's inequality, and the third inequality is owing to the fact that $0 \leq (f(X_i)-f_0(X_i))^2 \leq 4B_0^2$ for each $f \in \mathcal{F}$.

Let $\delta > 0$ and let $\mathcal{F}_\delta$ be an $L^2(\mathcal{D})$ $(B_0\delta)$-cover of $\mathcal{F}$ satisfying $|\mathcal{F}_\delta| = N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))$. Then it follows from Cauchy-Schwarz inequality that for each $f \in \mathcal{F}$, there exists $f_\delta \in \mathcal{F}_\delta$ such that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f_0(X_i))^2-\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f_\delta(X_i)-f_0(X_i))^2,\\ &\leq\Big(\frac{1}{n}\sum_{i=1}^n(f(X_i)+f_\delta(X_i)-2f_0(X_i))^2\\ &\quad\times(f(X_i)-f_\delta(X_i))^2\Big)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\right)^{1/2}\\ &\leq 4B_0^2\delta.\end{aligned}$$

By a same argument, we obtain

$$\begin{aligned}&\frac{1}{B_0^2n}\left(-\sum_{i=1}^n(f(X_i)-f_0(X_i))^4+\sum_{i=1}^n(f_\delta(X_i)-f_0(X_i))^4\right)\\ &\leq 32B_0^2\delta.\end{aligned}$$

Combining (42) with above two inequalities yields

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}_{\mathcal{D}}^{\lambda})-2\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda})\Big]-20B_0^2\delta\\ &\leq\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\Big[\max_{f\in\mathcal{F}_\delta}\frac{3}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f_0(X_i))^2\\ &\quad-\frac{1}{4B_0^2n}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\Big]. \qquad (28)\end{aligned}$$

In order to estimate the expectation in (28), we consider the following probability conditioning on $\mathcal{D}=\{(X_i,Y_i)\}_{i=1}^n$

$$\begin{aligned}&\mathrm{Pr}_{\varepsilon}\Big\{\frac{3}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f_0(X_i))^2\\ &\qquad>t+\frac{1}{4B_0^2n}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\Big\}.\end{aligned}$$

For a fixed sample $\mathcal{D}=\{(X_i,Y_i)\}_{i=1}^n$, the random variables $\{\varepsilon_i(f(X_i)-f_0(X_i))^2\}_{i=1}^n$ are independent and satisfy

$$\mathbb{E}_{\varepsilon}\big[\varepsilon_i(f(X_i)-f_0(X_i))^2\big]=0,$$

and for each $1 \leq i \leq n$,

$$\begin{aligned}-(f(X_i)-f_0(X_i))^2&\leq\varepsilon_i(f(X_i)-f_0(X_i))^2\\ &\leq(f(X_i)-f_0(X_i))^2.\end{aligned}$$

Consequently, it follows from Hoeffding's inequality [97, Lemma D.2] that

$$\begin{aligned}&\mathrm{Pr}_{\varepsilon}\Big\{\frac{3}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f_0(X_i))^2\\ &\qquad>t+\frac{1}{4B_0^2n}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\Big\}\\ &\leq\exp\left(-\frac{\left(\frac{nt}{3}+\frac{1}{12B_0^2}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\right)^2}{2\sum_{i=1}^n(f(X_i)-f_0(X_i))^4}\right)\\ &\leq\exp\left(-\frac{nt}{18B_0^2}\right),\end{aligned}$$

where we used the numeric inequality that $(a+y)^2/y \geq 4a$ for each $a>0$. Then with the aid of the above estimate of the tail probability, it follows that

$$\begin{aligned}&\mathbb{E}_{\varepsilon}\Big[\max_{f\in\mathcal{F}_\delta}\frac{3}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f_0(X_i))^2\\ &\quad-\frac{1}{4B_0^2n}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\Big]\\ &\leq T+N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))\int_T^\infty\exp\left(-\frac{nt}{18B_0^2}\right)dt\\ &=T+\frac{18B_0^2}{n}N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))\exp\left(-\frac{nT}{18B_0^2}\right).\end{aligned}$$

By setting $T=18B_0^2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))n^{-1}$, we deduces

$$\begin{aligned}&\mathbb{E}_{\varepsilon}\Big[\max_{f\in\mathcal{F}_\delta}\frac{3}{n}\sum_{i=1}^n\varepsilon_i(f(X_i)-f_0(X_i))^2\\ &\quad-\frac{1}{4B_0^2n}\sum_{i=1}^n(f(X_i)-f_0(X_i))^4\Big]\\ &\leq\frac{18B_0^2}{n}(1+\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))). \qquad (29)\end{aligned}$$

Combining (28) and (29) implies that

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}_{\mathcal{D}}^{\lambda})-2\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda})\Big]\\ &\leq\frac{18B_0^2}{n}(1+\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D})))+20B_0^2\delta. \qquad (30)\end{aligned}$$

This completes the proof of (24).

Step (II). Recall the $L^2(\mathcal{D})$ $(B_0\delta)$-cover $\mathcal{F}_\delta$ of the hypothesis class $\mathcal{F}$. There exists $f_\delta \in \mathcal{F}_\delta$ such that

$$\frac{1}{n}\sum_{i=1}^n|f_\delta(X_i)-\widehat{f}_{\mathcal{D}}^{\lambda}(X_i)|^2\leq(B_0\delta)^2,$$

which implies

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^n\xi_i(\widehat{f}_{\mathcal{D}}^{\lambda}(X_i)-f_\delta(X_i))\right]\\ &\leq\mathbb{E}_{\mathcal{D}}^{1/2}\left[\frac{1}{n}\sum_{i=1}^n\xi_i^2\right]\mathbb{E}_{\mathcal{D}}^{1/2}\left[\frac{1}{n}\sum_{i=1}^n(\widehat{f}_{\mathcal{D}}^{\lambda}(X_i)-f_\delta(X_i))^2\right]\\ &\leq B_0\sigma\delta, \qquad (31)\end{aligned}$$

and

$$\begin{aligned}&\widehat{R}_{\mathcal{D}}^{1/2}(f_\delta)\\ &\leq\left(\frac{1}{n}\sum_{i=1}^n(f_\delta(X_i)-\widehat{f}_{\mathcal{D}}^{\lambda}(X_i))^2\right)^{1/2}+\widehat{R}_{\mathcal{D}}^{1/2}(\widehat{f}_{\mathcal{D}}^{\lambda})\end{aligned}$$

$$\leq B_0\delta + \widehat{R}^{1/2}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}}), \tag{32}$$

where we used Cauchy-Schwarz inequality and Assumption 2. Consequently, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}^{\lambda}_{\mathcal{D}}(X_i)\right]\\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}^{\lambda}_{\mathcal{D}}(X_i)-f_0(X_i))\right]\\
&\leq \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(f_\delta(X_i)-f_0(X_i))\right]+B_0\sigma\delta\\
&\leq \mathbb{E}_{\mathcal{D}}\left[\frac{\widehat{R}^{1/2}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})+B_0\delta}{\sqrt{n}}\psi(f_\delta)\right]+B_0\sigma\delta\\
&\leq \left(\mathbb{E}^{1/2}_{\mathcal{D}}\left[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})\right]+B_0\delta\right)\frac{1}{\sqrt{n}}\mathbb{E}^{1/2}_{\mathcal{D}}\left[\psi^2(f_\delta)\right]+B_0\sigma\delta\\
&\leq \frac{1}{4}\mathbb{E}_{\mathcal{D}}\left[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})\right]+\frac{2}{n}\mathbb{E}_{\mathcal{D}}\left[\psi^2(f_\delta)\right]\\
&\quad+\frac{1}{4}B_0^2\delta^2+B_0\sigma\delta.
\end{aligned} \tag{33}$$

Here, the first inequality holds from (31), the second inequality is from (32), where

$$\psi(f_\delta) := \frac{\sum_{i=1}^{n}\xi_i(f_\delta(X_i)-f_0(X_i))}{\sqrt{n}\widehat{R}^{1/2}_{\mathcal{D}}(f_\delta)}.$$

The third inequality follows from Cauchy-Schwarz inequality, while the last one is owing to the inequality $ab \leq a^2/4 + b^2$ for $a, b > 0$. Observe that for each fixed f independent of ξ, the random variable $\psi(f)$ is sub-Gaussian with variance proxy σ^2. Then using Lemma 9 gives that

$$\mathbb{E}_\xi\left[\psi^2(f_\delta)\right] \leq \mathbb{E}_\xi\left[\max_{f\in\mathcal{F}_\delta}\psi^2(f)\right] \leq 4\sigma^2(\log|\mathcal{F}_\delta|+1). \tag{34}$$

Combining (33) and (34) yields (26). ■

Proof of Lemma 3 Using Lemma 15, by setting $k = 0$, we obtain the estimate in $L^2(\mu_X)$-norm. Further, setting $k = 1$ yields the estimate for the first-order derivative. This completes the proof. ■

Proof of Theorem 1 According to Lemma 4.2, we set the hypothesis class $\mathcal{F}$ as ReQU neural networks $\mathcal{F} = \mathcal{N}(L, S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then there exists $f \in \mathcal{F}$ such that $\|f - \phi\|_{L^2(\mu_X)} \leq CN^{-s}$ and $\|\nabla f\|_{L^2(\nu_X)} \leq C$. By using Lemma 12 and set $\delta = 1/n$, we find

$$\begin{aligned}
&\log N(B_0 n^{-1}, \mathcal{F}, L^2(\mathcal{D}))\\
&\qquad\lesssim LS\log S\log n \lesssim N^d\log^2 N\log n.
\end{aligned}$$

Substituting these estimates into Lemma 2 yields

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\|\widehat{f}^{\lambda}_{\mathcal{D}} - f_0\|^2_{L^2(\mu_X)}\right]\\
&\qquad\lesssim CN^{-2s} + C\lambda + C\log n\frac{N^d\log^2 N}{n\log^{-1}(n)}.
\end{aligned}$$

Letting $N = \mathcal{O}\left(n^{\frac{1}{d+2s}}\right)$ and $\lambda = \mathcal{O}\left(n^{-\frac{2s}{d+2s}}\log^3 n\right)$ deduces the desired result. ■

Appendix D
Proofs of Results in Section V

In this section, we show proofs of theoretical results in Section V. The proofs for the deep Sobolev regressor are shown in Section D-A, and proofs for semi-supervised deep Sobolev regressor are shown in Section D-B.

Proof of Lemma 4 It follows from (22) that

$$\begin{aligned}
&\lambda(\nabla(f^\lambda - f_0), \nabla h)_{L^2(\nu_X)} + (f^\lambda - f_0, h)_{L^2(\mu_X)}\\
&\quad = \lambda(\Delta f_0 + \nabla f_0\cdot\nabla(\log q), h)_{L^2(\nu_X)}, \quad \forall h \in H^1(\nu_X),
\end{aligned}$$

where we used Lemma 8 and Assumption 4. By setting $h = f^\lambda - f_0 \in H^1(\nu_X)$ and using Cauchy-Schwarz inequality, we derive

$$\begin{aligned}
&\lambda\|\nabla(f^\lambda - f_0)\|^2_{L^2(\nu_X)} + \|f^\lambda - f_0\|^2_{L^2(\mu_X)}\\
&\leq \lambda\|\Delta f_0 + \nabla f_0\cdot\nabla(\log q)\|_{L^2(\nu_X)}\|f^\lambda - f_0\|_{L^2(\nu_X)}\\
&\leq \lambda\kappa^{1/2}\|\Delta f_0 + \nabla f_0\cdot\nabla(\log q)\|_{L^2(\nu_X)}\\
&\quad\times\|f^\lambda - f_0\|_{L^2(\mu_X)}
\end{aligned} \tag{35}$$

which implies immediately

$$\|f^\lambda - f_0\|_{L^2(\mu_X)} \leq \lambda\kappa^{1/2}\|\Delta f_0 + \nabla f_0\cdot\nabla(\log q)\|_{L^2(\nu_X)}. \tag{36}$$

Substituting (36) into (35) deduces the estimate for the derivative, which completes the proof. ■

A. Deep Sobolev Regressor

Proof of Lemma 5 For simplicity of notation, we define the empirical inner-product and norm based on the sample $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, respectively, as

$$\begin{aligned}
(u, v)_{L^2(\mathcal{D})} &= \frac{1}{n}\sum_{i=1}^{n}u(X_i)v(X_i),\\
\|u\|^2_{L^2(\mathcal{D})} &= \frac{1}{n}\sum_{i=1}^{n}u^2(X_i),
\end{aligned}$$

for each $u, v \in L^\infty(\mu_X)$. Then we define the excess risk and its empirical counterpart, respectively, as

$$R(f) = \|f - f_0\|^2_{L^2(\mu_X)} \quad\text{and}\quad \widehat{R}_{\mathcal{D}}(f) = \|f - f_0\|^2_{L^2(\mathcal{D})}.$$

The proof is divided into five parts which are denoted by (I) to (V):

(I) We first relate the excess risk with its empirical counterpart:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[R(\widehat{f}^{\lambda}_{\mathcal{D}}) - \widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})\right]\\
&\leq 4B_0^2\left\{\left(\frac{2\log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))}{n}\right)^{\frac{1}{2}} + \delta\right\}.
\end{aligned} \tag{37}$$

(II) We next derive the following inequality for preparation:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}^{\lambda}_{\mathcal{D}}(X_i)\right]\\
&\leq \frac{B_0^2+\sigma^2}{\log^{-1}n}\left\{\left(\frac{2\log N(B_0\delta, \mathcal{F}, L^2(\mathcal{D}))}{n}\right)^{\frac{1}{2}} + \delta\right\}.
\end{aligned} \tag{38}$$

(III) With the help of variational inequality, we obtain the following inequality:

$$\begin{aligned}&\lambda\mathbb{E}_{\mathcal{D}}\Big[\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D}}-f_0)\|^2_{L^2(\nu_X)}\Big]+\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})\Big]\\ &\quad\le\frac{1}{8}\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D}})\Big]+2\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}^{\lambda}_{\mathcal{D}}(X_i)\right]\\ &\qquad+c\Big\{\beta\lambda^2+\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)\Big\},\end{aligned}\tag{39}$$

where c is an absolute positive constant. Here the constant β and the approximation error $\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)$ is defined as

$$\begin{aligned}&\beta=\kappa\Big\{\|\Delta f_0\|^2_{L^2(\nu_X)}+\|\nabla f_0\cdot\nabla(\log q)\|^2_{L^2(\nu_X)}\Big\},\\ &\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)\\ &\quad=\inf_{f\in\mathcal{F}}\Big\{\|f-f_0\|^2_{L^2(\mu_X)}+\lambda\|\nabla(f-f_0)\|^2_{L^2(\nu_X)}\Big\}.\end{aligned}$$

(IV) Combining (37), (38) and (39), we obtain an estimate for $L^2(\mu_X)$-error:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[\|\widehat{f}^{\lambda}_{\mathcal{D}}-f_0\|^2_{L^2(\mu_X)}\Big]\\ &\quad\lesssim\beta\lambda^2+\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)+\varepsilon_{\mathrm{gen}}(\mathcal{F},n),\end{aligned}\tag{40}$$

and an estimate for $L^2(\nu_X)$-error of the gradient:

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D}}-f_0)\|^2_{L^2(\nu_X)}\Big]\\ &\quad\lesssim\beta\lambda+\frac{\varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)}{\lambda}+\frac{\varepsilon_{\mathrm{gen}}(\mathcal{F},n)}{\lambda}.\end{aligned}\tag{41}$$

Here the generalization error $\varepsilon_{\mathrm{gen}}(\mathcal{F},n)$ is defined as

$$\begin{aligned}&\varepsilon_{\mathrm{gen}}(\mathcal{F},n)\\ &\quad=\frac{B_0^2+\sigma^2}{\log^{-1}n}\inf_{\delta>0}\left\{\left(\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n}\right)^{\frac12}+\delta\right\}.\end{aligned}$$

Step (I). Given a ghost sample $\mathcal{D}'=\{(X_i',Y_i')\}_{i=1}^n$, where $\{X_i'\}_{i=1}^n$ are independently and identically drawn from μ_X. Further, the ghost sample $\mathcal{D}'$ is independent of $\mathcal{D}=\{(X_i,Y_i)\}_{i=1}^n$. Let $\varepsilon=\{\varepsilon_i\}_{i=1}^n$ be a set of Rademacher variables and independent of $\mathcal{D}$ and $\mathcal{D}'$. Since that $\widehat{f}^{\lambda}_{\mathcal{D}}\in\mathrm{conv}(\mathcal{F})$, by the technique of symmetrization, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D}})-\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})\Big]\le\mathbb{E}_{\mathcal{D}}\Big[\sup_{f\in\mathrm{conv}(\mathcal{F})}R(f)-\widehat{R}_{\mathcal{D}}(f)\Big]\\ &=\mathbb{E}_{\mathcal{D}}\Bigg[\sup_{f\in\mathrm{conv}(\mathcal{F})}\mathbb{E}_{\mathcal{D}'}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i')-f_0(X_i'))^2\right]\\ &\quad-\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-f_0(X_i))^2\Bigg]\\ &\le\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\Bigg[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}(f(X_i')-f_0(X_i'))^2\\ &\quad-\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-f_0(X_i))^2\Bigg]\\ &=\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\mathcal{D}'}\mathbb{E}_{\varepsilon}\Bigg[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Big((f(X_i')-f_0(X_i'))^2\\ &\quad-(f(X_i)-f_0(X_i))^2\Big)\Bigg]\\ &=2\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i)-f_0(X_i))^2\right]\\ &\le4B_0\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(f(X_i)-f_0(X_i))\right]\\ &=4B_0\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(X_i)\right]\\ &=4B_0\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(X_i)\right],\end{aligned}\tag{42}$$

where the second inequality follows from the convexity of supremum and Jensen's inequality, and the third inequality holds from Ledoux-Talagrand contraction inequality [97, Lemma 5.7] and the fact that $0\le|f(X_i)-f_0(X_i)|\le2B_0$ for each $f\in\mathrm{conv}(\mathcal{F})$ and each $1\le i\le n$. The last equality invokes the fact that the Rademacher complexity of the convex hull of $\mathcal{F}$ is equal to that of $\mathcal{F}$.

Let $\delta>0$ and let $\mathcal{F}_\delta$ be an $L^2(\mathcal{D})$ $(B_0\delta)$-cover of $\mathcal{F}$ satisfying $|\mathcal{F}_\delta|=N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))$. Then it follows from Cauchy-Schwarz inequality that for each $f\in\mathcal{F}$, there exists $f_\delta\in\mathcal{F}_\delta$ such that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(X_i)-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if_\delta(X_i)\\ &\le\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i)-f_\delta(X_i))^2\right)^{1/2}\le B_0\delta.\end{aligned}$$

Combining (42) with the above inequality yields

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D}})-\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D}})\Big]\\ &\le4B_0\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}_\delta}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(X_i)\right]+4B_0^2\delta\\ &\le4B_0^2\left(\frac{2\log|\mathcal{F}_\delta|}{n}\right)^{\frac12}+4B_0^2\delta\\ &=4B_0^2\left(\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n}\right)^{\frac12}+4B_0^2\delta,\end{aligned}\tag{43}$$

where the last inequality holds from Massart's lemma [97, Theroem 3.7]. This completes the proof of (37).

Step (II). According to [98, Lemma 4], the Gaussian complexity can be bounded by the Rademacher complexity, that is,

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}^{\lambda}_{\mathcal{D}}(X_i)\right]\\ &\le\mathbb{E}_{\mathcal{D}}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\xi_if(X_i)\right]\\ &\le\sigma(\log n)\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(X_i)\right],\end{aligned}\tag{44}$$

where $\varepsilon=\{\varepsilon_i\}_{i=1}^n$ is a set of Rademacher variables and independent of $\mathcal{D}$. By the same argument as (42) and (43), we have

$$\mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_if(X_i)\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right] \\
&\le B_0\left(\frac{2\log N(B_0\delta,\mathcal{F},L^2(\mathcal{D}))}{n}\right)^{\frac{1}{2}} + \delta. \qquad (45)
\end{aligned}
$$

Combining (44) and (45) completes the proof of (38).

Step (III). For each element $f \in \operatorname{conv}(\mathcal{F})$, by the convexity of $\operatorname{conv}(\mathcal{F})$ we have $\widehat{f}_{\mathcal{D}}^{\lambda} + t(f - \widehat{f}_{\mathcal{D}}^{\lambda}) \in \operatorname{conv}(\mathcal{F})$ for each $t \in [0,1]$. Now the optimality of $\widehat{f}_{\mathcal{D}}^{\lambda}$ yields that for each $t \in [0,1]$

$$
\widehat{L}_{\mathcal{D}}^{\lambda}(\widehat{f}_{\mathcal{D}}^{\lambda}) - \widehat{L}_{\mathcal{D}}^{\lambda}(\widehat{f}_{\mathcal{D}}^{\lambda} + t(f - \widehat{f}_{\mathcal{D}}^{\lambda})) \le 0,
$$

which implies

$$
\begin{aligned}
&\lim_{t\to 0+}\frac{1}{t}\Big(\widehat{L}_{\mathcal{D}}^{\lambda}(\widehat{f}_{\mathcal{D}}^{\lambda}) - \widehat{L}_{\mathcal{D}}^{\lambda}(\widehat{f}_{\mathcal{D}}^{\lambda} + t(f - \widehat{f}_{\mathcal{D}}^{\lambda}))\Big) \\
&= \lambda(\nabla\widehat{f}_{\mathcal{D}}^{\lambda}, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f))_{L^2(\nu_X)} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_{\mathcal{D}}^{\lambda}(X_i) - Y_i)(\widehat{f}_{\mathcal{D}}^{\lambda}(X_i) - f(X_i)) \le 0.
\end{aligned}
$$

Therefore, it follows from (1) that for each $f \in \operatorname{conv}(\mathcal{F})$,

$$
\begin{aligned}
&\lambda(\nabla\widehat{f}_{\mathcal{D}}^{\lambda}, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f))_{L^2(\nu_X)} + (\widehat{f}_{\mathcal{D}}^{\lambda} - f_0, \widehat{f}_{\mathcal{D}}^{\lambda} - f)_{L^2(\mathcal{D})} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}_{\mathcal{D}}^{\lambda}(X_i) - f(X_i)). \qquad (46)
\end{aligned}
$$

For the first term in the left-hand side of (46), it follows from the linearity of inner-product that

$$
\begin{aligned}
&\lambda(\nabla\widehat{f}_{\mathcal{D}}^{\lambda}, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f))_{L^2(\nu_X)} \\
&= \lambda(\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0) + \nabla f_0, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0) - \nabla(f - f_0))_{L^2(\nu_X)} \\
&= \lambda\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2 + \lambda(\nabla f_0, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f))_{L^2(\nu_X)} \\
&\quad - \lambda(\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0), \nabla(f - f_0))_{L^2(\nu_X)}. \qquad (47)
\end{aligned}
$$

Then using Lemma 8 and Assumption 4, one obtains easily

$$
\begin{aligned}
&-\lambda(\nabla f_0, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f))_{L^2(\nu_X)} \\
&= \lambda(\Delta f_0, \widehat{f}_{\mathcal{D}}^{\lambda} - f)_{L^2(\nu_X)} \\
&\quad + \lambda(\nabla f_0 \cdot \nabla(\log q), \widehat{f}_{\mathcal{D}}^{\lambda} - f)_{L^2(\nu_X)} \\
&\le \lambda\kappa^{1/2}\Big\{\|\Delta f_0\|_{L^2(\nu_X)} + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}\Big\} \\
&\quad \times \Big\{R^{1/2}(\widehat{f}_{\mathcal{D}}^{\lambda}) + \|f - f_0\|_{L^2(\mu_X)}\Big\} \\
&\le 9\lambda^2\kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\} \\
&\quad + \frac{1}{16}R(\widehat{f}_{\mathcal{D}}^{\lambda}) + \frac{1}{2}\|f - f_0\|_{L^2(\mu_X)}^2, \qquad (48)
\end{aligned}
$$

where the first inequality holds from Cauchy-Schwarz inequality and the triangular inequality, and the last inequality is due to $ab \le \epsilon a^2 + b^2/(4\epsilon)$ for $a, b, \epsilon > 0$. Similarly, we also find that

$$
\begin{aligned}
&\lambda(\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0), \nabla(f - f_0))_{L^2(\nu_X)} \\
&\le \frac{\lambda}{2}\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2 + \frac{\lambda}{2}\|\nabla(f - f_0)\|_{L^2(\nu_X)}^2. \qquad (49)
\end{aligned}
$$

Using (47), (48) and (49) yields

$$
\begin{aligned}
&\frac{\lambda}{2}\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2 \\
&\le \lambda(\nabla\widehat{f}_{\mathcal{D}}^{\lambda}, \nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f))_{L^2(\nu_X)} + \frac{1}{16}R(\widehat{f}_{\mathcal{D}}^{\lambda}) \\
&\quad + \left\{\frac{1}{2}\|f - f_0\|_{L^2(\mu_X)}^2 + \frac{\lambda}{2}\|\nabla(f - f_0)\|_{L^2(\nu_X)}^2\right\} \\
&\quad + 9\lambda^2\kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\}. \qquad (50)
\end{aligned}
$$

We next turn to consider the second term in the left-hand side of (46). By Cauchy-Schwarz inequality and AM-GM inequality we have

$$
\begin{aligned}
&(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0, \widehat{f}_{\mathcal{D}}^{\lambda} - f)_{L^2(\mathcal{D})} \\
&= \widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda}) - (\widehat{f}_{\mathcal{D}}^{\lambda} - f_0, f - f_0)_{L^2(\mathcal{D})} \\
&\ge \frac{1}{2}\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda}) - \frac{1}{2}\widehat{R}_{\mathcal{D}}(f). \qquad (51)
\end{aligned}
$$

Combining (46), (50) and (51) and taking expectation with respect to $\mathcal{D} \sim \mu^n$ implies the following inequality for each $f \in \operatorname{conv}(\mathcal{F})$

$$
\begin{aligned}
&\frac{\lambda}{2}\mathbb{E}_{\mathcal{D}}\Big[\|\nabla(\widehat{f}_{\mathcal{D}}^{\lambda} - f_0)\|_{L^2(\nu_X)}^2\Big] + \frac{1}{2}\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] \\
&\le \frac{1}{16}\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] + \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_{\mathcal{D}}^{\lambda}(X_i)\right] \\
&\quad + \left\{\|f - f_0\|_{L^2(\mu_X)}^2 + \frac{\lambda}{2}\|\nabla(f - f_0)\|_{L^2(\nu_X)}^2\right\} \\
&\quad + 9\lambda^2\kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\},
\end{aligned}
$$

where we used the fact that $\mathbb{E}[\widehat{R}_{\mathcal{D}}(f)] = R(f)$ and $\mathbb{E}\left[\sum_{i=1}^{n}\xi_i f(X_i)\right] = 0$ for each fixed function $f \in L^{\infty}(\Omega)$. Since that $\mathcal{F} \subseteq \operatorname{conv}(\mathcal{F})$, it is apparent that this inequality also holds for each element in $\mathcal{F}$. Taking infimum with respect to $f \in \mathcal{F}$ obtains the inequality (39).

Step (IV). Using (38) and (39), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] \\
&\le \frac{1}{4}\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] + c\Big\{\beta\lambda^2 + \varepsilon_{\mathrm{app}}(\mathcal{F},\lambda) + \varepsilon_{\mathrm{gen}}(\mathcal{F},n)\Big\},
\end{aligned}
$$

where c is an absolute positive constant and β is a positive constant defined as

$$
\beta = \kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0 \cdot \nabla(\log q)\|_{L^2(\nu_X)}^2\Big\}.
$$

Consequently, by the estimate in (37), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] \le \mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] + c'\varepsilon_{\mathrm{gen}}(\mathcal{F},n) \\
&\le \frac{1}{4}\mathbb{E}_{\mathcal{D}}\Big[R(\widehat{f}_{\mathcal{D}}^{\lambda})\Big] + c\Big\{\beta\lambda^2 + \varepsilon_{\mathrm{app}}(\mathcal{F},\lambda)\Big\} \\
&\quad + (c + c')\varepsilon_{\mathrm{gen}}(\mathcal{F},n).
\end{aligned}
$$

This completes the proof of (40). Finally, combining (39) and (40) achieves (41). ■

Proof of Lemma 6 A direct conclusion of Lemma 15. ■

Proof of Theorem 2 According to Lemma 5.3, we set the hypothesis class $\mathcal{F}$ as ReQU neural networks $\mathcal{F} = \mathcal{N}(L,S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then there exists $f \in \mathcal{F}$ such that

$$
\|f - \phi\|_{L^2(\mu_X)} \le CN^{-s},
$$

$$\|\nabla(f-\phi)\|_{L^2(\nu_X)} \le CN^{-(s-1)}.$$

By using Lemma 12 and set $\delta = 1/n$, we find

$$\begin{aligned}&\log N(B_0 n^{-1}, \mathcal{F}, L^2(\mathcal{D}))\\ &\quad\lesssim LS\log(S)(\log n) \lesssim N^d \log^2 N \log n. \end{aligned}\tag{52}$$

Substituting these estimates into Lemma 5 yields

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[\|\widehat{f}^{\lambda}_{\mathcal{D}} - f_0\|^2_{L^2(\mu_X)}\Big]\\ &\quad\lesssim \beta\lambda^2 + CN^{-2s} + C\lambda N^{-2(s-1)}\\ &\qquad + C\log n\left(\frac{N^d\log^2 N\log n}{n}\right)^{\frac{1}{2}}.\end{aligned}$$

Setting $N = \mathcal{O}\left(n^{\frac{1}{d+4s}}\right)$, and letting the regularization parameter be $\lambda = \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right)$ deduce the desired result.

B. Semi-Supervised Deep Sobolev Regressor

Proof of Lemma 7 Before proceeding, we first define the empirical inner-product and norm based on the sample $\mathcal{S} = \{Z_i\}_{i=1}^m$ as

$$(u,v)_{L^2(\mathcal{S})} = \frac{1}{m}\sum_{i=1}^m u(Z_i)v(Z_i),\ \|u\|^2_{L^2(\mathcal{S})} = \frac{1}{m}\sum_{i=1}^m u^2(Z_i),$$

for each $u, v \in L^\infty(\nu_X)$. The proof is divided into four parts which are denoted by (I) to (IV):

(I) By a same argument as (I) in the proof of Lemma 5, we deduces

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|^2_{L^2(\nu_X)} - \|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|^2_{L^2(\mathcal{S})}\Big]\\ &\quad\le \varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m).\end{aligned}\tag{53}$$

Here the generalization error $\varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m)$ associated to the regularization term are defined as

$$\begin{aligned}&\varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m)\\ &\quad= B_1^2\inf_{\delta>0}\left\{\max_{1\le k\le d}\frac{N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m} + \delta\right\}.\end{aligned}$$

(II) By the technique of symmetrization and Green's formula, it holds that

$$\begin{aligned}&-\lambda\mathbb{E}_{\mathcal{S}}\Big[(\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\Big]\\ &\quad\le \frac{1}{16}\mathbb{E}_{\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] + c\Big\{\tilde{\beta}\lambda^2 + \varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m)\Big\},\end{aligned}\tag{54}$$

where c is an absolute positive constant. Here the constant $\tilde{\beta}$ is defined as

$$\tilde{\beta} = \kappa\Big\{\|\Delta f_0\|^2_{L^2(\nu_X)} + \|\nabla f_0\cdot\nabla(\log q)\|^2_{L^2(\nu_X)} + B_1^2\Big\}.$$

(III) With the aid of the variational inequality and (54), we have

$$\begin{aligned}&\lambda\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|^2_{L^2(\mathcal{S})}\Big] + \mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big]\\ &\quad\le \frac{1}{8}\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] + 2\mathbb{E}_{\mathcal{D},\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^n \xi_i\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}(X_i)\right]\\ &\qquad + c\Big\{\beta\lambda^2 + \varepsilon_{\mathrm{app}}(\mathcal{F},\lambda) + \varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m)\Big\},\end{aligned}\tag{55}$$

where c is an absolute positive constant

(IV) Applying (37), (38), (53) and (55), we conclude the final results.

Step (I). By a same argument as *Step (I)* in the proof of Lemma 5, we deduce the following inequality

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[\|D_k(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|^2_{L^2(\nu_X)} - \|\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0\|^2_{L^2(\mathcal{S})}\Big]\\ &\quad\le 4B_{1,k}^2\inf_{\delta>0}\left\{\frac{N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m} + \delta\right\},\end{aligned}$$

for each $1\le k\le d$. Summing over these equalities obtains (53) immediately.

Step (II). Given a ghost sample $\mathcal{S}' = \{Z_i'\}_{i=1}^m$, where $\{Z_i'\}_{i=1}^m$ are independently and identically distributed random variables from ν_X. Further, the ghost sample $\mathcal{S}'$ is independent of $\mathcal{S} = \{Z_i\}_{i=1}^m$. Let $\varepsilon = \{\varepsilon_i\}_{i=1}^m$ be a set of Rademacher variables and independent of $\mathcal{S}$ and $\mathcal{S}'$. Then by the technique of symmetrization, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[(D_kf_0, D_k\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})_{L^2(\nu_X)} - (D_kf_0, D_k\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})_{L^2(\mathcal{S})}\Big]\\ &\le \mathbb{E}_{\mathcal{S}}\Big[\sup_{f\in\mathrm{conv}(\mathcal{F})}(D_kf_0, D_kf)_{L^2(\nu_X)} - (D_kf_0, D_kf)_{L^2(\mathcal{S})}\Big]\\ &= \mathbb{E}_{\mathcal{S}}\Big[\sup_{f\in\mathrm{conv}(\mathcal{F})}\mathbb{E}_{\mathcal{S}'}\left[\frac{1}{m}\sum_{i=1}^m D_kf_0(Z_i')D_kf(Z_i')\right]\\ &\quad - \frac{1}{m}\sum_{i=1}^m D_kf_0(Z_i)D_kf(Z_i)\Big]\\ &\le \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}'}\Big[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{m}\sum_{i=1}^m D_kf_0(Z_i')D_kf(Z_i')\\ &\quad - \frac{1}{m}\sum_{i=1}^m D_kf_0(Z_i)D_kf(Z_i)\Big]\\ &= \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\varepsilon}\Big[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{m}\sum_{i=1}^m\varepsilon_i\Big(D_kf_0(Z_i')D_kf(Z_i')\\ &\quad - D_kf_0(Z_i)D_kf(Z_i)\Big)\Big]\\ &= \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathrm{conv}(\mathcal{F})}\frac{1}{m}\sum_{i=1}^m\varepsilon_iD_kf_0(Z_i)D_kf(Z_i)\right]\\ &= \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\varepsilon_iD_kf_0(Z_i)D_kf(Z_i)\right],\end{aligned}\tag{56}$$

where the second inequality follows from the Jensen's inequality, and the last equality invokes the fact that the Rademacher complexity of the convex hull is equal to that of the original set. According to Ledoux-Talagrand contraction inequality [97, Lemma 5.7], we have

$$\begin{aligned}&\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\varepsilon_iD_kf_0(Z_i)D_kf(Z_i)\right]\\ &\quad\le B_{1,k}\mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\varepsilon_iD_kf(Z_i)\right].\end{aligned}\tag{57}$$

Let $\delta > 0$ and let $(D_k\mathcal{F})_\delta$ be an $L^2(\mathcal{S})$ $(B_{1,k}\delta)$-cover of $D_k\mathcal{F}$. Suppose $|(D_k\mathcal{F})_\delta| = N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))$. Then it follows from Cauchy-Schwarz inequality that for each $D_kf \in D_k\mathcal{F}$, there exists $(D_kf)_\delta \in (D_k\mathcal{F})_\delta$ such that

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i D_kf(Z_i) - \frac{1}{m}\sum_{i=1}^{m}\varepsilon_i (D_kf)_\delta(Z_i)\\ &\leq \left(\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i^2\right)^{1/2}\left(\frac{1}{m}\sum_{i=1}^{m}(D_kf(Z_i)-(D_kf)_\delta(Z_i))^2\right)^{1/2}\\ &\leq B_{1,k}\delta,\end{aligned}$$

which implies

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\varepsilon}\left[\sup_{D_kf\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i D_kf(Z_i)\right]\\ &\leq \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\varepsilon}\left[\sup_{D_kf\in(D_k\mathcal{F})_\delta}\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i D_kf(Z_i)\right] + B_{1,k}\delta\\ &\leq B_{1,k}\left(\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m}\right)^{1/2} + B_{1,k}\delta, \qquad (58)\end{aligned}$$

where the last inequality holds from Massart's lemma [97, Theroem 3.7]. Combining (56), (57) and (58) deduces

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[(D_kf_0, D_k\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})_{L^2(\nu_X)} - (D_kf_0, D_k\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})_{L^2(\mathcal{S})}\Big]\\ &\leq B_{1,k}^2\left(\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m}\right)^{1/2} + B_{1,k}^2\delta.\end{aligned}$$

Summing over this equation for $1 \leq k \leq d$ yields

$$\begin{aligned}&\mathbb{E}_{\mathcal{S}}\Big[(\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0))_{L^2(\nu_X)}\\ &\quad - (\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0))_{L^2(\mathcal{S})}\Big]\\ &\leq \sum_{k=1}^{d}B_{1,k}^2\inf_{\delta>0}\left\{\left(\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m}\right)^{1/2} + \delta\right\}.\end{aligned}$$

Combining this with Lemma 8 and Assumption 4, we find that for each $\delta > 0$,

$$\begin{aligned}&-\lambda\mathbb{E}_{\mathcal{S}}\Big[(\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\Big]\\ &\leq \lambda\mathbb{E}_{\mathcal{S}}\Big[-(\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\nu_X)}\Big]\\ &\quad + \lambda\sum_{k=1}^{d}B_{1,k}^2\left\{\left(\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m}\right)^{1/2} + \delta\right\}\\ &= \lambda\mathbb{E}_{\mathcal{S}}\Big[(\Delta f_0, \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f)_{L^2(\nu_X)}\\ &\quad + (\nabla f_0\cdot\nabla(\log q), \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f)_{L^2(\nu_X)}\Big]\\ &\quad + \lambda\sum_{k=1}^{d}B_{1,k}^2\left\{\left(\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m}\right)^{1/2} + \delta\right\}\\ &\leq \lambda\kappa^{1/2}\mathbb{E}_{\mathcal{S}}\Big[\Big\{\|\Delta f_0 + \nabla f_0\cdot\nabla(\log q)\|_{L^2(\nu_X)}\Big\}\\ &\quad \times\Big\{R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})^{1/2} + \|f - f_0\|_{L^2(\mu_X)}\Big\}\Big]\\ &\quad + \lambda\sum_{k=1}^{d}B_{1,k}^2\left\{\left(\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m}\right)^{1/2} + \delta\right\}\\ &\leq 9\lambda^2\kappa\Big\{\|\Delta f_0\|_{L^2(\nu_X)}^2 + \|\nabla f_0\cdot\nabla(\log q)\|_{L^2(\nu_X)}^2\Big\}\\ &\quad + \frac{1}{16}\mathbb{E}_{\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] + \frac{1}{2}\|f - f_0\|_{L^2(\mu_X)}^2 + \lambda^2\left(\sum_{k=1}^{d}B_{1,k}^2\right)\\ &\quad + \frac{1}{4}\sum_{k=1}^{d}B_{1,k}^2\left\{\max_{1\leq k\leq d}\frac{2\log N(B_{1,k}\delta, D_k\mathcal{F}, L^2(\mathcal{S}))}{m} + \delta^2\right\},\end{aligned}$$

where the second inequality holds from Cauchy-Schwarz inequality and Assumption 1, and the last inequality is due to the inequality $ab \leq \epsilon a^2 + b^2/(4\epsilon)$ for $a, b, \epsilon > 0$. This completes the proof of (54).

Step (III). For each element $f \in \mathrm{conv}(\mathcal{F})$, by the convexity of $\mathrm{conv}(\mathcal{F})$ we have $\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} + t(f - \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}) \in \mathrm{conv}(\mathcal{F})$ for each $t \in [0,1]$. Now the optimality of $\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}$ yields that for each $t \in [0,1]$

$$\widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(\widehat{f}^{\lambda}_{\mathcal{D}}) - \widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} + t(f - \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})) \leq 0,$$

which implies

$$\begin{aligned}&\lim_{t\to0^+}\frac{1}{t}\Big(\widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}) - \widehat{L}^{\lambda}_{\mathcal{D},\mathcal{S}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} + t(f - \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}))\Big)\\ &= \lambda(\nabla\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}(X_i) - Y_i)(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}(X_i) - f(X_i)) \leq 0.\end{aligned}$$

Therefore, it follows from (1) that for each $f \in \mathrm{conv}(\mathcal{F})$,

$$\begin{aligned}&\lambda(\nabla\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\\ &+(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0, \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f)_{L^2(\mathcal{D})}\\ \leq\ &\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}(X_i) - f(X_i)). \qquad (59)\end{aligned}$$

For the first term in the left-hand side of (59), we have

$$\begin{aligned}&\lambda(\nabla\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\\ &= \lambda(\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0) + \nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0) - \nabla(f - f_0))_{L^2(\mathcal{S})}\\ &= \lambda\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|_{L^2(\mathcal{S})}^2 + \lambda(\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\\ &\quad - \lambda(\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0), \nabla(f - f_0))_{L^2(\mathcal{S})}, \qquad (60)\end{aligned}$$

According to Cauchy-Schwarz inequality and AM-GM inequality, one obtains easily

$$\begin{aligned}&\lambda(\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0), \nabla(f - f_0))_{L^2(\mathcal{S})}\\ &\leq \frac{\lambda}{2}\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|_{L^2(\mathcal{S})}^2 + \frac{\lambda}{2}\|\nabla(f - f_0)\|_{L^2(\mathcal{S})}^2. \qquad (61)\end{aligned}$$

Using (60) and (61), and taking expectation with respect to $\mathcal{S} \sim \nu_X^m$ yield

$$\begin{aligned}&\frac{\lambda}{2}\mathbb{E}_{\mathcal{S}}\Big[\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|_{L^2(\mathcal{S})}^2\Big]\\ &\leq \lambda\mathbb{E}_{\mathcal{S}}\Big[(\nabla\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\Big]\\ &\quad + \frac{\lambda}{2}\|\nabla(f - f_0)\|_{L^2(\nu_X)}^2\\ &\quad - \lambda\mathbb{E}_{\mathcal{S}}\Big[(\nabla f_0, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\Big].\end{aligned}$$

Combining this estimate with (54) implies

$$\frac{\lambda}{2}\mathbb{E}_{\mathcal{S}}\Big[\|\nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0)\|_{L^2(\mathcal{S})}^2\Big]$$

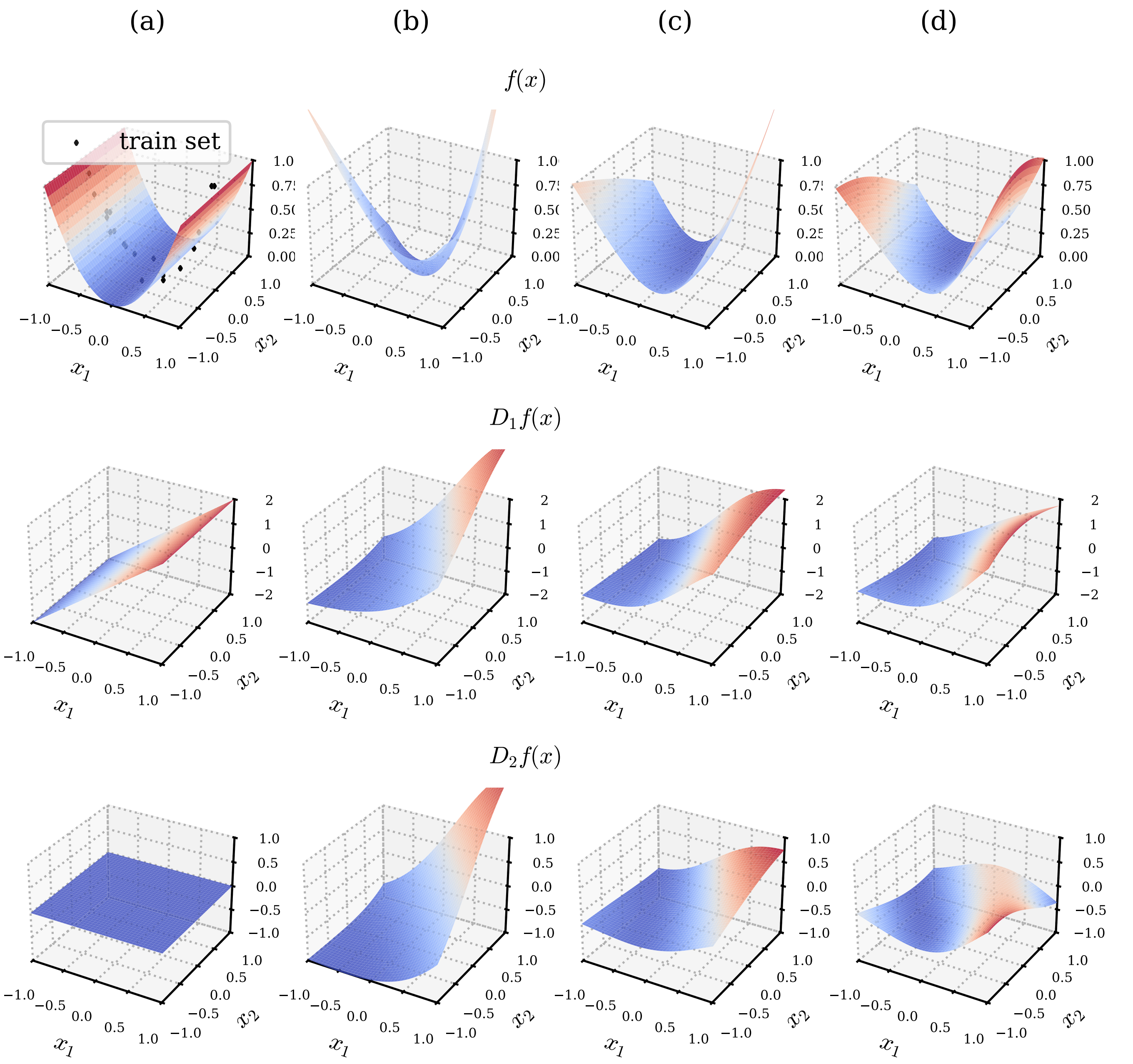

Fig. 3. Effect of the regularization technique on function fitting for a toy problem $f_0(x) = x_1^2$. (a) Landscape of the primitive function and its partial derivatives. The train samples are plotted in black dots. (b) least-squares fitting estimation. (c) DORE estimation. (d) SDORE estimation.

$$
\begin{aligned}
&\leq \lambda \mathbb{E}_{\mathcal{S}}\Big[(\nabla \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}, \nabla(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f))_{L^2(\mathcal{S})}\Big] \\
&\quad + \frac{\lambda}{2}\|\nabla(f - f_0)\|^2_{L^2(\nu_X)} + \frac{1}{16}\mathbb{E}_{\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] \\
&\quad + c\Big\{\tilde{\beta}\lambda^2 + \varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m)\Big\}. \qquad (62)
\end{aligned}
$$

We next turn to consider the second term in the left-hand side of (59). By Cauchy-Schwarz inequality and AM-GM inequality we have

$$
\begin{aligned}
&(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0, \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f)_{L^2(\mathcal{D})} \\
&= \widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}) - (\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0, f - f_0)_{L^2(\mathcal{D})} \\
&\geq \frac{1}{2}\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}}) - \frac{1}{2}\widehat{R}_{\mathcal{D}}(f),
\end{aligned}
$$

which implies by taking expectation with respect to $\mathcal{D} \sim \mu^n$ that for each $f \in \mathcal{F}$,

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\mathcal{D}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] \\
&\leq \frac{1}{2}R(f) + \mathbb{E}_{\mathcal{D}}\Big[(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0, \widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f)_{L^2(\mathcal{D})}\Big]. \qquad (63)
\end{aligned}
$$

Combining (59), (62) and (63) yields (55).

Step (IV). Using (38) and (55), we have

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] \leq \frac{1}{4}\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] \\
&\quad + c\Big\{\tilde{\beta}\lambda^2 + \varepsilon_{\mathrm{app}}(\mathcal{F}, \lambda) + \varepsilon_{\mathrm{gen}}(\mathcal{F}, n) + \varepsilon^{\mathrm{reg}}_{\mathrm{gen}}(\nabla\mathcal{F}, m)\Big\}.
\end{aligned}
$$

Then according to the above inequality and (37), it follows that

$$\begin{aligned}\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] &\le \mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[\widehat{R}_{\mathcal{D}}(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] + c'\varepsilon_{\rm gen}(\mathcal{F},n)\\ &\le \frac{1}{2}\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[R(\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}})\Big] + (2c+c')\varepsilon_{\rm gen}(\mathcal{F},n)\\ &\quad + 2c\Big\{\tilde{\beta}\lambda^2 + \varepsilon_{\rm app}(\mathcal{F},\lambda) + \varepsilon_{\rm gen}^{\rm reg}(\nabla\mathcal{F},m)\Big\},\end{aligned}$$

which implies

$$\begin{aligned}&\mathbb{E}_{\mathcal{D},\mathcal{S}}\Big[\|\widehat{f}^{\lambda}_{\mathcal{D},\mathcal{S}} - f_0\|^2_{L^2(\mu_X)}\Big]\\ &\lesssim \tilde{\beta}\lambda^2 + \varepsilon_{\rm app}(\mathcal{F},\lambda) + \varepsilon_{\rm gen}(\mathcal{F},n) + \varepsilon_{\rm gen}^{\rm reg}(\nabla\mathcal{F},m). \qquad (64)\end{aligned}$$

Finally, combining (53), (55) and (64) completes the proof. ■

Proof of Theorem 3 According to Lemma 6, we set the hypothesis class $\mathcal{F}$ as ReQU neural networks $\mathcal{F} = \mathcal{N}(L,S)$ with $L = \mathcal{O}(\log N)$ and $S = \mathcal{O}(N^d)$. Then there exists $f \in \mathcal{F}$ such that

$$\|f-\phi\|_{L^2(\mu_X)} \le CN^{-s}, \quad \|\nabla(f-\phi)\|_{L^2(\nu_X)} \le CN^{-(s-1)}.$$

By using Lemma 14 and set $\delta = 1/n$, we find

$$\begin{aligned}&\log N(B_{1,k}n^{-1}, D_k\mathcal{F}, L^2(\mathcal{D}))\\ &\lesssim L^2 S\log S\log n \lesssim N^d\log^3 N\log n.\end{aligned}$$

Substituting these estimates and (52) into Lemma 5 yields

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}\Big[\|\widehat{f}^{\lambda}_{\mathcal{D}} - f_0\|^2_{L^2(\mu_X)}\Big]\\ &\lesssim \tilde{\beta}\lambda^2 + CN^{-2s} + C\lambda N^{-2(s-1)}\\ &\quad + C\log n\left(\frac{N^d\log N\log n}{n}\right)^{\frac{1}{2}} + C\log n\frac{N^d\log^3 N}{m}.\end{aligned}$$

Setting $N = \mathcal{O}\left(n^{\frac{1}{d+4s}}\right)$, and letting the regularization parameter be $\lambda = \mathcal{O}\left(n^{-\frac{s}{d+4s}}\log^2 n\right)$ deduce the desired result. ■

Appendix E
Proofs in Results in Section VI

Proof of Corollary 1 A direct conclusion of Theorem 2. ■

Proof of Corollary 2 By Markov's inequality [97, Theorem C.11], the following inequality holds for each $\epsilon > 0$

$$\begin{aligned}&\lim_{n\to\infty}\Pr\Big\{\|D_k\widehat{f}^{\lambda}_{\mathcal{D}} - D_kf_0\|_{L^2(\nu_X)} > \epsilon\Big\}\\ &\le \lim_{n\to\infty}\frac{\mathbb{E}_{\mathcal{D}}\Big[\|D_k\widehat{f}^{\lambda}_{\mathcal{D}} - D_kf_0\|_{L^2(\nu_X)}\Big]}{\epsilon} = 0, \qquad (65)\end{aligned}$$

where the equality follows from Corollary 1. For each irrelevant variable $k \notin \mathcal{I}(f_0)$, one has $\|D_kf_0\|_{L^2(\nu_X)} = 0$. Then (65) deduces that for each $\epsilon > 0$

$$\begin{aligned}&\lim_{n\to\infty}\Pr\Big\{\|D_k\widehat{f}^{\lambda}_{\mathcal{D}}\|_{L^2(\nu_X)} > \epsilon\Big\}\\ &\le \lim_{n\to\infty}\Pr\Big\{\|D_k\widehat{f}^{\lambda}_{\mathcal{D}} - D_kf_0\|_{L^2(\nu_X)} > \epsilon\Big\} = 0,\end{aligned}$$

which implies $\|D_k\widehat{f}^{\lambda}_{\mathcal{D}}\|_{L^2(\nu_X)}$ goes to 0 in probability, and thus

$$\lim_{n\to\infty}\Pr\Big\{\mathcal{I}(\widehat{f}^{\lambda}_{\mathcal{D}}) \subseteq \mathcal{I}(f_0)\Big\} = 1. \qquad (66)$$

On the other hand, for each relevant variable $k \in \mathcal{I}(f_0)$, it follows from (65) that

$$\begin{aligned}&\lim_{n\to\infty}\Pr\Big\{\|D_k\widehat{f}^{\lambda}_{\mathcal{D}}\|_{L^2(\nu_X)} > \epsilon + \|D_kf_0\|_{L^2(\nu_X)}\Big\}\\ &\le \lim_{n\to\infty}\Pr\Big\{\|D_k\widehat{f}^{\lambda}_{\mathcal{D}} - D_kf_0\|_{L^2(\nu_X)} > \epsilon\Big\} = 0,\end{aligned}$$

where we used the triangular inequality. Since $\|D_kf_0\|_{L^2(\nu_X)} > 0$, we find that for each $\epsilon > 0$

$$\lim_{n\to\infty}\Pr\Big\{\|D_k\widehat{f}^{\lambda}_{\mathcal{D}}\|_{L^2(\nu_X)} > \epsilon\Big\} = 1,$$

As a consequence,

$$\lim_{n\to\infty}\Pr\Big\{\mathcal{I}(f_0) \subseteq \mathcal{I}(\widehat{f}^{\lambda}_{\mathcal{D}})\Big\} = 1. \qquad (67)$$

Combining (66) and (67) completes the proof. ■

Appendix F
Additional Experiments Results

In this section, we present several supplementary numerical examples to complement the numerical studies in Section VI.

A. Additional Examples for Derivative Estimation

Example 3: We consider a toy problem in two-dimensions, where the support of the marginal distribution μ_X approximately coincides with the coordinate subspace $[0,1] \times \{0\}$. Precisely the first element of the covariate is uniformly sampled from $[-1,1]$, whereas the second one is drawn from a Gaussian distribution $N(0,0.05)$. The underlying regression function is $f_0(x) = x_1^2$, and labels are generated by $Y = f_0(X) + \xi$, where the noise term $\xi \sim N(0,0.1)$. The regularization parameter is set as $\lambda = 1.0 \times 10^{-4}$.

In all cases the accuracy on the $\mathrm{supp}(\mu_X)$ is high, see Figure 3 (top), but the least-squares regressor fails to extend the approximation and smoothness outside the support, as the least-squares loss is insensible to errors out of $\mathrm{supp}(\mu_X)$. While the landscape of DORE is smoother compared to the least-squares regressor, SDORE further extends the smoothness to $[-1,1]^2$ as it utilizes unlabeled samples from ν_X.

We examine the partial derivative estimation with respect to x_1 and x_2 on $[-1,1]^2$ and display the result in Figure 3. As expected, the least-squares one is unstable compared to the DORE and SDORE. Also, we can tell from the bottom right of Figure 3 that x_2 is the irrelevant variable.

B. Additional Examples for Variable Selection

Example 4: Consider the regression function $f_0(x) = 2x_1^2 + e^{x_2} + 2\sin(x_3) + 2\cos(x_4+1)$, with observations $Y = f_0(X) + \xi$, where $X \in \mathbb{R}^{10}$ and ξ is a white noise, sampled from a Gaussian distribution with the signal to noise ratio to be 25. The first four elements of X are drawn from the uniform distribution on $[0,1]$, and the rest noise variables are drawn from the uniform distribution on $[0,0.05]$. The regularization parameter λ is set as 1.0×10^{-4} for SDORE.

We repeat the process for least-squares regressor and SDORE, respectively, and evaluate the models on a test set with sample size 1000. We report the estimated mean square of partial derivative by both estimators with respect to x_i,

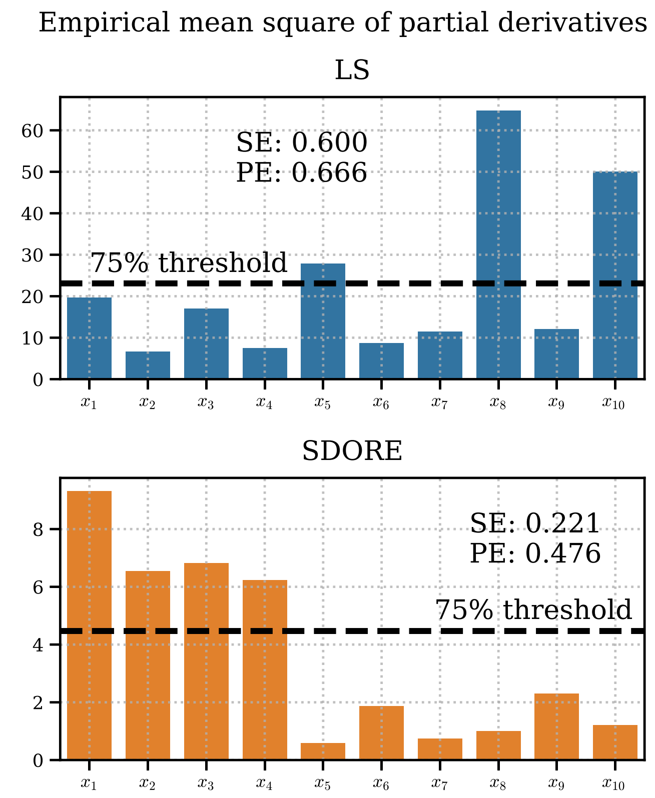

Fig. 4. (left) Empirical mean square of the partial derivatives estimated by least-squares regression (LS) and SDORE on a variable selection problem in $\mathbb{R}^{10}$ where f_0 is dependent on the x_1 to x_4. The dashed line is the 75 % quantile threshold for variable selection. (center) Mean variable selection error for the estimated derivative function on test set. (right) Root mean squared prediction error for the primitive function on test set.

the mean selection error (mean of false positive rate and false negative rate) the root mean squared prediction error on the primitive function in Figure 4. The results indicate least-squares regression fails to identify the correct dependent variables and has larger prediction error. In contrast, SDORE yields smaller prediction error, and points out that x_1 to x_4 are the relevant variables.

Acknowledgment

The authors would like to express their sincere gratitude to the Editor-in-Chief, Associate Editor, and all reviewers for their careful review of the manuscript. Their insightful comments and constructive suggestions have significantly improved the quality and clarity of this articles. The numerical calculations in this articles have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression* (Springer Series in Statistics (SSS)), vol. 1. New York, NY, USA: Springer, 2002.

[2] L. Wasserman, *All of Nonparametric Statistics* (Springer Texts in Statistics), 1st ed., New York, NY, USA: Springer, 2006.

[3] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics (SSS)). New York, NY, USA: Springer, 2009.

[4] H. Drucker and Y. L. Cun, "Double backpropagation increasing generalization performance," in *Proc. IJCNN Seattle Int. Joint Conf. Neural Netw.*, Jul. 1991, pp. 145–150.

[5] H. Drucker and Y. Le Cun, "Improving generalization performance using double backpropagation," *IEEE Trans. Neural Netw.*, vol. 3, no. 6, pp. 991–997, Nov. 1992.

[6] S. Rifai, P. Vincent, X. M'uller, X. Glorot, and Y. Bengio, "Contractive auto-encoders: Explicit invariance during feature extraction," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, WA, USA, Jun. 2011, pp. 833–840.

[7] V. Nagarajan and J. Z. Kolter, "Gradient descent gan optimization is locally stable," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 1–11.

[8] K. A. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Jan. 2017, pp. 1–11.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.

[10] Y. Gao, J. Huang, Y. Jiao, J. Liu, X. Lu, and Z. Yang, "Deep generative learning via Euler particle transport," in *Proc. 2nd Math. Sci. Mach. Learn. Conf.*, vol. 145, J. Bruna, J. Hesthaven, and L. Zdeborova, Eds., Aug. 2022, pp. 336–368.

[11] C. Lyu, K. Huang, and H.-N. Liang, "A unified gradient regularization family for adversarial examples," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2015, pp. 301–309.

[12] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., May 2017, pp. 2266–2276.

[13] A. G. Ororbia II, D. Kifer, and C. L. Giles, "Unifying adversarial training algorithms with data gradient regularization," *Neural Comput.*, vol. 29, no. 4, pp. 867–887, Apr. 2017.

[14] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, "Sensitivity and generalization in neural networks: An empirical study," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2018, pp. 1–21.

[15] D. Jakubovitz and R. Giryes, "Improving DNN robustness to adversarial attacks using Jacobian regularization," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, vol. 11216, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Springer, Sep. 2018, pp. 525–541.

[16] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients," in *Proc. 32th AAAI Conf. Artif. Intell.*, Palo Alto, CA, USA, Jan. 2017, pp. 1–13.

[17] R. W. Shephard, *Cost and Production Functions* (Lecture Notes in Economics and Mathematical Systems), vol. 194. Berlin, Germany: Springer, 1981.

[18] V. Rondonotti, J. S. Marron, and C. Park, "Sizer for time series: A new approach to the analysis of trends," *Electron. J. Statist.*, vol. 1, no. none, pp. 268–289, Jan. 2007.

[19] J. O. Ramsay and B. W. Silverman, *Applied Functional Data Analysis: Methods and Case Studies* (Springer Series in Statistics (SSS)). New York, NY, USA: Springer, 2002.

[20] S. Banerjee, A. E. Gelfand, and C. F. Sirmans, "Directional rates of change under spatial process models," *J. Amer. Stat. Assoc.*, vol. 98, no. 464, pp. 946–954, Dec. 2003.

[21] H.-G. Müller and F. Yao, "Additive modelling of functional gradients," *Biometrika*, vol. 97, no. 4, pp. 791–805, Dec. 2010.

[22] X. Dai, H.-G. Müller, and W. Tao, "Derivative principal components for representing the time dynamics of longitudinal and functional data," *Statistica Sinica*, vol. 28, no. 3, pp. 1583–1609, 2018.

[23] L. Rosasco, M. Santoro, S. Mosci, A. Verri, and S. Villa, "A regularization approach to nonlinear variable selection," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, vol. 9, Y. W. Teh and M. Titterington, Eds., May 2010, pp. 653–660.

[24] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa, "Is there sparsity beyond additive models?," *IFAC Proc. Volumes*, vol. 45, no. 16, pp. 971–976, Jul. 2012.

[25] L. Rosasco, S. Villa, S. Mosci, M. Santoro, and A. Verri, "Nonparametric sparsity and regularization," *J. Mach. Learn. Res.*, vol. 14, no. 52, pp. 1665–1714, 2013.

[26] Q. Hu, S. Shu, and J. Zou, "A new variational approach for inverse source problems," *Numer. Math., Theory, Methods Appl.*, vol. 12, no. 2, pp. 331–347, 2018.

[27] K. D. Brabanter, J. D. Brabanter, B. D. Moor, and I. Gijbels, "Derivative estimation with local polynomial fitting," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 281–301, Jan. 2013.

[28] N. E. Heckman and J. O. Ramsay, "Penalized regression with model-based penalties," *Can. J. Statist.*, vol. 28, no. 2, pp. 241–258, Jun. 2000.

[29] Z. Liu and M. Li, "On the estimation of derivatives using plug-in kernel ridge regression estimators," *J. Mach. Learn. Res.*, vol. 24, no. 266, pp. 1–37, Jan. 2023.

[30] H.-G. Müller, U. Stadtmuller, and T. Schmitt, "Bandwidth choice and confidence intervals for derivatives of noisy data," *Biometrika*, vol. 74, no. 4, p. 743, Dec. 1987.

[31] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ, USA: Princeton Univ. Press, 1961.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[33] R. Liu, K. Li, and M. Li, "Estimation and hypothesis testing of derivatives in smoothing spline ANOVA models," 2023, *arXiv:2308.13905*.

[34] B. Bauer and M. Kohler, "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Ann. Statist.*, vol. 47, no. 4, pp. 2261–2285, Aug. 2019.

[35] R. Nakada and M. Imaizumi, "Adaptive approximation and generalization of deep neural network with intrinsic dimensionality," *J. Mach. Learn. Res.*, vol. 21, no. 174, pp. 1–38, 2020.

[36] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, Aug. 2020.

[37] M. Kohler and S. Langer, "On the rate of convergence of fully connected deep neural network regression estimates," *Ann. Statist.*, vol. 49, no. 4, pp. 2231–2249, Aug. 2021.

[38] M. H. Farrell, T. Liang, and S. Misra, "Deep neural networks for estimation and inference," *Econometrica*, vol. 89, no. 1, pp. 181–213, 2021.

[39] M. Kohler, A. Krzyzak, and S. Langer, "Estimation of a function of low local dimensionality by deep neural networks," *IEEE Trans. Inf. Theory*, vol. 68, no. 6, pp. 4032–4042, Jun. 2022.

[40] Y. Jiao, G. Shen, Y. Lin, and J. Huang, "Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors," *Ann. Statist.*, vol. 51, no. 2, pp. 691–716, Apr. 2023.

[41] G. Wahba, *Spline Models for Observational Data* (CBMS-NSF Regional Conference Series in Applied Mathematics). Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1990.

[42] M. Kohler and A. Krzyzak, "Nonparametric regression estimation using penalized least squares," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3054–3058, Jul. 2001.

[43] M. Köhler, A. Krzyżak, and D. Schäfer, "Application of structural risk minimization to multivariate smoothing spline regression estimates," *Bernoulli*, vol. 8, no. 4, pp. 475–489, Aug. 2002.

[44] C. J. Stone, "Additive regression and other nonparametric models," *Ann. Statist.*, vol. 13, no. 2, pp. 689–705, Jun. 1985.

[45] R. A. DeVore and G. G. Lorentz, *Constructive Approximation* (Grundlehren der mathematischen Wissenschaften), vol. 303. Berlin, Germany: Springer, 1993.

[46] E. Weinan and B. Yu, "The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems," *Commun. Math. Statist.*, vol. 6, no. 1, pp. 1–12, 2018.

[47] J.-P. Florens, M. Ivaldi, and S. Larribeau, "Sobolev estimation of approximate regressions," *Econ. Theory*, vol. 12, no. 5, pp. 753–772, Dec. 1996.

[48] P. Hall and A. Yatchew, "Nonparametric estimation when data on derivatives are available," *Ann. Statist.*, vol. 35, no. 1, pp. 300–323, Feb. 2007.

[49] P. Hall and A. Yatchew, "Nonparametric least squares estimation in derivative families," *J. Econometrics*, vol. 157, no. 2, pp. 362–374, Aug. 2010.

[50] J. Newell and J. Einbeck, "A comparative study of nonparametric derivative estimators," in *Proc. 22nd Int. Workshop Stat. Model.*, Barcelona, Spain, Jul. 2007, pp. 453–456.

[51] Y. Liu and K. D. Brabanter, "Derivative estimation in random design," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Jan. 2018, pp. 1–10.

[52] Y. Liu and K. D. Brabanter, "Smoothed nonparametric derivative estimation using weighted difference quotients," *J. Mach. Learn. Res.*, vol. 21, no. 65, pp. 1–45, Jan. 2020.

[53] J. Epperson, "On the Runge example," *Amer. Math. Month.*, vol. 94, no. 4, pp. 329–341, 1987.

[54] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications* (Monographs on Statistics and Applied Probability), 1st ed., London, U.K.: Chapman & Hall, 1996.

[55] E. Masry, "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *J. Time Ser. Anal.*, vol. 17, no. 6, pp. 571–599, Nov. 1996.

[56] E. Masry, "Multivariate regression estimation local polynomial fitting for time series," *Stochastic Processes Appl.*, vol. 65, no. 1, pp. 81–101, Dec. 1996.

[57] A. Amiri and B. Thiam, "Regression estimation by local polynomial fitting for multivariate data streams," *Stat. Papers*, vol. 59, no. 2, pp. 813–843, Jun. 2018.

[58] P. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach* (Monographs on Statistics and Applied Probability), 1st ed., New York, NY, USA: Chapman & Hall, 1993.

[59] R. L. Eubank, *Nonparametric Regression Spline Smoothing*. Boca Raton, FL, USA: CRC Press, 1999.

[60] S. Zhou and D. A. Wolfe, "On derivative estimation in spline regression," *Statistica Sinica*, vol. 10, no. 1, pp. 93–108, Jan. 2000.

[61] T. Tao, *Analysis II* (Texts and Readings in Mathematics (TRIM)), vol. 38, 4th ed., Singapore: Springer, 2022.

[62] X. Zhu and A. B. Goldberg, *Introduction To Semi-supervised Learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning (SLAIML)). Cham, Switzerland: Springer, 2009.

[63] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, Feb. 2020.

[64] T. Zhang, "The value of unlabeled data for classification problems," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 1–23.

[65] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[66] L. Wasserman and J. Lafferty, "Statistical analysis of semi-supervised regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds., Dec. 2007, pp. 1–15.

[67] D. Azriel, L. D. Brown, M. Sklar, R. Berk, A. Buja, and L. Zhao, "Semi-supervised linear regression," *J. Amer. Statist. Assoc.*, vol. 117, no. 540, pp. 2238–2251, 2022.

[68] I. Livne, D. Azriel, and Y. Goldberg, "Improved estimators for semi-supervised high-dimensional regression model," *Electron. J. Statist.*, vol. 16, no. 2, pp. 5437–5487, Jan. 2022.

[69] S. Song, Y. Lin, and Y. Zhou, "A general M-estimation theory in semi-supervised framework," *J. Amer. Stat. Assoc.*, vol. 119, no. 546, pp. 1065–1075, Apr. 2024.

[70] S. Deng, Y. Ning, J. Zhao, and H. Zhang, "Optimal and safe estimation for high-dimensional semi-supervised learning," *J. Amer. Stat. Assoc.*, vol. 119, no. 548, pp. 2748–2759, Oct. 2024.

[71] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods* (Texts in Applied Mathematics (TAM)). New York, NY, USA: Springer, 2008.

[72] L. C. Evans, *Partial Differential Equations* (Graduate Studies in Mathematics), vol. 19. Providence, RI, USA: American Mathematical Society, 2010.

[73] D. Yarotsky, "Optimal approximation of continuous functions by very deep ReLU networks," in *Proc. Conf. Learn. Theory*, 2018, pp. 639–649.

[74] D. Yarotsky and A. Zhevnerchuk, "The phase diagram of approximation rates for deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Jan. 2020, pp. 13005–13015.

[75] Z. Shen, H. Yang, and S. Zhang, "Nonlinear approximation via compositions," *Neural Netw.*, vol. 119, pp. 74–84, Nov. 2019.

[76] Z. Shen, "Deep network approximation characterized by number of neurons," *Commun. Comput. Phys.*, vol. 28, no. 5, pp. 1768–1811, Jun. 2020.

[77] J. Lu, Z. Shen, H. Yang, and S. Zhang, "Deep network approximation for smooth functions," *SIAM J. Math. Anal.*, vol. 53, no. 5, pp. 5465–5506, Jan. 2021.

[78] P. Petersen and F. Voigtlaender, "Optimal approximation of piecewise smooth functions using deep ReLU neural networks," *Neural Netw.*, vol. 108, pp. 296–330, Dec. 2018.

[79] B. L. Bo Li, S. T. Shanshan Tang, and H. Y. Haijun Yu, "Better approximations of high dimensional smooth functions by deep neural networks with rectified power units," *Commun. Comput. Phys.*, vol. 27, no. 2, pp. 379–411, Jan. 2020.

[80] B. L. Bo Li, S. T. Shanshan Tang, and H. Y. Haijun Yu, "PowerNet: Efficient representations of polynomials and smooth functions by deep neural networks with rectified power units," *J. Math. Study*, vol. 53, no. 2, pp. 159–191, Jan. 2020.

[81] C. D. C. Duan, Y. J. Y. Jiao, Y. L. Y. Lai, D. L. D. Li, X. L. X. Lu, and J. Z. Y. J. Z. Yang, "Convergence rate analysis for deep Ritz method," *Commun. Comput. Phys.*, vol. 31, no. 4, pp. 1020–1048, Jan. 2022.

[82] G. Shen, Y. Jiao, Y. Lin, and J. Huang, "Differentiable neural networks with RePU activation: With applications to score estimation and isotonic regression," 2023, *arXiv:2305.00608*.

[83] J. Huang, Y. Jiao, Z. Li, S. Liu, Y. Wang, and Y. Yang, "An error analysis of generative adversarial networks for learning distributions," *J. Mach. Learn. Res.*, vol. 23, no. 116, pp. 1–43, Jan. 2022.

[84] Y. Jiao, Y. Wang, and Y. Yang, "Approximation bounds for norm constrained neural networks with applications to regression and GANs," *Appl. Comput. Harmon. Anal.*, vol. 65, pp. 249–278, Jul. 2023.

[85] C. J. Stone, "Optimal global rates of convergence for nonparametric regression," *Ann. Statist.*, vol. 10, no. 4, pp. 1040–1053, Dec. 1982.

[86] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, Oct. 1999.

[87] S. Hashem, "Optimal linear combinations of neural networks," *Neural Netw.*, vol. 10, no. 4, pp. 599–614, Jun. 1997.

[88] I. Gühring, G. Kutyniok, and P. Petersen, "Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms," *Anal. Appl.*, vol. 18, no. 5, pp. 803–859, 2020.

[89] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization With PDE Constraints* (Mathematical Modelling: Theory and Applications), 1st ed., Dordrecht, The Netherlands: Springer, 2009.

[90] F. Tröltzsch, *Optimal Control of Partial Differential Equations: Theory, Methods and Applications* (Graduate Studies in Mathematics), vol. 112. Providence, RI, USA: American Mathematical Society, 2010.

[91] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[92] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, vol. 9. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[93] P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," *J. Mach. Learn. Res.*, vol. 20, no. 63, pp. 1–17, 2019.

[94] T. Bagby, L. Bos, and N. Levenberg, "Multivariate simultaneous approximation," *Constructive Approximation*, vol. 18, no. 4, pp. 569–577, Dec. 2002.

[95] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2011.

[96] T. Liang, A. Rakhlin, and K. Sridharan, "Learning with square loss: Localization through offset Rademacher complexity," in *Proc. 28th Conf. Learn. Theory*, Paris, France, Jun. 2015, pp. 1260–1285.

[97] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed., Cambridge, MA, USA: MIT Press, 2018.

[98] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.

Zhao Ding received the B.Sc. degree in mathematics and applied mathematics from Wuhan University, Wuhan, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics.

His research interests include the theory of deep learning and generative learning based on diffusion models.

Chenguang Duan received the B.S. degree in information and computational science from Wuhan University, Wuhan, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics. He has authored research articles, including *Journal of Computational and Applied Mathematics* and *Communications in Computational Physics*. His research interests lie in deep learning theory, generative models, and scientific machine learning.

Yuling Jiao received the B.Sc. degree in applied mathematics from Shangqiu Normal University, Shangqiu, China, in 2008, and the Ph.D. degree in applied mathematics from Wuhan University, Wuhan, China, in 2014.

He is currently a Full Professor with the School of Artificial Intelligence, Wuhan University. His research works have been published in journals and conferences, including *SIAM Journal on Mathematical Analysis*, *SIAM Journal on Control and Optimization*, *SIAM Journal on Numerical Analysis*, *SIAM Journal on Scientific Computing*, *SIAM Journal on Mathematics of Data Science*, *Applied and Computational Harmonic Analysis*, IEEE TRANSACTIONS ON INFORMATION THEORY, *Annals of Statistics*, *Journal of the American Statistical Association*, *Statistical Science*, *Inverse Problems*, IEEE TRANSACTIONS ON SIGNAL PROCESSING, *Nature Communications*, ICML, and NeurIPS. His research interests include machine learning and scientific computing.

Jerry Zhijian Yang received the B.S. and M.S. degrees from Peking University, Beijing, China, in 1999 and 2001, respectively, and the Ph.D. degree in applied and computational mathematics from Princeton University, Princeton, NJ, USA, in 2006.

He completed his Post-Doctoral Research at California Institute of Technology, Pasadena, CA, USA, in 2008. He is currently a Full Professor with the School of Mathematics and Statistics and Wuhan Institute for Math & Al, Wuhan University, Wuhan, China. He has authored or co-authored research articles, including *Numerische Mathematik*, *Physical Review B*, *SIAM Multiscale Modeling and Simulation*, *SIAM Journal on Control and Optimization*, *Journal of Computational Physics*, *The Journal of Chemical Physics*, *International Journal for Numerical Methods in Engineering*, and *Communications in Computational Physics*. His research interests include multiscale modeling and simulation, machine learning, and scientific computing.