A Theoretical Analysis of Using Gradient Data for Sobolev Training in RKHS

Zain ul Abdeen, Ruoxi Jia, Vassilis Kekatos, Ming Jin

The Bradley Department of Electrical and Computer Engineering
Virginia Tech, Blacksburg, VA, USA

Abstract:
Recent works empirically demonstrated that incorporating target derivatives, in addition to the conventional usage of target values, during the training process improves the accuracy of the predictor and data efficiency. Despite the successful application of gradient data in the learning process, very little is understood theoretically about their performance guarantee. In this paper, our goal is to highlight (i) the limitations of gradient data on their performance guarantees, especially in low-data regimes, and (ii) the extent to which the gradients affect the learning rate. Our result implies that in a low-data regime, if the Lipschitz of the target function is below a threshold, gradient data for Sobolev training outperforms the classical training in terms of sample efficiency. For a target function with a large Lipschitz constant, there is a threshold for training data size beyond which the gradient data perform better than conventional training. The convergence behavior of gradient data for Sobolev training is studied, and the learning rate of order $\mathcal{O}(n^{-\frac{1}{2}+\epsilon})$ is derived. Experiments are conducted to determine the effect of gradient data in the learning process.

Keywords: Sobolev training, gradient data, learning rate, reproducing kernel Hilbert space.

1. INTRODUCTION

Gradient data have recently attracted growing attention in different domains Parag et al. (2022); Novara et al. (2022); Tsay (2021); Singh et al. (2020); Cocola and Hand (2020); Raissi et al. (2019). *Sobolev training* introduced by Czarnecki et al. (2017) adds an additional term to the loss definition that penalizes deviation of the estimated function gradient from the gradient of the target function. Figure 1 illustrates how Sobolev differs from classical training in the context of supervised learning. The main idea is that the distance between the estimated and target functions is quantified by the difference in both their output values and gradients with respect to its inputs. Several works have demonstrated the successes of Sobolev training Son et al. (2021); Buchholz (2022). For instance, tasks such as distillation compresses a target model into a smaller one so that the two models exhibit similar behavior Louati et al. (2022); Srinivas and Fleuret (2018). In the prediction of a synthetic gradient for training complex models, especially in the low-data regime, Sobolev training yields higher-accuracy models compared to classical training Jaderberg et al. (2017). The predictive power of Gaussian processes improves with the use of information about gradients and the Hessian matrix for Bayesian optimization Wu et al. (2017). Sobolev training has also been used in actor-critic-based reinforcement learning methods, where the action-value function is estimated using gradient information Parag et al. (2022); D'Oro and Jaśkowski (2020). The idea of gradient data has been successfully employed in various other domains Jalali et al. (2022); Vlassis et al. (2020); Son et al. (2021); Bouhlel et al. (2020).

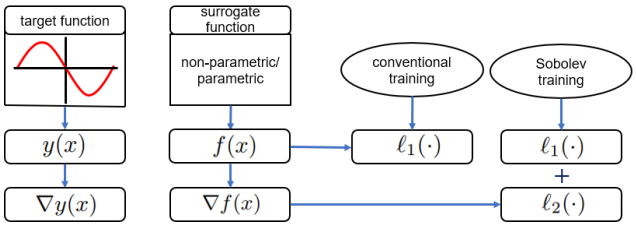


Fig. 1. Illustration of conventional and Sobolev training. $\ell_1(\cdot)$ and $\ell_2(\cdot)$ represents the loss function.

Despite the success and wide applicability of gradient data, there is little theoretical understanding and guarantees of when and by how much functional estimation accuracy improves using Sobolev training. In this paper, we analyze the effect of incorporating gradient data in supervised learning. We consider the ridge regression problem in the reproducing kernel Hilbert space (RKHS). The ability to approximate functions by nonparametric functional representation is provided by RKHS, thus rendering RKHS an important tool in many areas, especially kernel methods Schölkopf and Smola (2002). Kernel methods are among the most popular and frequently used tools in modeling complex relations, having substantial influence on several fields of control theory, machine learning and statistics Thorpe et al. (2022); van Waarde and Sepulchre (2022); Anjanapura Venkatesh et al. (2021); Sun et al. (2018). Representer's theorem, which permits one to parameterize the optimal solution by finitely many coefficients, is the essential property that renders kernel methods computationally practical

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

and the optimization over RKHS tractable. Because of this characteristic, RKHS is a desirable option for control problems Dubey et al. (2020); Huang et al. (2018). Recently, kernel-based algorithms involving kernel derivatives have been used to address numerous learning tasks Szabó and Sriperumbudur (2019); Wang et al. (2022). An approach based on random Fourier features to approximate kernel derivative is proposed in Fang et al. (2022); Sriperumbudur and Szabó (2015). The convergence analysis of gradient data for the learning algorithm in RKHS has been explored Shi et al. (2010), that uses a sampling operator for sample error and an integral operator in Sobolev space for the approximation error. An approach from convex analysis in the framework of multitask vector learning was employed for error analysis of Sobolev training in RKHS Sheng et al. (2018). However, the previous analyses are unable to explain the advantages of gradient data in terms of sample efficiency. The goal of this work is to understand whether incorporating gradient data into the learning process always improves prediction accuracy and the extent to which it affects learning rates.

The following is a summary of our contributions:

(1) We analytically express how including the gradient reduces learning error and derive the learning rate of order $\mathcal{O}(n^{-\frac{1}{2}+\epsilon})$.
(2) The generalization error bound exhibits a precise relation between Sobolev and classical training, hence enabling one to identify the condition when the gradient data can be resourceful or harmful.
(3) We observe that there exists a threshold for Lipschitz constant of the target function below which the gradient data improves the performance guarantee, especially in the low-data regime. For the target function with a larger Lipschitz constant there is a threshold of training data size beyond which the gradient data for Sobolev training performs better than the classical training.

The rest of the paper is organized as follows. Section II describes the problem setup. In Section III, we present our main results on learning rates for Sobolev training and briefly sketch the main ideas of the proof. In Section IV, we conduct numerical experiments to analyze the effect of gradient data in the learning process. Section V concludes the paper with some future directions.

2. PROBLEM SETTING AND PRELIMINARIES

Let $X \subset \mathbb{R}^d$ be a compact convex set and $Y \subset \mathbb{R}^{d+1}$. Instead of the standard supervised learning setup where one is given a training dataset consisting of data pairs $\{x_i, y_i\}_{i=1}^n$, we consider the case where the training dataset is augmented by gradient data as $\{x_i, y_i, \tilde{y}_i\}_{i=1}^n$ (where, $\tilde{y}_i$ is sample value for the gradient of target function at x_i). To leverage such additional information, the learning algorithm we study here is expressed as a regularized regression problem introduced as:

$$f_z = \arg\min_{f\in\mathcal{H}_K} \{\frac{1}{n}(\sum_{i=1}^n (y_i^0 - f(x_i))^2 + \beta\sum_{i=1}^n \|\tilde{y}_i - \nabla f(x_i)\|_2^2) + \lambda\|f\|_{\mathcal{H}_K}^2\}, \tag{1}$$

where $\lambda > 0$is the regularization parameter. In statistical learning theory, parameter λ generally depends on the sample size as for example $\lambda = \lambda(n)$with $\lim_{n\to\infty}\lambda(n) = 0$. The parameter β represents the inclusion of gradient data; if $\beta = 0$, the problem (1) reduces to the conventional learning algorithm. $\mathcal{H}_K$ is the reproducing kernel Hilbert space associated with the kernel $K : X \times X \to \mathbb{R}$, characterized by the following two properties:

(1) The operator $K_x = K(x,\cdot) \in \mathcal{H}_K, \quad \forall x \in X.$
(2) For $\alpha = 0, 1, \ldots, d$, the reproducing property
$$\langle (D^\alpha K)_x, f\rangle_{\mathcal{H}_K} = D^\alpha f(x), \quad \forall x \in X.$$

An appealing property of RKHS is that its geometry makes optimization problem (1) defined over $\mathcal{H}_K$ computationally tractable. The explicit form of the estimator is given by the Representer theorem. The Representer theorem for problem (1) was provided in Zhou (2008); Shi et al. (2010), demonstrated that minimization over the potentially infinite dimensional space $\mathcal{H}_K$ can be achieved in a finite-dimensional subspace produced by $\{K_{x_i}(\cdot)\}$, and their partial derivatives. To account for noise in gradient data and study its effect on the learning algorithm, we assume the training set $\{(x_i, y_i, \tilde{y}_i)\}_{i=1}^n \in Z^n$ is drawn independently from a nondegenerate Borel probability measure ρ on $X \times Y$. The generalization error for a function $f : X \to Y$ is defined as:

$$\varepsilon(f) = \int_{X\times Y} (y^0 - f(x))^2 + \beta\|\tilde{y} - \nabla f(x)\|_2^2 d\rho \tag{2}$$

The ρ regression function minimizes generalization error and is defined as:

$$F_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X. \tag{3}$$

Here $\rho(\cdot|x)$ is the conditional distribution of ρ. If we denote $y \in Y$ as $y = (y^0, \tilde{y})$, where $\tilde{y} = (y^1, \ldots, y^d)$. We have

$$F_\rho(x) = (f_\rho(x), \tilde{f}_\rho(x)) \tag{4}$$

where

$$f_\rho(x) = \int_{[-M,M]} y^0 d\rho(y|x)$$

and

$$\tilde{f}_\rho(x) = \left(\int_{[-B,B]} y^1 d\rho(y|x), \ldots, \int_{[-B,B]} y^d d\rho(y|x)\right).$$

The efficiency of algorithm (1) is measured by the difference between f_z and the regression function f_ρ. To measure the efficiency of the algorithm, we reformulate problem (1), let us define:

$$\bar{\mathcal{H}}_K = \left\{\bar{f} = (f, f_1, \ldots, f_d)^\top : f \in \mathcal{H}_K, f_\alpha \in \mathcal{H}_K^\alpha\right\},$$

where $\mathcal{H}_K^\alpha = \left\{f_\alpha(x) = \frac{\partial f(x)}{\partial x^\alpha} : \quad f(x) \in \mathcal{H}_K\right\}$, and $\alpha = 1, 2, \ldots, d$. Then, we can rewrite the algorithm (1) as:

$$\bar{f}_z = \arg\min_{\bar{f}\in\bar{\mathcal{H}}_K} \varepsilon_z(\bar{f}) + \lambda\|f\|_{\mathcal{H}_K}^2, \tag{5}$$

where,

$$\varepsilon_z(\bar{f}) = \frac{1}{n}\sum_{i=1}^n \left\{(y_i^0 - f_z(x_i))^2 + \beta\|\tilde{y}_i - \nabla f_z(x_i)\|_2^2\right\}.$$

Algorithm (5) is neither the same as the usual least square regression nor the multitask learning model since it uses the penalty $\|f\|_{\mathcal{H}_K}^2$ not $\|\bar{f}\|_{\mathcal{H}_K}^2$. Let $L_2(\rho)$ be the class

of all square integrable functions with respect to the measure ρ with the norm defined as $\|f\|_\rho = \|f\|_{L^2(\rho)} = \left(\int_X |f(x)|^2 d\rho x\right)^{1/2} < \infty$. One can see that

$$\|\bar{f}_z - F_\rho\|^2_{\rho,\beta} = \varepsilon(\bar{f}_z) - \varepsilon(F_\rho). \tag{6}$$

where

$$\varepsilon(\bar{f}) = \int_Z (y_i^0 - f(x_i))^2 + \beta\|\tilde{y}_i - \nabla f(x_i)\|_2^2 d\rho,$$

and

$$\|\bar{f}_z - F_\rho\|^2_{\rho,\beta} = \|f_z - f_\rho\|^2_\rho + \beta\|\nabla f_z - \nabla f_\rho\|^2_\rho.$$

We would anticipate that $\bar{f}_z$ is a good approximation to the minimizer F_ρ of the error $\varepsilon(\bar{f})$. As the minimization (5) is taken for the discrete quantity ε_z, the approximation of F_ρ by $\bar{f}_z$ involves the capacity of the function space $\bar{\mathcal{H}}_K$. Here capacity is measured by covering number.

Definition 1. *Sheng et al. (2018) The covering number $\mathcal{N}(S,\eta)$ is defined as the smallest positive integer l such that there are l disks in metric space S with radius η covering S. We call compact subset E of a metric space has a logarithmic complexity exponent $s \geq 0$, if there is a constant $c_s > 0$ such that the closed ball of radius R centered at origin i.e. $\mathcal{B}_R = \left\{f \in \mathcal{H}_K : \|f\|^2_{\mathcal{H}_K} \leq R\right\}$ satisfies*

$$\log \mathcal{N}(\mathcal{B}_R, \eta) \leq c_s \left(\frac{R}{\eta}\right)^s, \quad \forall \eta > 0. \tag{7}$$

Remark 1. *Note that inequality (7) is standard to measure the capacity of $\mathcal{H}_K$ Shi (2022); Cucker and Zhou (2007). When X is a bounded domain and $K \in C^r(X \times X)$, condition (7) holds true with $s = \frac{2n}{r}$. In particular, if $K \in C^\infty(X \times X)$, inequality (7) is valid for an arbitrary small $s > 0$.*

3. MAIN RESULTS

In this section, we state our main results under some basic assumptions and sketch the main ideas of our proof.

Assumption 1. *Let $M, B > 0$, be given positive real numbers, $\rho(y^0|x)$ and $\rho(\tilde{y}|x)$ is almost everywhere supported on $[-M, M]$ and $[-B, B]^d$ respectively, that is $|y^0| \leq M$ and $|y^i| \leq B$. It follows from the definition (3) of F_ρ that $|f_\rho| \leq M$ and $|\tilde{f}_\rho| \leq B$.*

Assumption 2. *Let $K \in C^2(X \times X)$, and K be the Mercer kernel—continuous symmetric function K such that the matrix $(K(x_i, x_j))_{i,j}^d$ is positive semidefinite for any finite set of points $\{x_1, \ldots, x_d\} \subset X$. Then $\mathcal{H}_K$ can be embedded into both $C^1(X)$ and $C^1(X)$, and the following relations hold:*

$$\left|\frac{\partial f(x)}{\partial x^\alpha}\right| \leq \kappa \|f\|_{\mathcal{H}_K}, \quad \forall x \in X, \quad \forall \alpha = 0, 1, \ldots, d, \tag{8}$$

where,

$$\kappa = \sup_{x,y \in X,\ 0 \leq \alpha, \beta \leq d} \sqrt{\left|\frac{\partial^2 K(x,y)}{\partial x^\alpha \partial x^\beta}\right|}.$$

3.1 Learning rates for Sobolev training

Let $\mathcal{H}^1_{\rho_x}$ be the Sobolev space consisting of the functions $f \in L^2_{\rho x}$ with all partial derivatives belonging to $L^2_{\rho x}$, whose norm $\|f\|_{\mathcal{H}^1_{\rho x}}$ is induced by the inner product

$$\langle f, g\rangle_{\mathcal{H}^1_{\rho x}} = \int_X f(x)g(x) + \beta \nabla_x f(x) \cdot \nabla_x g(x) d\rho_x,$$

where, ρ_x is the marginal distribution of ρ on X. Define an integral operator $L = L_{k,\rho x} : \mathcal{H}^1_{\rho x} \to \mathcal{H}^1_{\rho x}$ associated with kernel K, $x \in X$, $f \in \mathcal{H}^1_{\rho_x}$ and Borel measure ρx by

$$Lf(x) = \int_X K_x(y)f(y) + \beta \nabla(K_x)(y) \cdot \nabla f(y) d\rho_x(y).$$

When ρ is perfect i.e., $\tilde{f}_\rho(x) = \nabla_x f_\rho(x)$, and

$$\|f_z - f_\rho\|_\rho = \|f_z - f_\rho\|_{\mathcal{H}^1_{\rho x}}. \tag{9}$$

Theorem 1. *Let $X \subset \mathbb{R}^n$, is compact convex set with diameter τ, the kernel K satisfies Eq. (7) and $D(\lambda) \leq \lambda^\gamma$ for some $0 < \gamma \leq 1$. Thus, for $0 < \delta < 1$, there is a set $V_R \subset Z^n$, with $\rho(V_R) \leq \delta$, such that for all $z \in \mathcal{W}(R)\backslash V_R$, and $f_z \in B_R$ we have*

$$\begin{aligned}\|f_z - f_\rho\|^2_\rho \leq \left(\frac{\tau^2}{\tau^2 + \pi^2\beta}\right)\{&\frac{4M^2}{\lambda}\left((\kappa+3)^2(1+d\beta)\right) \\ &v^*(n, \delta/2) + \frac{22\kappa^2(1+d\beta)D(\lambda)\log\frac{2}{\delta}}{3n\lambda} \\ &+ 2D(\lambda) + \frac{66(M^2 + d\beta B^2)\log\frac{2}{\delta}}{n}\},\end{aligned} \tag{10}$$

where

$$v^*(n, \delta/2) \leq \max\left\{\frac{55}{n}\log\left(\frac{2}{\delta}\right), \left(\frac{55 c_s \kappa^s}{n}\right)^{\frac{1}{1+s}}\right\}.$$

Proof. We moved the detailed proof to an Appendix. □

The error bound (10) is obtained by the combination of Lemma 2 and Lemma 3 along with the use of the Poincare inequality [see, Lemma 4.1; Constantine (2015)].

Remark 2. *The Poincare inequality characterizes the relation about the variance of a function and its derivative in the spirit of the Sobolev inequality. It is a standard technical assumption for investigating the empirical convergence of the error analysis. In our analysis, Poincare inequality helps us to understand that the additional data for function gradients lead to improvements in the learning performance of the algorithm.*

Remark 3. *Looking at the expression on the right side of (10), we observe that if $\beta \neq 0$ the factor $\left(\frac{\tau^2}{\tau^2+\pi^2\beta}\right) < 1$, thus the gradient data helps to improve sample efficiency comparative to classical training problem.*

Remark 4. *If $\lambda = n^{\frac{-1}{(1+2r)(1+s)}}$, then the convergence rate is of order $\mathcal{O}\left(n^{-\frac{2r}{(1+2r)(1+s)}}\right)$. For C^∞ kernels, s can be arbitrary small. Then the convergence rate would be $\mathcal{O}(n^{-1/2+\epsilon})$ for any $\epsilon > 0$, achieved with $r = 1/2$.*

Remark 5. *For sample error estimate, we require the confidence $\mathcal{N}(\eta)\exp\{\frac{n\eta}{55}\}$ to be atmost δ for $0 < \delta < 1$. To realize this confidence, $v^*(n, \delta)$ is defined, obtained by the unique solution of the equation $\log\mathcal{N}(\eta) - \frac{n\eta}{55} = \log\delta$.*

From Theorem (1), a convergence property of gradient data for Sobolev training follows.

Corollary 1. *Let $0 < \delta < 1$ be arbitrary. Choose $\lambda = \lambda(n)$ satisfies $\lambda(n) \to 0$, $\lim_{n\to\infty} n\lambda(n) \geq 1$, and $v^*(n, \delta/2)/\lambda(n) \to 0$. If $D(\lambda) \to 0$, then for any $\epsilon > 0$, there is some $n_{\delta,\epsilon}$ such that with confidence $1 - \delta$, the following results holds.*

$$\|f_z - f_\rho\|^2_\rho \leq \epsilon, \quad \forall n \geq n_{\delta,\epsilon} \tag{11}$$

3.2 Proof framework

In this subsection, we sketch the framework of proof for Theorem 1. The idea of error decomposition has been used in the analysis of regularization scheme. In order to estimate the error $\|f_z - f_\rho\|$, we need the intermediate function. Let $\bar{f}_\lambda$ be a data free limit of (1) defined by

$$\bar{f}_\lambda = \arg\min_{\bar{f}\in\mathcal{H}_K} \left\{ \|\bar{f} - F_\rho\|^2_{\rho,\beta} + \lambda\|f\|^2_{\mathcal{H}_K} \right\}. \tag{12}$$

Then, the error decomposition follows from the relation $\varepsilon(\bar{f}_z) - \varepsilon(F_\rho) \le \varepsilon(\bar{f}_\lambda) - \varepsilon(F_\rho) + \lambda\|f_\lambda\|^2_{\mathcal{H}_K}$, which can be bounded by

$$\begin{aligned}\{\varepsilon(\bar{f}_\lambda) - \varepsilon(F_\rho) + \lambda\|f_\lambda\|^2_{\mathcal{H}_K}\} \\ + \left\{\varepsilon(\bar{f}_z) - \varepsilon_z(\bar{f}_z) + \varepsilon_z(\bar{f}_\lambda) - \varepsilon(\bar{f}_\lambda)\right\}\end{aligned} \tag{13}$$

The first term in the (13) is called the *regularization error* and the second term is called *sample error*. The regularization error for regularizing function $\bar{f}_\lambda$ is defined as

$$D(\lambda) = \varepsilon(\bar{f}_\lambda) - \varepsilon(F_\rho) + \lambda\|f_\lambda\|^2_{\mathcal{H}_K}. \tag{14}$$

The decay rate of regularization error is important for bounding the first term in (13), and also crucial for bounding the sample error. The decay of $\lambda(n)$ as $n \to \infty$ determines the size of hypothesis space and hence the sample error estimate. Therefore, we need to understand the choice of the parameter λ from the bound for $D(\lambda)$. The sample error in (13) can be written as

$$\begin{aligned}&\varepsilon(\bar{f}_z) - \varepsilon_z(\bar{f}_z) + \varepsilon_z(\bar{f}_\lambda) - \varepsilon(\bar{f}_\lambda) = \\ &\left\{E(\xi_1) - \frac{1}{n}\sum_{i=1}^n \xi_1(z_i)\right\} + \left\{\frac{1}{n}\sum_{i=1}^n \xi_2(z_i) - E(\xi_2)\right\},\end{aligned} \tag{15}$$

where

$$\begin{aligned}\xi_1 =&(f_z(x) - y^0)^2 + \beta\|\nabla f_z(x) - \tilde{y}\|^2_2 \\ &- \left((f_\rho(x) - y^0)^2 + \beta\|\nabla f_\rho(x) - \tilde{y}\|^2_2\right).\end{aligned}$$

$$\begin{aligned}\xi_2 =&(f_\lambda(x) - y^0)^2 + \beta\|\nabla f_\lambda(x) - \tilde{y}\|^2_2 \\ &- \left((f_\rho(x) - y^0)^2 + \beta\|\nabla f_\rho(x) - \tilde{y}\|^2_2\right).\end{aligned}$$

The first term on the right hand side of (15) is depending on ξ_1, which is not a fixed random variable since the function f_z is changing with the sample $\mathbf{z}$, we shall bound this using covering number of the ball $\mathcal{B}_R$. While the second term on the right-hand side of (15) depends on fixed random variable ξ_2 on (Z, ρ), we shall use the Bernstein inequality to bound this term. To do this however we need the bounds for $\|f_\lambda\|_{\mathcal{H}_K}$.

Lemma 1. *For $\lambda > 0$,*

$$\|f_\lambda\|_{\mathcal{H}_K} \le \sqrt{\frac{D(\lambda)}{\lambda}}$$

Proof. Since $\bar{f}_\lambda$ is the minimizer of (12), thus by definition of $D(\lambda)$, we have

$$\lambda\|f_\lambda\|^2_{\mathcal{H}_K} \le \varepsilon(\bar{f}_\lambda) - \varepsilon(F_\rho) + \lambda\|f_\lambda\|^2_{\mathcal{H}_K} = D(\lambda)$$

Thus the inequality holds. □

Lemma 2. *Let ξ_2 be a fixed random variable on a probability space Z and satisfying $|\xi_2(z) - E(\xi_2)| \le 2c_1$ for almost all $z \in Z$. For every $0 < \delta < 1$, with confidence atleast $1 - \delta$, there holds*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \xi_2(z_i) - E(\xi_2) \le& \frac{11\kappa^2(1+d\beta)D(\lambda)\log\frac{1}{\delta}}{3n\lambda} + D(\lambda) \\ &+ \frac{33(M^2 + d\beta B^2)\log\frac{1}{\delta}}{n}.\end{aligned} \tag{16}$$

Where c_1 is given by

$$c_1 = \left(\kappa\sqrt{\frac{D(\lambda)}{\lambda}} + 3M\right)^2 + \beta\left(\kappa\sqrt{\frac{dD(\lambda)}{\lambda}} + 3\sqrt{d}B\right)^2.$$

Proof. We moved the detailed proof to an Appendix. □

Now, we present the error analysis deal with error term $\varepsilon(\bar{f}_z) - \varepsilon_z(\bar{f}_z)$ on the right-hand side of (15), which is more difficult to deal with because ξ_1 is not really a single variable, since the function f_z depends on the the sample $\mathbf{z}$ itself. We will use the idea of ERM to bound this term by mean of covering number.

For $R > 0$, denote $\mathcal{W}(R) := \{\mathbf{z} \in Z^n : \|f_z\|_{\mathcal{H}_K} \le R\}$, and define $\mathcal{F}_R$ to be the set of functions from Z to $\mathbb{R}$.

$$\begin{aligned}\mathcal{F}_R := \{&(f_z(x) - y^0)^2 + \beta\|\nabla f_z(x) - \tilde{y}\|^2_2 \\ &- \left((f_\rho(x) - y^0)^2 + \beta\|\nabla f_\rho(x) - \tilde{y}\|^2_2\right) : f \in \mathcal{B}_R\}.\end{aligned}$$

Lemma 3. *Consider the set $\mathcal{F}_R$. Each function $g \in \mathcal{F}_R$ has the form $g(z) = (f_z(x) - y^0)^2 + \beta\|\nabla f_z(x) - \tilde{y}\|^2_{\mathbb{R}^d} - \left((f_\rho(x) - y^0)^2 + \beta\|\nabla f_\rho(x) - \tilde{y}\|^2_2\right)$, such that there is a set V'_R of measure $\delta/2$ and $f_z \in \mathcal{B}_R$. Then for $\epsilon > 0$ and $R \ge M$,*

$$\begin{aligned}E(\xi_1) - \frac{1}{n}\sum_{i=1}^n \xi_1(z_i) &= \varepsilon(\bar{f}_z) - \varepsilon(F_\rho) - \left(\varepsilon_z(\bar{f}_z) - \varepsilon_z(F_\rho)\right) \\ &\le \frac{1}{2}\left(\varepsilon(\bar{f}_z) - \varepsilon(F_\rho)\right) + 2R^2(\kappa+3)^2(1+d\beta)v^*(n,\delta/2).\end{aligned} \tag{17}$$

Proof. We moved the detailed proof to an Appendix. □

4. NUMERICAL EXPERIMENTS

In this section, experiments are conducted to investigate the performance of gradient data for Sobolev training in comparison to classical training.We consider the task of regression on a set of well-known low-dimensional functions used for bench marking optimization methods, where information about the derivatives is available during learning. Since the task is standard regression, we choose all losses to be L2 errors. In our experiments, we use the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/2\gamma^2)$ with $\gamma = 1$. Note that related error curves and standard deviations are obtained by running the experiments 100 times.

1. Effect of the Lipschitz constant on RMSE: For the first experiment, we considered the Styblinski-Tang function as a target function defined over the interval $[-1, 1]$. We compared the root-mean-square error (RMSE) at the test data (say $n = 20$) received by Sobolev training and the classical learning algorithm, with different Lipschitz constants of the target function for a fixed number of training data size (say $n = 30$) sampled randomly. Looking at the plots in Fig. 2, we make the following important observations. First, there exists a threshold such that if the

Lipschitz constant of the target function is less than the threshold value, the Sobolev training algorithm outperforms the classical learning algorithm. Second, the variation in the error with Sobolev training is more profound than the classical algorithm, as the Lipschitz constant of the target function increases.

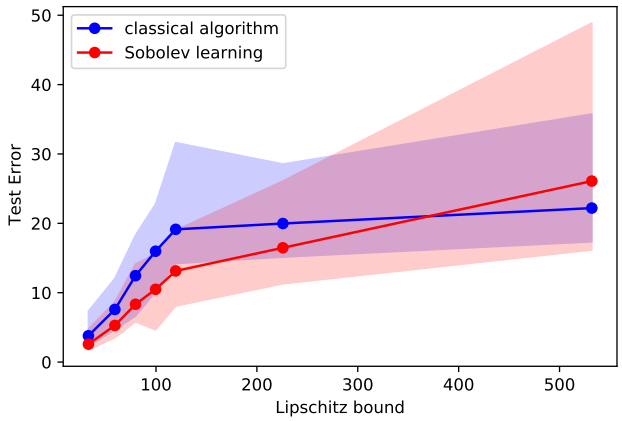
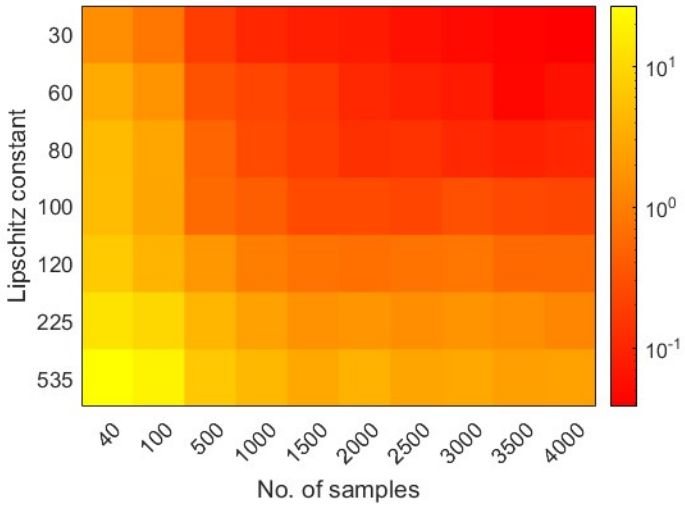


Fig. 2. Empirical comparison of root mean square error with different Lipschitz constant between classical and Sobolev training algorithms.

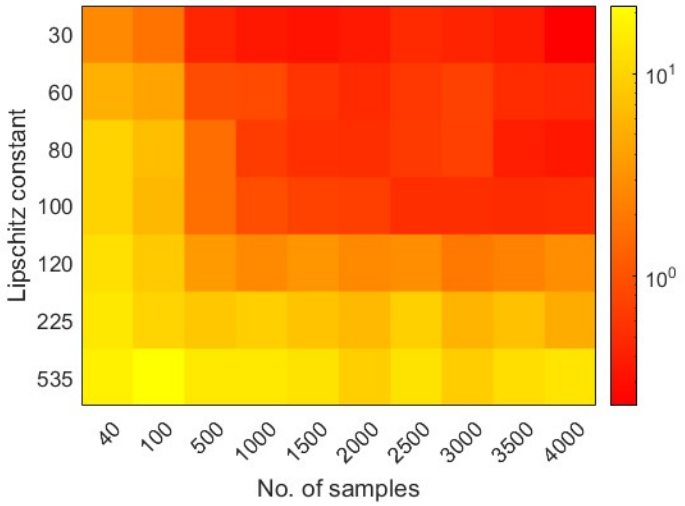
2. Effect of training data amount on RMSE and Lipschitz constant: Fig. 3 shows the performance of gradient data for Sobolev training and the classical training algorithm from the perspective of the test error in relation to the number of training samples and the Lipschitz constant. In both plots, the amount of training data is plotted on the x-axis, the Lipschitz bound is plotted on the y-axis, and the color-bar elucidates the test error in ascending order. The trend along horizontal direction depicts as the number of samples increases, the error decreases, whereas the trend along the vertical direction displays as the Lipschitz constant of the target function increases the test error increases. In general it is observed that the error with gradient data is less than that of the classical algorithm. The improvement with gradient data is more substantial compared to the classical algorithm as we increase the size of the training samples. Previously, in Fig. 2, we have shown that there is a threshold value for the Lipschitz constant of the target function, beyond which the classical algorithm outperforms the Sobolev training algorithm. Fig. 3a and Fig. 3b shows the fascinating trade-off between the Lipschitz bound and training data size and indicates that there exists a threshold in the context of training data. If the training data size is greater than the threshold, Sobolev training always outperforms the classical learning algorithm, and vice versa.

5. CONCLUSION AND FUTURE WORK

In this paper, we have studied the generalization properties of gradient data for Sobolev training in RKHS. The excess error converges at a faster rate of $\mathcal{O}(n^{-\frac{1}{2}+\epsilon})$. The experimental results delineate the limitations of gradient data for Sobolev training. If the Lipschitz constant of the target function is below a threshold, Sobolev training outperforms classical training in terms of sample efficiency.



(a) Heatmap for Sobolev learning algorithm



(b) Heatmap for classical learning algrithm

Fig. 3. Test error is plotted against different number of training size and Lipschitz bound. The deep red color depicts the minimum error value and the bright yellow color represents the maximum error.

For a target function with a large Lipschitz constant, there is a threshold for training data size beyond which Sobolev training performs better than conventional training. Building on the present work, one direction that we are currently pursuing is developing a nonparametric actor-critic reinforcement learning (RL) algorithm based on the kernel method.

REFERENCES

Anjanapura Venkatesh, A.K., Shilton, A., Rana, S., Gupta, S., and Venkatesh, S. (2021). Kernel functional optimisation. *Advances in Neural Information Processing Systems*, 34, 4725–4737.

Bouhlel, M.A., He, S., and Martins, J.R. (2020). Scalable gradient–enhanced artificial neural networks for airfoil shape design in the subsonic and transonic regimes. *Structural and Multidisciplinary Optimization*, 61(4), 1363–1376.

Buchholz, S. (2022). Kernel interpolation in sobolev spaces is not consistent in low dimensions. In *Conference on Learning Theory*, 3410–3440. PMLR.

Cocola, J. and Hand, P. (2020). Global convergence of sobolev training for overparameterized neural networks. In *International Conference on Machine Learning, Optimization, and Data Science*, 574–586. Springer.

Constantine, P.G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM.

Cucker, F., Smale, S., et al. (2002). Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of computational Mathematics*, 2(4), 413–428.

Cucker, F. and Zhou, D.X. (2007). *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

Czarnecki, W.M., Osindero, S., Jaderberg, M., Swirszcz, G., and Pascanu, R. (2017). Sobolev training for neural networks. *Advances in Neural Information Processing Systems*, 30.

D'Oro, P. and Jaśkowski, W. (2020). How to learn a useful critic? model-based action-gradient-estimator policy optimization. *Advances in Neural Information Processing Systems*, 33, 313–324.

Dubey, A. et al. (2020). Kernel methods for cooperative multi-agent contextual bandits. In *International Conference on Machine Learning*, 2740–2750. PMLR.

Fang, K., Huang, X., Liu, F., and Yang, J. (2022). End-to-end kernel learning via generative random fourier features. *Pattern Recognition*, 109057.

Huang, Z., Guo, Y., Arief, M., Lam, H., and Zhao, D. (2018). A versatile approach to evaluating and testing automated vehicles based on kernel methods. In *2018 Annual American Control Conference (ACC)*, 4796–4802. IEEE.

Jaderberg, M., Czarnecki, W.M., Osindero, S., Vinyals, O., Graves, A., Silver, D., and Kavukcuoglu, K. (2017). Decoupled neural interfaces using synthetic gradients. In *International conference on machine learning*, 1627–1635. PMLR.

Jalali, M., Singh, M.K., Kekatos, V., Giannakis, G.B., and Liu, C.C. (2022). Fast inverter control by learning the opf mapping using sensitivity-informed gaussian processes. *arXiv preprint arXiv:2202.07500*.

Louati, H., Bechikh, S., Louati, A., Aldaej, A., and Said, L.B. (2022). Joint design and compression of convolutional neural networks as a bi-level optimization problem. *Neural Computing and Applications*, 1–23.

Novara, C., Nicolì, A., and Calafiore, G.C. (2022). Nonlinear system identification in sobolev spaces. *International Journal of Control*, 1–16.

Parag, A., Kleff, S., Saci, L., Mansard, N., and Stasse, O. (2022). Value learning from trajectory optimization and sobolev descent: A step toward reinforcement learning with superlinear convergence properties. In *International Conference on Robotics and Automation*.

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

Schölkopf, B. and Smola, A.J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press. URL `/bib/scholkopf/scholkopf2002learning/lwk.pdf`.

Sheng, B., Wang, J., and Xiang, D. (2018). Error analysis on hérmite learning with gradient data. *Chinese Annals of Mathematics, Series B*, 39(4), 705–720.

Shi, L., Guo, X., and Zhou, D.X. (2010). Hermite learning with gradient data. *Journal of computational and applied mathematics*, 233(11), 3046–3059.

Shi, X. (2022). *Applications of Neural Tangent Kernel Theory in Deep Learning*. Ph.D. thesis, Northeastern University.

Singh, M.K., Gupta, S., Kekatos, V., Cavraro, G., and Bernstein, A. (2020). Learning to optimize power distribution grids using sensitivity-informed deep neural networks. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 1–6. IEEE.

Son, H., Jang, J.W., Han, W.J., and Hwang, H.J. (2021). Sobolev training for physics informed neural networks. *arXiv e-prints*, arXiv–2101.

Srinivas, S. and Fleuret, F. (2018). Knowledge transfer with jacobian matching. In *International Conference on Machine Learning*, 4723–4731. PMLR.

Sriperumbudur, B. and Szabó, Z. (2015). Optimal rates for random fourier features. *Advances in neural information processing systems*, 28.

Sun, Y., Gilbert, A., and Tewari, A. (2018). But how does it work in theory? linear svm with random features. *Advances in Neural Information Processing Systems*, 31.

Szabó, Z. and Sriperumbudur, B. (2019). On kernel derivative approximation with random fourier features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 827–836. PMLR.

Thorpe, A., Lew, T., Oishi, M., and Pavone, M. (2022). Data-driven chance constrained control using kernel distribution embeddings. In *Learning for Dynamics and Control Conference*, 790–802. PMLR.

Tsay, C. (2021). Sobolev trained neural network surrogate models for optimization. *Computers & Chemical Engineering*, 153, 107419.

van Waarde, H. and Sepulchre, R. (2022). Training lipschitz continuous operators using reproducing kernels. In *Learning for Dynamics and Control Conference*, 221–233. PMLR.

Vlassis, N.N., Ma, R., and Sun, W. (2020). Geometric deep learning for computational mechanics part i: anisotropic hyperelasticity. *Computer Methods in Applied Mechanics and Engineering*, 371, 113299.

Wang, S., Yu, X., and Perdikaris, P. (2022). When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449, 110768.

Wu, A., Aoi, M.C., and Pillow, J.W. (2017). Exploiting gradients and hessians in bayesian optimization and bayesian quadrature. *arXiv preprint arXiv:1704.00060*.

Wu, Q., Ying, Y., and Zhou, D.X. (2006). Learning rates of least-square regularized regression. *Foundations of computational mathematics*, 6(2), 171–192.

Zhou, D.X. (2008). Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2), 456–463.

Appendix A. SUPPLEMENTARY MATERIAL

Proposition 1. *Cucker and Zhou (2007) Let ξ be a random variable on a probability space Z with mean $E(\xi)=\mu$ and variance $\sigma^2(\xi)=\sigma^2$, and satisfying $|\xi(z)-E(\xi)|\le M$ for almost all $z\in Z$. Then for all $\epsilon>0$,*

Bernstein

$$Prob_{z\in Z^n}\left\{\frac{1}{n}\sum_{i=1}^{n}\xi(z_i)-\mu\le\epsilon\right\}\le\exp\left\{-\frac{n\epsilon^2}{2(\sigma^2+\frac{1}{3}M\epsilon)}\right\}.$$

Hoeffding

$$Prob_{z\in Z^n}\left\{\frac{1}{n}\sum_{i=1}^{n}\xi(z_i)-\mu\le\epsilon\right\}\le\exp\left\{-\frac{n\epsilon^2}{2M^2}\right\}.$$

Lemma 4. *Sheng et al. (2018) If $\bar{f}_{z,\lambda}$ is unique solution of (5). Then, it satisfy the following inequality*

$$\|f_{z,\lambda}\|_{\mathcal{H}_K}\le\frac{M}{\sqrt{(\lambda)}}. \tag{A.1}$$

Proof. For proof see ref Sheng et al. (2018). □

Lemma 5. *Wu et al. (2006) Suppose a random variable ξ on Z satisfies $\mu=E(\xi)\ge 0$, and $\boldsymbol{z}=(z_i)_{i=1}^n$ are independent samples. If $|\xi-\mu|\le B$ almost everywhere and $E(\xi^2)\le c_\xi E(\xi)$ for some $c_\xi\ge 0$, then for every $\epsilon>0$ and $0<\alpha\le 1$, there holds*

$$Prob_{\boldsymbol{z}\in Z^n}\left\{\frac{\mu-\frac{1}{n}\sum_{i=1}^{n}\xi(z_i)}{\sqrt{\mu+\epsilon}}\ge\alpha\sqrt{\epsilon}\right\}\le\left\{-\frac{\alpha^2 n\epsilon}{2c_\xi+\frac{2}{3}B}\right\}.$$

Lemma 6. *Wu et al. (2006) Let $\mathcal{G}$ be a set of functions on Z such that for some $c_\rho\ge 0$, $|g-E(g)|\le B$, almost everywhere and $E(g^2)\le c_\rho E(g)$ for each $g\in\mathcal{G}$. Then for every $\epsilon>0$ and $0<\alpha\le 1$,*

$$Prob_{\boldsymbol{z}\in Z^n}\left\{\sup_{g\in\mathcal{G}}\frac{E(g)-\frac{1}{n}\sum_{i=1}^{n}g(z_i)}{\sqrt{E(g)+\epsilon}}\ge 4\alpha\sqrt{\epsilon}\right\}\le\mathcal{N}(\mathcal{G},\alpha\epsilon)\left\{-\frac{\alpha^2 n\epsilon}{2c_\rho+\frac{2}{3}B}\right\}.$$

Appendix B. MISSING PROOFS

B.1 Proof of lemma 2

As

$$\begin{aligned}|\xi_2| &= |(f_\lambda(x)-y^0)^2-(f_\rho(x)-y^0)^2\\ &\quad+\beta\left(\|\nabla f_\lambda(x)-\tilde{y}\|_2^2-\|\nabla f_\rho(x)-\tilde{y}\|_2^2\right)|\\ &\le|(f_\lambda(x)-f_\rho(x))\left[(f_\lambda(x)-y)+(f_\rho(x)-y)\right]|\\ &\quad+\beta\|\nabla f_\lambda(x)-\nabla f_\rho(x)\|_2(\|\nabla f_\lambda(x)-\tilde{y}\|_2\\ &\quad+\|\nabla f_\rho(x)-\tilde{y}\|_2)|\\ &\le(\|f_\lambda\|_2+3M)^2+\beta\left(\|\nabla f_\lambda\|_2+3\sqrt{d}B\right)^2.\end{aligned} \tag{B.1}$$

Furthermore, by lemma (1) we have

$$\|f_\lambda\|_2=\sqrt{|f_\lambda(x)|^2}\le\kappa\|f_\lambda\|_{\mathcal{H}_K}\le\kappa\sqrt{\frac{D(\lambda)}{\lambda}}$$

$$\|\nabla f_\lambda\|_2=\sqrt{\sum_{i=1}^{d}\left|\frac{\partial f_\lambda(x_i)}{\partial x_i}\right|^2}\le\sqrt{d}\kappa\|f_\lambda\|_{\mathcal{H}_K}\le\kappa\sqrt{\frac{dD(\lambda)}{\lambda}} \tag{B.2}$$

So, now inequality (B.1) takes the form

$$\begin{aligned}|\xi_2|&\le\left(\kappa\sqrt{\frac{D(\lambda)}{\lambda}}+3M\right)^2+\beta\left(\kappa\sqrt{\frac{dD(\lambda)}{\lambda}}+3\sqrt{d}B\right)^2\\ &=:c_1\end{aligned} \tag{B.3}$$

Hence $|\xi_2-E(\xi_2)|\le 2c_1$. Moreover, we have

$$E(\xi_2^2)\le c_1\|\bar{f}_\lambda-F_\rho\|_{\rho,\beta}^2$$

which implies that $\sigma^2(\xi_2)\le E(\xi_2^2)\le c_1 D(\lambda)$. Now by proposition (1) we apply the Bernstein inequality to ξ_2. It asserts that for any $\epsilon>0$,

$$\frac{1}{n}\sum_{i=1}^{n}\xi_2(z_i)-E(\xi_2)\le\epsilon$$

with confidence atleast

$$1-\exp\left\{-\frac{n\epsilon^2}{2(\sigma^2(\xi_2)+\frac{2c_1}{3}\epsilon)}\right\}\ge 1-\exp\left\{-\frac{n\epsilon^2}{2(c_1D(\lambda)+\frac{2c_1}{3}\epsilon)}\right\}. \tag{B.4}$$

Let ϵ^* be the unique positive solution of the quadratic equation

$$\epsilon^2-\epsilon\frac{4c}{3n}\log\left(\frac{1}{\delta}\right)-\frac{2c_1D(\lambda)}{n}\log\left(\frac{1}{\delta}\right)=0.$$

Then with confidence $1-\delta$ there holds $\frac{1}{n}\sum_{i=1}^{n}\xi_2(z_i)-E(\xi_2)\le\epsilon^*$. And

$$\begin{aligned}\epsilon^*&=\frac{2c_1}{3n}\log\left(\frac{1}{\delta}\right)+\sqrt{\left(\frac{2c_1}{3n}\log\left(\frac{1}{\delta}\right)\right)^2+\frac{2c_1D(\lambda)}{n}\log\left(\frac{1}{\delta}\right)}\\ &\le\frac{4c_1}{3n}\log\left(\frac{1}{\delta}\right)+\sqrt{\frac{2c_1D(\lambda)}{n}\log\left(\frac{1}{\delta}\right)}\\ &\le\frac{11c_1\log\left(\frac{1}{\delta}\right)}{6n}+D(\lambda)\\ &\le\frac{11\kappa^2(1+d\beta)D(\lambda)\log\frac{1}{\delta}}{3n\lambda}+D(\lambda)\\ &\quad+\frac{33(M^2+d\beta B^2)\log\frac{1}{\delta}}{n}.\end{aligned}$$

The last inequality follows after inserting the value of c_1, which is bounded by $\frac{2\kappa^2(1+d\beta)D(\lambda)}{\lambda}+18(M^2+d\beta B^2)$.

B.2 Proof of lemma 3

Consider the set $\mathcal{F}_R$. Each function $g\in\mathcal{F}_R$ has the form $g(z)=(f(x)-y^0)^2+\beta\|\nabla f(x)-\tilde{y}\|_2^2-\left((f_\rho(x)-y^0)^2+\beta\|\nabla f_\rho(x)-\tilde{y}\|_2^2\right)$, such that $f\in\mathcal{B}_R$. Hence $E(g)=\varepsilon(\bar{f})-\varepsilon(F_\rho)\ge 0$, $E_z(g)=\varepsilon_z(\bar{f})-\varepsilon_z(F_\rho)$. And

$$\begin{aligned}|g(z)|&=|(f(x)-y^0)^2-(f_\rho(x)-y^0)^2\\ &\quad+\beta\left(\|\nabla f(x)-\tilde{y}\|_2^2-\|\nabla f_\rho(x)-\tilde{y}\|_2^2\right)|\\ &\le|(f(x)-f_\rho(x))\left[(f(x)-y)+(f_\rho(x)-y)\right]|\\ &\quad+\beta\|\nabla f(x)-\nabla f_\rho(x)\|_2\\ &\quad(\|\nabla f(x)-\tilde{y}\|_2+\|\nabla f_\rho(x)-\tilde{y}\|_2)|\\ &\le(\|f\|_2+3M)^2+\beta\left(\|\nabla f\|_2+3\sqrt{d}B\right)^2.\end{aligned} \tag{B.5}$$

Furthermore,

$$\|f\|_{\mathbb{R}}=\sqrt{|f(x)|^2}\le\kappa\|f\|_{\mathcal{H}_K}\le\kappa R$$

$$\|\nabla f\|_{\mathbb{R}^d} = \sqrt{\sum_{i=1}^{d} \left|\frac{\partial f(x_i)}{\partial x_i}\right|^2} \leq \sqrt{d}\kappa \|f\|_{\mathcal{H}_K} \leq \sqrt{d}\kappa R \quad \text{(B.6)}$$

So, now inequality (B.5) takes the form

$$\begin{aligned} |g(z)| &\leq (\kappa R + 3M)^2 + \beta\left(\sqrt{d}\kappa R + 3\sqrt{d}B\right)^2 \\ &\leq R^2(\kappa+3)^2(1+d\beta) =: c_2 \end{aligned} \quad \text{(B.7)}$$

So we have $|g(z) - E(g)| \leq 2c_2$ almost everywhere. In addition,

$$E(g^2) \leq c_2 \|\bar{f}_z - F_\rho\|^2_{\rho,\beta}$$

Thus, $E(g^2) \leq c_2 E(g)$, for each $g \in \mathcal{F}_R$. Now by Lemma (6) with $\alpha = 1/4$ to the function set $\mathcal{F}_R$. We deduce that

$$\begin{aligned} &\sup_{f\in\mathcal{B}_R} \frac{\varepsilon(\bar{f}) - \varepsilon(F_\rho) - \left(\varepsilon_z(\bar{f}) - \varepsilon_z(F_\rho)\right)}{\sqrt{\varepsilon(\bar{f}) - \varepsilon(F_\rho) + \epsilon}} \\ &= \sup_{g\in\mathcal{B}_R} \frac{E(g) - \frac{1}{n}\sum_{i=1}^n g(z_i)}{\sqrt{E(g)+\epsilon}} \leq \sqrt{\epsilon} \end{aligned} \quad \text{(B.8)}$$

with confidence at-least

$$\begin{aligned} &1 - \mathcal{N}\left(\mathcal{F}_R, \frac{\epsilon}{4}\right) \exp\left\{-\frac{n\epsilon}{32c_2 + \frac{64c_2}{3}}\right\} \\ &\geq 1 - \mathcal{N}\left(\mathcal{F}_R, \frac{\epsilon}{4}\right) \exp\left\{\frac{-n\epsilon}{55\left(R^2(\kappa+3)^2(1+d\beta)\right)}\right\} \end{aligned}$$

Now we have to bound the covering number $\mathcal{N}\left(\mathcal{F}_R, \frac{\epsilon}{4}\right)$. To accomplish so, we should point forth that

$$\begin{aligned} &|(f_1(x) - y^0)^2 + \beta\|\nabla f_1(x) - \tilde{y}\|_2^2 \\ &\quad - \left((f_2(x) - y^0)^2 + \beta\|\nabla f_2(x) - \tilde{y}\|_2^2\right)| \\ &\quad \leq 2(M + \kappa R)\|f_1 - f_2\|_\infty \\ &\quad + 2\beta\sqrt{d}(\kappa R + B)\sqrt{\sum_{i=1}^{d} \|\frac{\partial f_1}{\partial x_i} - \frac{\partial f_2}{\partial x_i}\|_\infty^2} \\ &\quad \leq 2\kappa(M + \kappa R)\|f_1 - f_2\|_{\mathcal{H}_K} \\ &\quad + 2d\beta\kappa(B + \kappa R)\|f_1 - f_2\|_{\mathcal{H}_K} \\ &\quad = 2\kappa\left((M + \kappa R) + d\beta(B + \kappa R)\right)\|f_1 - f_2\|_{\mathcal{H}_K}. \end{aligned} \quad \text{(B.9)}$$

Since an $\left(\frac{\eta}{2\kappa R((M+\kappa R)+d\beta(\kappa R+B))}\right)$-covering of $\mathcal{B}_1$ yields an $\left(\frac{\eta}{2\kappa((M+\kappa R)+d\beta(\kappa R+B))}\right)$ covering of $\mathcal{B}_R$. We see that for any $\eta > 0$, an $\left(\frac{\eta}{2\kappa R((M+\kappa R)+d\beta(\kappa R+B))}\right)$- covering of $\mathcal{B}_1$ provides an η-covering of $\mathcal{F}_R$. That is

$$\begin{aligned} \mathcal{N}\left(\mathcal{F}_R, \frac{\epsilon}{4}\right) &\leq \mathcal{N}\left(\mathcal{B}_R, \frac{\epsilon}{8\kappa R((M+\kappa R)+d\beta(\kappa R+B))}\right) \\ &\leq \mathcal{N}\left(\mathcal{B}_R, \frac{2\epsilon}{2\kappa R^2((\kappa+3)^2+d\beta(\kappa+3)^2)}\right). \end{aligned} \quad \text{(B.10)}$$

Now by inequality (7) we have

$$\begin{aligned} &Prob_{z\in Z^n}\left\{\sup_{f\in\mathcal{B}_R} \frac{\varepsilon(\bar{f}) - \varepsilon(F_\rho) - \left(\varepsilon_z(\bar{f}) - \varepsilon_z(F_\rho)\right)}{\sqrt{\varepsilon(\bar{f}) - \varepsilon(F_\rho) + \epsilon}} \leq \sqrt{\epsilon}\right\} \\ &\geq 1 - \mathcal{N}\left(\mathcal{B}_R, \frac{\epsilon}{2\kappa R^2((\kappa+3)^2+d\beta(\kappa+3)^2)}\right) \\ &\quad \exp\left\{\frac{-n\epsilon}{55R^2\left((\kappa+3)^2+d\beta(\kappa+3)^2\right)}\right\} \\ &\geq 1 - \exp\left\{c_s\left(\frac{2\kappa R^2(\kappa+3)^2(1+d\beta)}{\epsilon}\right)^s\right\} \\ &\quad \exp\left\{\frac{-n\epsilon}{55R^2(\kappa+3)^2(1+d\beta)}\right\}. \end{aligned} \quad \text{(B.11)}$$

Note that from (B.11) we have $\sqrt{\varepsilon(\bar{f}) - \varepsilon(F_\rho) + \epsilon}\sqrt{\epsilon} \leq \frac{1}{2}\left(\varepsilon(\bar{f}) - \varepsilon(F_\rho)\right) + \epsilon$ with confidence

$$\begin{aligned} &1 - \left(\exp\left\{c_s\left(\frac{2\kappa R^2(\kappa+3)^2(1+d\beta)}{\epsilon}\right)^s\right\}\right) \\ &\quad \times \left(\exp\left\{\frac{-n\epsilon}{55R^2(\kappa+3)^2(1+d\beta)}\right\}\right). \end{aligned} \quad \text{(B.12)}$$

For $\delta > 0$, choose

$$\begin{aligned} \delta &= \left(\exp\left\{c_s\left(\frac{2\kappa R^2(\kappa+3)^2(1+d\beta)}{\epsilon}\right)^s\right\}\right) \\ &\quad \times \left(\exp\left\{\frac{-n\epsilon}{55R^2(\kappa+3)^2(1+d\beta)}\right\}\right). \end{aligned}$$

Thus we have unique ϵ obtain by following quadratic equation,

$$\begin{aligned} &\epsilon^{s+1} - \epsilon^s \frac{55R^2(\kappa+3)^2(1+d\beta)\log\left(\frac{1}{\delta}\right)}{n} \\ &\quad - \frac{55c_s 2^s \kappa^s R^{2s+2}\left((\kappa+3)^2(1+d\beta)\right)^{s+1}}{n} = 0. \end{aligned} \quad \text{(B.13)}$$

Then by famous result [Lemma 7; Cucker et al. (2002)], we have

$$\epsilon \leq 2R^2(\kappa+3)^2(1-d\beta)v^*(n,\delta) \quad \text{(B.14)}$$

where, $v^*(n,\delta) = \max\left\{\frac{55}{n}\log\left(\frac{1}{\delta}\right), \left(\frac{55c_s\kappa^s}{n}\right)^{\frac{1}{1+s}}\right\}$. Thus for ϵ there is a set $V_R' \subset Z^n$ of measure at most $\frac{\delta}{2}$ such that ,

$$\begin{aligned} &\varepsilon(\bar{f}) - \varepsilon(F_\rho) - \left(\varepsilon_z(\bar{f}) - \varepsilon_z(F_\rho)\right) \\ &\quad \leq \frac{1}{2}\left(\varepsilon(\bar{f}) - \varepsilon(F_\rho)\right) + 2R^2(\kappa+3)^2(1+d\beta)v^*(n,\delta/2). \end{aligned} \quad \text{(B.15)}$$

In particular, when $z \in \mathcal{W}(R)\backslash V_R'$, and $f_z \in B_R$, we have

$$\begin{aligned} &E(\xi_1) - \frac{1}{n}\sum_{i=1}^n \xi_1(z_i) = \varepsilon(\bar{f}_z) - \varepsilon(F_\rho) - \left(\varepsilon_z(\bar{f}_z) - \varepsilon_z(F_\rho)\right) \\ &\quad \leq \frac{1}{2}\left(\varepsilon(\bar{f}_z) - \varepsilon(F_\rho)\right) + 2R^2(\kappa+3)^2(1+d\beta)v^*(n,\delta/2). \end{aligned} \quad \text{(B.16)}$$

B.3 Proof of theorem 1

By proposition (2) with δ replaced by $\delta/2$. We can find another set $V_R^{''} \subset Z^n$ of measure at most $\delta/2$ such that for all $z \in Z^n \in V_R^{''}$,

$$\frac{1}{n}\sum_{i=1}^{n}\xi_2(z_i) - E(\xi_2) \leq \frac{11\kappa^2(1+d\beta)D(\lambda)\log\frac{2}{\delta}}{3n\lambda} + D(\lambda) + \frac{33(M^2+d\beta B^2)\log\frac{2}{\delta}}{n}.$$

Combining above bound with the bound of Lemma 3, we see that for $z \in \mathcal{W}(R)\backslash\left(V_R = V_R^{'}\bigcup V_R^{''}\right)$,

$$\begin{aligned}\varepsilon(\bar{f}_z) - \varepsilon_z(\bar{f}_z) + \varepsilon_z(\bar{f}_\lambda) - \varepsilon(\bar{f}_\lambda) &\leq \frac{1}{2}\left(\varepsilon(\bar{f}_z) - \varepsilon(F_\rho)\right) + 2R^2(\kappa+3)^2(1+d\beta)v^*(n,\delta/2) \\ &+ \frac{11\kappa^2(1+d\beta)D(\lambda)\log\frac{2}{\delta}}{3n\lambda} + D(\lambda) \\ &+ \frac{33(M^2+d\beta B^2)\log\frac{2}{\delta}}{n}.\end{aligned} \tag{B.17}$$

Now the above inequality together with (13), and lemma (4), tells us that $R = \frac{M}{\sqrt{\lambda}} > M$ for $\lambda \in (0,1]$ and we have

$$\begin{aligned}\|\bar{f}_z - F_\rho\|_\rho^2 &\leq 4R^2\left((\kappa+3)^2(1+d\beta)\right)v^*(n,\delta/2) \\ &+ \frac{22\kappa^2(1+d\beta)D(\lambda)\log\frac{2}{\delta}}{3n\lambda} + 2D(\lambda) \\ &+ \frac{66(M^2+d\beta B^2)\log\frac{2}{\delta}}{n} \\ &= \frac{4M^2}{\lambda}\left((\kappa+3)^2(1+d\beta)\right)v^*(n,\delta/2) \\ &+ \frac{22\kappa^2(1+d\beta)D(\lambda)\log\frac{2}{\delta}}{3n\lambda} + 2D(\lambda) \\ &+ \frac{66(M^2+d\beta B^2)\log\frac{2}{\delta}}{n}.\end{aligned} \tag{B.18}$$

Thus by Poincare inequality [see, Constantine (2015); lemma 4.1], we have

$$\|f_z - f_\rho\|_\rho^2 \leq \frac{\tau^2}{\pi^2}\|\nabla f_z - \nabla f_\rho\|_\rho^2 \tag{B.19}$$

Now, (B.18) takes the form,

$$\begin{aligned}\|f_z - f_\rho\|_\rho^2 \leq \left(\frac{\tau^2}{\tau^2+\pi^2\beta}\right)\Big(&\frac{4M^2}{\lambda}\left((\kappa+3)^2(1+d\beta)\right)v^*(n,\delta/2) \\ &+ \frac{22\kappa^2(1+d\beta)D(\lambda)\log\frac{2}{\delta}}{3n\lambda} + 2D(\lambda) \\ &+ \frac{66(M^2+d\beta B^2)\log\frac{2}{\delta}}{n}\Big).\end{aligned} \tag{B.20}$$