Optimal Learning Rates for Distribution Regression

Zhiying Fang
School of Data Science, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, Email: zyfang4-c@my.cityu.edu.hk
Zheng-Chu Guo
School of Mathematical Sciences, Zhejiang University
Hangzhou 310027, P. R. China, Email: guozhengchu@zju.edu.cn
Ding-Xuan Zhou
School of Data Science and Department of Mathematics, City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, Email: mazhou@cityu.edu.hk

Abstract

We study a learning algorithm for distribution regression with regularized least squares. This algorithm, which contains two stages of sampling, aims at regressing from distributions to real valued outputs. The first stage sample consists of distributions and the second stage sample is obtained from these distributions. To extract information from samples, we embed distributions to a reproducing kernel Hilbert space (RKHS) and use the second stage sample to form the regressor by a tool of mean embedding. We show error bounds in the L^2-norm and prove that the regressor is a good approximation to the regression function. We derive a learning rate which is optimal in the setting of standard least squares regression and improve the existing work. Our analysis is achieved by using a novel second order decomposition to bound operator norms.

Keywords: Distribution regression, reproducing kernel Hilbert space, mean embedding, integral operator, optimal learning rate

1 Introduction

Explosion of information not only causes curse of dimensionality, but also generates various patterns of useful data. Considering this perspective, classical regression may not be suitable to solve some problems such as those dealing with functional data or matrix-valued data. A research problem has arisen recently, called distribution regression, trying to make predictions dealing with data of probability measures. In the sense of regression, input data are no longer vectors in Euclidean spaces but probability distributions on a compact metric space $\tilde{X}$ on which a reproducing kernel k will be introduced below. Distinct from the standard regression setting, distribution regression has two stages of sampling. For the first stage, we are given data $\tilde{D} = \{(x_i, y_i)\}_{i=1}^l \subset X \times Y$ where X is the input space of probability distributions on $\tilde{X}$ and $Y = \mathbb{R}$ is the output space, that is, each x_i represents a distribution and y_i is the corresponding label. For the second and essential stage, we obtain samples $\left\{\{x_{i,j}\}_{j=1}^N\right\}_{i=1}^l$ from distributions $\{x_i\}_{i=1}^l$ accordingly, where each $x_{i,j}$ is a point in $\tilde{X}$.

To illustrate ideas of the above two-stages sampling process for distribution regression, we describe three examples. The first one is for analysis of functional data where X is the set of probability density functions on the interval $\tilde{X} = [0,1]$ and, for each i, a sample $\{x_{i,j} = x_i(j/N)\}_{j=1}^N$ is given by the values of the probability density function x_i at the sampling points $\{j/N\}_{j=1}^N$. Our target here is to learn a functional from X to Y through the sample $\left\{\left(\{x_i(j/N)\}_{j=1}^N, y_i\right)\right\}_{i=1}^l \subset \mathbb{R}^N \times Y$ instead of $\{(x_i, y_i)\}_{i=1}^l \subset X \times Y$. The second example on medical applications is borrowed from [24] where X is a pool of patients identified with a set of probability distributions on $\tilde{X} = [0,1]$ and for the ith patient x_i in a sample $\{x_i\}_{i=1}^l$, $\{x_{i,j}\}_{j=1}^N$ is given by blood tests made for x_i periodically at moments $\{j/N\}_{j=1}^N$. Here $\{y_i\}_{i=1}^l$ are the values of some health indicator of the patients. Our goal is to learn a mapping from the set of blood tests to the health indicator by observations on a large group of patients. The last example is a classical learning problem of multiple instance learning [7, 8, 20] where each instance in one bag is an i.i.d. sample drawn from an unknown distribution related to the bag.

In this paper we are interested in a kernel method induced by an important tool called *mean embedding* [2] to transform information. We embed the set of (probability) distributions to a reproducing kernel Hilbert space H and then learn a functional relation from embeddings to outputs. Let $(H = H(k), \| \cdot \|_H)$ be a reproducing kernel Hilbert space (RKHS) with a Mercer kernel $k : \tilde{X} \times \tilde{X} \to \mathbb{R}$ (meaning that k is symmetric, continuous and positive semidefinite). Then the mean embedding of a (probability) distribution x on $\tilde{X}$ is defined to be an element in the RKHS H given by

$$\mu_x = \int_{\tilde{X}} k(\cdot, s) dx(s) \in H.$$

Through this transformation, kernel methods for processing data on Euclidean spaces can be extended to those on the space of probability measures. When k is a characteristic kernel (see [12] and references therein), this transformation is injective meaning that for two distributions P and Q, $\|\mu_P - \mu_Q\|_H = 0$ if and only if $P = Q$. Hence the mean embedding approach enables functional analysis of distribution regression. In particular, the injectivity of mean embedding has been found to be very useful in many statistical applications that require unique representations of distributions including homogeneity testing [12], independence testing [11, 14], and dimensionality reduction [10]. Let X denote the set of Borel probability measures on $\tilde{X}$. We denote

$$X_\mu = \mu(X) = \{\mu_x : x \in X\} \subset H \tag{1}$$

the set of mean embeddings which is a separable compact set of continuous functions on $\tilde{X}$ [24]. Let ρ be a Borel probability measure on $Z = X_\mu \times Y$. For a function $f : X_\mu \to Y$, $f(\mu_x)$ represents the prediction of y based on μ_x. The least squares regression problem aims to find the minimizer of the expect risk

$$\mathcal{E}(f) = \int_Z (f(\mu_x) - y)^2 d\rho \tag{2}$$

over all measurable functions. This minimizer is referred to as the regression function defined by

$$f_\rho(\mu_x) = \int_Y y d\rho(y|\mu_x), \quad \mu_x \in X_\mu \tag{3}$$

with $\rho(\cdot|\mu_x)$ being the conditional distribution of ρ induced at $\mu_x \in X_\mu$. Since ρ is unknown, we may approximate f_ρ based on the first stage sample $D = \{(\mu_{x_i}, y_i)\}_{i=1}^l$ (mean embeddings of $\tilde{D} = \{(x_i, y_i)\}_{i=1}^l$) according to the least squares regression setting. The distribution regression setting considered in this paper is different: the probability distributions $\{x_i\}_{i=1}^l$ are still unknown, each of them is approximately available by a random sample $\{x_{i,j}\}_{j=1}^N$ of size $N \in \mathbb{N}$. So the distribution regression aims to learn the regression function f_ρ from the sample $\hat{D} = \left\{ \left(\{x_{i,j}\}_{j=1}^N, y_i \right) \right\}_{i=1}^l$. In this paper, we study the following least squares regularization algorithm in a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ associated with a Mercer kernel $K : X_\mu \times X_\mu \to \mathbb{R}$, given by

$$f_{\hat{D},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{l} \sum_{i=1}^l (f(\mu_{\hat{x}_i}) - y_i)^2 + \lambda \|f\|_K^2 \right\}, \tag{4}$$

where $\hat{x}_i = \frac{1}{N} \sum_{j=1}^N \delta_{x_{i,j}}$ is the empirical version of the distribution x_i determined by the sample $\{x_{i,j}\}_{j=1}^N$ and $\mu_{\hat{x}_i} = \frac{1}{N} \sum_{j=1}^N k(\cdot, x_{i,j})$ is its mean embedding, $\lambda > 0$ is a regularization parameter.

For the one stage sampling setting for algorithm (4), that is, the least squares regularization, or called the kernel ridge regression in statistics, is well studied in learning theory, e.g., [4, 21]. In this case, $\{x_i\}_{i=1}^l$ are vectors in Euclidean space, mininmax rates are derived recently by a novel integral operator approach [18]. General regularization algorithms [1] are also investigated which include kernel ridge regression as a special example, and can avoid the saturation phenomenon if the target function has enough regularity. Minimax learning rates are also established [3, 15, 17]. However, error analysis for the two stage sampling setting is more challenging. In [19], consistency for distribution regression algorithm (4) is derived via kernel density estimation. Recently, theoretical studies in [24] show that learning rates can be established under certain conditions on the target function and on the kernel K. We will compare our results with those in [24] in Section 3.

Our goal of this paper is to present minimax optimal learning rates for algorithm (4) under some mild conditions. The paper is organized as follows. We first state our main results in Section 2. In Section 3, we compare our results with those in the literature. Section 4 includes the error decomposition and our novel error analysis of bounding norms by second order decomposition. Proofs of main results are given in Section 5 and the appendix includes the proof of a lemma.

2 Main Results

Throughout the paper we assume that there exists a constant $M > 0$ such that $|y| \leq M$ almost surely. Also, kernels K and k are bounded, that is, the following two constants κ and B_k are finite

$$\kappa = \sup_{\mu_a \in X_\mu} \sqrt{K(\mu_a, \mu_a)}, \qquad B_k = \sup_{u \in \tilde{X}} k(u, u).$$

Let $h \in (0, 1]$ and $L > 0$. We assume the mapping $K_{(\cdot)} : X_\mu \to \mathcal{H}_K$ defined as $K_{(\mu_x)} = K(\mu_x, \cdot)$ is (h, L) Hölder continuous, that is

$$\|K_{\mu_a} - K_{\mu_b}\|_{\mathcal{H}_K} \leq L \|\mu_a - \mu_b\|_H^h, \quad \forall (\mu_a, \mu_b) \in X_\mu \times X_\mu. \tag{5}$$

Let $L^2_{\rho_{X_\mu}}$ be the Hilbert space of square-integrable functions with respect to ρ_{X_μ} from X_μ to Y, where ρ_{X_μ} is the marginal distribution of ρ on X_μ. Denote by $\| \cdot \|_\rho$ the corresponding norm of $L^2_{\rho_{X_\mu}}$ induced by the inner product $\langle f, g \rangle_{\rho_{X_\mu}} = \int_{X_\mu} f(s) g(s) d\rho_{X_\mu}(s)$. For the kernel $K : X_\mu \times X_\mu \to \mathbb{R}$, an integral operator L_K on $L^2_{\rho_{X_\mu}}$ is defined as

$$L_K(f) = \int_{X_\mu} K_{\mu_x} f(\mu_x) d\rho_{X_\mu}, \quad f \in L^2_{\rho_{X_\mu}}.$$

Since X_μ is compact and K is continuous, symmetric and positive semidefinite, L_K is compact and positive on $L^2_{\rho_{X_\mu}}$, and its rth power L_K^r is well defined for any $r > 0$.

Our error analysis is based on the following *regularity condition* imposed for the regression function

$$f_\rho = L_K^r (g_\rho) \quad \text{for some} \quad g_\rho \in L^2_{\rho_{X_\mu}}, \; r > 0. \tag{6}$$

It means f_ρ lies in the range of L_K^r and the special case $f_\rho \in \mathcal{H}_K$ corresponds to the choice $r = \frac{1}{2}$.

We use the *effective dimension* $\mathcal{N}(\lambda)$ to measure the complexity of $\mathcal{H}_K$ with respect to ρ_{X_μ}, which is defined to be the trace of the operator $L_K (L_K + \lambda I)^{-1}$ as

$$\mathcal{N}(\lambda) = \operatorname{Tr}\left(L_K (L_K + \lambda I)^{-1} \right), \quad \forall \lambda > 0.$$

We assume throughout the paper that $D = \{(\mu_{x_i}, y_i)\}_{i=1}^l$ is a sample independently drawn according to ρ, and $\{x_{i,j}\}_{j=1}^N$ is a sample independently drawn according to x_i for $i = 1, \cdots, l$. In this section, we state our main results on the difference between $f_{\hat{D},\lambda}$ and f_ρ with respect to L^2-norm in expectation taken for D and $\hat{D}$, to be proved in Section 4.

Theorem 1. *Assume $|y| \leq M$ almost surely and the regularity condition (6) with some $\frac{1}{2} \leq r \leq 1$ and the mapping $K_{(\cdot)} : X_\mu \to \mathcal{H}_K$ is (h, L) Hölder continuous with $h \in (0, 1]$ and $L > 0$. Then we have*

$$\begin{aligned} E\left[\left\| f_{\hat{D},\lambda} - f_\rho \right\|_\rho \right] \leq{} & \sqrt{3}(2 + \sqrt{\pi})^{\frac{1}{2}} L \frac{2^{\frac{h}{2}} B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}} N^{\frac{h}{2}}} \left\{ \frac{\kappa \sqrt{\mathcal{N}(\lambda)}}{\sqrt{l\lambda}} + 2\kappa L \left(2 + \sqrt{\pi}\right)^{\frac{1}{2}} (2B_k)^{\frac{h}{2}} \frac{1}{\lambda N^{\frac{h}{2}}} + 1 \right\} \\ & \times \left\{ M + 50M \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1 \right) \frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 8\kappa \lambda^{r - \frac{1}{2}} \|g_\rho\|_\rho \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1 \right)^{2r-1} + 2\sqrt{3}\kappa^{r + \frac{1}{2}} \|g_\rho\|_\rho \right\} \\ & + 30 \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1 \right)^2 \frac{M}{\kappa} \mathcal{B}_{l,\lambda} + (4 + \log 2) \lambda^r \|g_\rho\|_\rho \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1 \right)^{2r}, \end{aligned} \tag{7}$$

where $\mathcal{B}_{l,\lambda} = \frac{2\kappa}{\sqrt{l}} \left(\frac{\kappa}{\sqrt{l\lambda}} + \sqrt{\mathcal{N}(\lambda)} \right)$.

To derive explicit learning rates, we assume a *capacity assumption* that for some $\alpha > \frac{1}{2}$ and $c > 0$,

$$\mathcal{N}(\lambda) \leq c\lambda^{-\frac{1}{2\alpha}}, \quad \forall \lambda > 0. \tag{8}$$

By taking $\lambda = l^{-\frac{2\alpha}{4\alpha r+1}}$ and $N = l^{\frac{2\alpha(1+2r)}{h(4\alpha r+1)}}$, we get the following minimax optimal learning rates for the distribution regression algorithm (4).

Theorem 2. *Assume $|y| \leq M$ almost surely, the regularity condition (6) holds with $\frac{1}{2} \leq r \leq 1$, the capacity assumption (8) holds with $\alpha > \frac{1}{2}$ and the mapping $K_{(\cdot)} : X_\mu \to \mathcal{H}_K$ is (h, L) Hölder continuous with $h \in (0, 1]$ and $L > 0$. Then by taking $\lambda = l^{-\frac{2\alpha}{4\alpha r+1}}$ and $N = l^{\frac{2\alpha+4\alpha r}{h(4\alpha r+1)}}$, we have*

$$E\left[\left\|f_{\hat{D},\lambda} - f_\rho\right\|_\rho\right] \leq Cl^{-\frac{2\alpha r}{4\alpha r+1}}, \tag{9}$$

where the constant C is independent of l or N and will be given explicitly in the proof.

3 Comparison and Discussion

In this section, we compare our results with those in the existing literature. Analysis for minimax learning rates has been well established for the standard regularized least squares regression. For distribution regression, to the best of our knowledge, [24] is the only existing work containing learning rates for algorithm (4). In [24], consistency of the two stage sampling setup has been proved and probabilistic error bounds and explicit learning rates are also provided for algorithm (4). Let $\{\lambda_i\}$ be the positive eigenvalues of L_K arranged in a decreasing order, if the eigenvalues λ_i satisfy $\lambda_i \approx i^{-2\alpha}$ with $\alpha > \frac{1}{2}$, and regression function satisfies the regularity condition (6) with $\frac{1}{2} < r \leq 1$, then Theorem 5 in [24], asserts that with $N = l^{\frac{2\alpha+4\alpha r}{h(4\alpha r+1)}} \log l$ and $\lambda = l^{-\frac{2\alpha}{4\alpha r+1}}$,

$$\left\|f_{\hat{D},\lambda} - f_\rho\right\|_\rho^2 = \mathcal{O}_p\left(l^{-\frac{4\alpha r}{4\alpha r+1}}\right), \tag{10}$$

which reaches optimal minimax rates for one stage regression setup [4] when $\frac{1}{2} < r \leq 1$.
Theorem 9 in [24] states that under the *regularity condition* (6) for $r = \frac{1}{2}$ and by taking $N = l^{\frac{6}{5h}} \log l$,

$$\left\|f_{\hat{D},\lambda} - f_\rho\right\|_\rho^2 = \mathcal{O}_p\left(l^{-\frac{2}{5}}\right). \tag{11}$$

However this convergence rate (11) is suboptimal. Actually in our analysis, the *capacity assumption* is always true with $\alpha = \frac{1}{2}$ and $c = \kappa^2$. Since the eigenvalues of operator $L_K(L_K + \lambda I)^{-1}$ are $\left\{\frac{\lambda_i}{\lambda_i+\lambda}\right\}_i$, the trace of this operator can be bounded by $\mathcal{N}(\lambda) = \sum_i \frac{\lambda_i}{\lambda_i+\lambda} \leq \sum_i \frac{\lambda_i}{\lambda} = \frac{\mathrm{Tr}(L_K)}{\lambda} \leq \kappa^2\lambda^{-1}$. With the choice of $r = \frac{1}{2}$ and $\alpha = \frac{1}{2}$, and by taking $N = l^{\frac{1}{h}}$, the optimal learning rate can be reached to

$$E\left[\left\|f_{\hat{D},\lambda} - f_\rho\right\|_\rho^2\right] = \mathcal{O}\left(l^{-\frac{1}{2}}\right). \tag{12}$$

Theorem 1 and Theorem 2 show that our results hold for $\frac{1}{2} \leq r \leq 1$ and we obtain the minimax rates for algorithm (4) for $\frac{1}{2} \leq r \leq 1$ by a novel integral operator approach, which covers the case with $r = \frac{1}{2}$ $(f_\rho \in \mathcal{H}_K)$.

Another contribution of this paper is that we successfully remove a logarithmic term $\log l$ in error bounds in [24] as a logarithmic factor appearing in the classical regression setting in [4], which is also crucial to our analysis. Benefitting from the removal of this logarithmic term $\log l$ in Theorem 1, we eliminate a logarithmic term $\log l$ in the restriction on N for reaching the optimal minimax rates, that is, we only require $N = l^{\frac{2\alpha+4\alpha r}{h(4\alpha r+1)}}$.

It would be interesting to extend our analysis on distribution regression to other learning algorithms such as those in deep learning [25, 26].

4 Bounding Operator Norms by Second Order Decomposition

Our analysis is carried out by the following error decomposition

$$\left\|f_{\hat{D},\lambda} - f_\rho\right\|_\rho = \left\|f_{\hat{D},\lambda} - f_{D,\lambda} + f_{D,\lambda} - f_\rho\right\|_\rho \leq \left\|f_{\hat{D},\lambda} - f_{D,\lambda}\right\|_\rho + \|f_{D,\lambda} - f_\rho\|_\rho, \tag{13}$$

where $f_{D,\lambda}$ is the minimizer of the regularized least squares regularization scheme based on the first stage sample D (after mean embedding), that is,

$$f_{D,\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{l}\sum_{i=1}^{l} \left(f(\mu_{x_i}) - y\right)^2 + \lambda \|f\|_K^2 \right\}. \tag{14}$$

We analyze the error for the distribution regression algorithm by an integral operator approach. The Hölder continuity of the kernel K is only applied to the first term of (13) and this explains why N is related to h. Detailed estimate of the first term will be presented in Section 5. For the second term of (13), we successfully extend the range for r and maintain the optimal rate by using a novel second order decomposition of operators which was first introduced in [18]. It is crucial to our analysis and we also apply it to estimating the first term of (13).

We approximate the integral operator L_K by its first stage empirical version $L_{K,D(x)}$ defined as

$$L_{K,D(x)}(f) = \frac{1}{l}\sum_{i=1}^{l} f(\mu_{x_i}) K_{\mu_{x_i}} = \frac{1}{l}\sum_{i=1}^{l} \langle f, K_{\mu_{x_i}}\rangle_K K_{\mu_{x_i}} \quad f\in\mathcal{H}_K, \tag{15}$$

and it is conceivable that when l is large enough these two operators should be close. We define the sampling operator $S_D : \mathcal{H}_K \to \mathbb{R}^l$ as

$$S_D f = \left(f(\mu_{x_1}), \cdots, f(\mu_{x_l})\right)^T, \ f\in\mathcal{H}_K \tag{16}$$

and its dual operator $S_D^* : \mathbb{R}^l \to \mathcal{H}_K$ is given by

$$S_D^* c = \sum_{i=1}^{l} c_i K_{\mu_{x_i}}, \ \forall c\in\mathbb{R}^l. \tag{17}$$

Then by denoting the vector $y = (y_i)_{i=1}^l \in \mathbb{R}^l$, $f_{D,\lambda}$ has the following explicit form

$$f_{D,\lambda} = \left(L_{K,D(x)} + \lambda I\right)^{-1} \frac{1}{l} S_D^* y. \tag{18}$$

Improvements are based on the following second order decomposition for operators [15], which is applied for bounding the operator norms.

Lemma 3. *Let A and B be invertible operators on a Banach space. Then we have*

$$BA^{-1} = (B-A)\,B^{-1}\,(B-A)\,A^{-1} + (B-A)\,B^{-1} + I. \tag{19}$$

Applying this lemma to operators $A = L_{K,D(x)} + \lambda I$ and $B = L_K + \lambda I$ on $\mathcal{H}_K$, we can present a sharp bound for the term $\left\|(L_K+\lambda I)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\|$, which is essential in estimating $\|f_{D,\lambda} - f_\rho\|_\rho$ and will be proved in the appendix. For convenience, we use $\|\cdot\|$ to represent the operator norm. Denote the Gamma function by $\Gamma(u)$.

Lemma 4. *Let D be a sample drawn independently according to ρ. We have for $d \geq 0$,*

$$E\left[\left\|(L_K+\lambda I)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\|^d\right] \leq \left(2\Gamma(2d+1) + \log^{2d} 2\right)\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)^{2d}. \tag{20}$$

and for $d \geq 1$

$$E\left[\left\|(L_K+\lambda I)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^* y - L_{K,D(x)} f_\rho\right)\right\|^d\right] \leq \left(2\Gamma\left(d+1\right) + \log^{d} 2\right)\left(\frac{2M}{\kappa}\mathcal{B}_{l,\lambda}\right)^{d}. \tag{21}$$

Now we give the following estimate for the second term of (13).

Proposition 5. *Assume $|y| \leq M$ almost surely and the regularity condition (6) holds for some $\frac{1}{2} \leq r \leq 1$. Then the following estimate holds*

$$E\left[\|f_{D,\lambda} - f_\rho\|_\rho\right] \leq 30\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)^2 \frac{M}{\kappa}\mathcal{B}_{l,\lambda} + (4+\log 2)\,\lambda^r \|g_\rho\|_\rho \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)^{2r}. \tag{22}$$

Proof. We separate $\left\|f_{D,\lambda}-f_\rho\right\|_\rho$ into two terms

$$\begin{aligned}\left\|f_{D,\lambda}-f_\rho\right\|_\rho &= \left\|L_K^{\frac{1}{2}}\left\{\left(L_{K,D(x)}+\lambda I\right)^{-1}\frac{1}{l}S_D^* y-\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}+\lambda I\right)f_\rho\right\}\right\|_K \\ &\leq \left\|L_K^{\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}S_D^* y-L_{K,D(x)}f_\rho\right)\right\|_K+\lambda\left\|L_K^{\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{-1}L_K^r g_\rho\right\|_K,\end{aligned} \tag{23}$$

where we used the fact that for $g\in L^2_{\rho_{X_\mu}}$, $\|g\|_\rho=\left\|L_K^{\frac{1}{2}}g\right\|_K$.

For the first term of (23), by using the bound $\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-\frac{1}{2}}\right\|\leq 1$, we have

$$\begin{aligned}&\left\|L_K^{\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}S_D^* y-L_{K,D(x)}f_\rho\right)\right\|_K \\ =&\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(L_K+\lambda I\right)^{\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(L_K+\lambda I\right)^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^* y-L_{K,D(x)}f_\rho\right)\right\|_K \\ \leq&\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(L_K+\lambda I\right)\right\|\left\|\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^* y-L_{K,D(x)}f_\rho\right)\right\|_K,\end{aligned} \tag{24}$$

where we have used the fact found in [3] that for positive operators L_1 and L_2 on a Hilbert space and $s\in[0,1]$, $\|L_1^s L_2^s\|\leq\|L_1L_2\|^s$ with $s=\frac{1}{2}$ and the identity $\|L_1L_2\|=\|L_2L_1\|$.

Similarly, for the second term of (23), we get

$$\begin{aligned}&\lambda\left\|L_K^{\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{-1}L_K^r g_\rho\right\|_K \\ \leq&\lambda\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-\frac{1}{2}}\right\|\left\|\left(L_K+\lambda I\right)^{\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{-\frac{1}{2}}\right\|\left\|\left(L_{K,D(x)}+\lambda I\right)^{-\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{r-\frac{1}{2}}\right\| \\ &\left\|\left(L_{K,D(x)}+\lambda I\right)^{-r+\frac{1}{2}}\left(L_K+\lambda I\right)^{r-\frac{1}{2}}\right\|\left\|\left(L_K+\lambda I\right)^{-r+\frac{1}{2}}L_K^{r-\frac{1}{2}}\right\|\left\|L_K^{\frac{1}{2}}g_\rho\right\|_K \\ \leq&\lambda^r\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(L_K+\lambda I\right)\right\|^r\|g_\rho\|_\rho,\end{aligned} \tag{25}$$

where we have used bounds $\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-\frac{1}{2}}\right\|\leq 1$, $\left\|\left(L_{K,D(x)}+\lambda I\right)^{-\frac{1}{2}}\left(L_{K,D(x)}+\lambda I\right)^{r-\frac{1}{2}}\right\|\leq\lambda^{r-1}$ and $\|L_1^s L_2^s\|\leq\|L_1L_2\|^s$ with $s=\frac{1}{2}$ and $s=r-\frac{1}{2}$ respectively.

Combining (24) and (25), and using the Schwarz inequality, we get

$$\begin{aligned}E\left[\left\|f_{D,\lambda}-f_\rho\right\|_\rho\right]\leq&\left\{E\left[\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(L_K+\lambda I\right)\right\|^2\right]\right\}^{\frac{1}{2}} \\ &\times\left\{E\left[\left\|\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^* y-L_{K,D(x)}f_\rho\right)\right\|_K^2\right]\right\}^{\frac{1}{2}} \\ &+\lambda^r\|g_\rho\|_\rho\left[E\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}\left(L_K+\lambda I\right)\right\|^r\right].\end{aligned} \tag{26}$$

By applying Lemma 4 and the identity $\|L_1L_2\|=\|L_2L_1\|$, we have

$$E\left[\left\|f_{D,\lambda}-f_\rho\right\|_\rho\right]\leq 30\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^2\frac{M}{\kappa}\mathcal{B}_{l,\lambda}+\left(4+\log 2\right)\lambda^r\|g_\rho\|_\rho\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2r}.$$

This completes the proof of Proposition 5. □

5 Proofs of Main Results

For the sake of simplicity, let $E_{\mathbf{z}^l}[\cdot]$ denote the expectation with respect to $\mathbf{z}^l:=\{z_i:=(\mu_{x_i},y_i)\}_{i=1}^l$ and let $E_{\mathbf{x}^{N,l}|\mathbf{z}^l}[\cdot]$ denote the conditional expectation with respect to $\left\{\{x_{i,j}\}_{j=1}^N\right\}_{i=1}^l$, given $z_1,\cdots,z_l$, that is

$$E_{\mathbf{z}^l}=E_{\{(\mu_{x_i},y_i)\}_{i=1}^l},\quad E_{\mathbf{x}^{N,l}|\mathbf{z}^l}=E_{\{\{x_{i,j}\}_{j=1}^N\}_{i=1}^l|\{z_i\}_{i=1}^l}. \tag{27}$$

The Hölder continuity is only applied to the first term of (13), and the second order decomposition approach is also crucial for the analysis of the first term. We have the following bound for the first term of (13).

Proposition 6. *Assume $|y| \leq M$ almost surely and the regularity condition (6) holds with some $\frac{1}{2} \leq r \leq 1$, and the mapping $K_{(\cdot)} : X_\mu \to, \mathcal{H}_K$ is (h, L) Hölder continuous with $h \in (0, 1]$ and $L > 0$. Then following estimate holds*

$$\begin{aligned}
&E\left[\left\|f_{\hat{D},\lambda} - f_{D,\lambda}\right\|_\rho\right] \\
&\leq \sqrt{3}(2+\sqrt{\pi})^{\frac{1}{2}} L \frac{2^{\frac{h}{2}} B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}} N^{\frac{h}{2}}} \left\{\frac{\kappa\sqrt{\mathcal{N}(\lambda)}}{\sqrt{l\lambda}} + 2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}} (2B_k)^{\frac{h}{2}} \frac{1}{\lambda N^{\frac{h}{2}}} + 1\right\} \\
&\times \left\{M + 50M\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 8\kappa\lambda^{r-\frac{1}{2}} \|g_\rho\|_\rho \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)^{2r-1} + 2\sqrt{3}\kappa^{r+\frac{1}{2}} \|g_\rho\|_\rho\right\}.
\end{aligned} \tag{28}$$

Before proving Proposition 6, we first give the bound for $\|f_{D,\lambda}\|_K$ in expectation.

Proposition 7. *Assume $|y| \leq M$ almost surely and the regularity condition (6) holds with some $\frac{1}{2} \leq r \leq 1$. Then we have*

$$\left\{E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K^2\right]\right\}^{\frac{1}{2}} \leq \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)^{2r-1} \left\{\frac{25M}{\kappa}\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}} + 1\right)^{2-2r} + 4\lambda^{r-\frac{1}{2}} \|g_\rho\|_\rho\right\} + \sqrt{3}\kappa^{r-\frac{1}{2}} \|g_\rho\|_\rho. \tag{29}$$

Proof. We first bound $\|f_{D,\lambda}\|_K$ by the following two terms

$$\|f_{D,\lambda}\|_K \leq \|f_{D,\lambda} - f_\rho\|_K + \|f_\rho\|_K. \tag{30}$$

For $\|f_{D,\lambda} - f_\rho\|_K$, by the definition of $f_{D,\lambda}$ and the regularity condition (6) with some $\frac{1}{2} \leq r \leq 1$, we have

$$\begin{aligned}
\|f_{D,\lambda} - f_\rho\|_K &= \left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} \frac{1}{l} S_D^* y - \left(L_{K,D(x)} + \lambda I\right)^{-1} \left(L_{K,D(x)} + \lambda I\right) f_\rho\right\|_K \\
&\leq \left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} \left(\frac{1}{l} S_D^* y - L_{K,D(x)} f_\rho\right)\right\|_K + \lambda\left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} L_K^r g_\rho\right\|_K.
\end{aligned} \tag{31}$$

For the first term of (31), we apply $\|L_1^s L_2^s\| \leq \|L_1 L_2\|^s$ to get

$$\begin{aligned}
&\left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} \left(\frac{1}{l} S_D^* y - L_{K,D(x)} f_\rho\right)\right\|_K \\
\leq &\left\|\left(L_{K,D(x)} + \lambda I\right)^{-\frac{1}{2}}\right\| \left\|\left(L_{K,D(x)} + \lambda I\right)^{-\frac{1}{2}} (L_K + \lambda I)^{\frac{1}{2}}\right\| \left\|(L_K + \lambda I)^{-\frac{1}{2}} \left(\frac{1}{l} S_D^* y - L_{K,D(x)} f_\rho\right)\right\|_K \\
\leq &\frac{1}{\sqrt{\lambda}} \left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} (L_K + \lambda I)\right\|^{\frac{1}{2}} \left\|(L_K + \lambda I)^{-\frac{1}{2}} \left(\frac{1}{l} S_D^* y - L_{K,D(x)} f_\rho\right)\right\|_K.
\end{aligned}$$

In the same way, for the second term of (31), we have

$$\begin{aligned}
&\lambda\left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} L_K^r g_\rho\right\|_K \\
\leq &\lambda\left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} \left(L_{K,D(x)} + \lambda I\right)^{r-\frac{1}{2}}\right\| \left\|\left(L_{K,D(x)} + \lambda I\right)^{-r+\frac{1}{2}} (L_K + \lambda I)^{r-\frac{1}{2}}\right\| \\
\times &\left\|(L_K + \lambda I)^{-r+\frac{1}{2}} L_K^{r-\frac{1}{2}}\right\| \left\|L_K^{\frac{1}{2}} g_\rho\right\|_K \\
\leq &\lambda^{r-\frac{1}{2}} \left\|\left(L_{K,D(x)} + \lambda I\right)^{-1} (L_K + \lambda I)\right\|^{r-\frac{1}{2}} \|g_\rho\|_\rho.
\end{aligned}$$

By combining the above bounds with (30) and (31), we have

$$\begin{aligned}
E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K^2\right] \leq & \frac{3}{\lambda}E_{\mathbf{z}^l}\left[\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}(L_K+\lambda I)\right\|\left\|(L_K+\lambda I)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^* y - L_{K,D(x)}f_\rho\right)\right\|_K^2\right] \\
& + 3\lambda^{2r-1}E_{\mathbf{z}^l}\left[\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}(L_K+\lambda I)\right\|^{2r-1}\|g_\rho\|_\rho^2\right] + 3\|f_\rho\|_K^2 \\
\leq & \frac{3}{\lambda}\left\{E_{\mathbf{z}^l}\left[\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}(L_K+\lambda I)\right\|^2\right]\right\}^{\frac{1}{2}}\left\{E_{\mathbf{z}^l}\left[\left\|(L_K+\lambda I)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^* y - L_{K,D(x)}f_\rho\right)\right\|_K^4\right]\right\}^{\frac{1}{2}} \\
& + 3\lambda^{2r-1}E_{\mathbf{z}^l}\left[\left\|\left(L_{K,D(x)}+\lambda I\right)^{-1}(L_K+\lambda I)\right\|^{2r-1}\right]\|g_\rho\|_\rho^2 + 3\|f_\rho\|_K^2.
\end{aligned}$$

Finally, the desired result holds by applying (20) with $d=2$ and $d=2r-1$ respectively and (21) with $d=4$

$$\begin{aligned}
E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K^2\right] \leq & \frac{\left(576+12\log^4 2\right)M^2}{\kappa^2}\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^2\frac{\mathcal{B}_{l,\lambda}^2}{\lambda} \\
& +(6\Gamma(4r-1)\lambda^{2r-1}+3\log^{4r-2}2)\|g_\rho\|_\rho^2\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{4r-2}+3\|f_\rho\|_K^2.
\end{aligned}$$

This completes the proof. □

Some notations and useful results are needed for proving Proposition 6. We define the empirical version of $L_{K,D(x)}$ by using the second stage sample $\hat{D}$ as follows

$$L_{K,\hat{D}(x)}(f)=\frac{1}{l}\sum_{i=1}^{l}f(\mu_{\hat{x}_i})K_{\mu_{\hat{x}_i}}=\frac{1}{l}\sum_{i=1}^{l}\left\langle f,K_{\mu_{\hat{x}_i}}\right\rangle_K K_{\mu_{\hat{x}_i}} \quad f\in\mathcal{H}_K. \tag{32}$$

Another sampling operator $\hat{S}_D:\mathcal{H}_K\to\mathbb{R}^l$ associated with the second stage sample is defined as follows

$$\hat{S}_D f=\left(f(\mu_{\hat{x}_1}),\cdots,f(\mu_{\hat{x}_l})\right)^T, \quad f\in\mathcal{H}_K, \tag{33}$$

and its dual operator $\hat{S}_D^*:\mathbb{R}^l\to\mathcal{H}_K$ is given by

$$\hat{S}_D^* c=\sum_{i=1}^{l}c_i K_{\mu_{\hat{x}_i}}, \quad c\in\mathbb{R}^l. \tag{34}$$

For each $i\in\{1,\cdots,l\}$, the difference $\mu_{\hat{x}_i}-\mu_{x_i}\in H$ can be estimated by considering the random variable ζ on $\left(\tilde{X},x_i\right)$ with values in H given by $\zeta(s)=k(\cdot,s)$ for $s\in\tilde{X}$ which has mean μ_{x_i} and empirical mean $\mu_{\hat{x}_i}$ and satisfies $\|\zeta\|_H\leq\sqrt{B_k}$ almost surely. In fact, from Section A.1.10 in [23], we know that for each $1\leq i\leq l$, $\mathbb{P}_{\{x_{i,j}\}_{j=1}^N|x_i}\left(\|\mu_{\hat{x}_i}-\mu_{x_i}\|_H\leq\frac{\sqrt{2B_k}}{\sqrt{N}}+\epsilon\right)\geq 1-e^{-\frac{\epsilon^2 N}{2B_k}}$, or

$$\|\mu_{\hat{x}_i}-\mu_{x_i}\|_H\leq\frac{\sqrt{2B_k}}{\sqrt{N}}+\frac{\sqrt{2\theta B_k}}{\sqrt{N}}=\frac{(1+\sqrt{\theta})\sqrt{2B_k}}{\sqrt{N}} \tag{35}$$

with probability at least $1-e^{-\theta}$. Now we let $\xi=\|\mu_{\hat{x}_i}-\mu_{x_i}\|_H^{2h}$ with $h\in(0,1]$ which satisfies $0\leq\xi\leq 2^{2h}B_k^h$ almost surely. By applying the formula $E[\xi]=\int_0^\infty\mathbb{P}_{\{x_{i,j}\}_{j=1}^N|x_i}(\xi>t)\,dt=\int_0^{2^{2h}B_k^h}\mathbb{P}_{\{x_{i,j}\}_{j=1}^N|x_i}(\xi>t)\,dt=\int_0^{2^{2h}B_k^h}\mathbb{P}_{\{x_{i,j}\}_{j=1}^N|x_i}\left(\|\mu_{\hat{x}_i}-\mu_{x_i}\|_H>t^{\frac{1}{2h}}\right)dt$, we can bound the probability by 1 on the interval $\left[0,\left(\frac{2B_k}{N}\right)^h\right]$ and make a change of variable $t=\left(\frac{2B_k}{N}\right)^h\left(1+\sqrt{\theta}\right)^{2h}$ on the interval $\left(\left(\frac{2B_k}{N}\right)^h,2^{2h}B_k^h\right]$ to get the estimate

$$E_{\{x_{i,j}\}_{j=1}^N|x_i}\left[\|\mu_{\hat{x}_i}-\mu_{x_i}\|_H^{2h}\right]\leq\left(2+\sqrt{\pi}\right)\frac{2^h B_k^h}{N^h}. \tag{36}$$

Further by applying equation (36) under the assumption of (h, L) Hölder continuity of $K(\cdot)$, we have the following estimation,

$$\left\{E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|\frac{1}{l}\hat{S}_D^* y-\frac{1}{l}S_D^* y\right\|_K^2\right]\right\}^{\frac{1}{2}} \leq \left(2+\sqrt{\pi}\right)^{\frac{1}{2}} LM\frac{2^{\frac{h}{2}}B_k^{\frac{h}{2}}}{N^{\frac{h}{2}}}. \tag{37}$$

Also, combining the result in Section A.1.11 in [23] with (36), we can derive

$$\left\{E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_{K,D(x)}-L_{K,\hat{D}(x)}\right\|^2\right]\right\}^{\frac{1}{2}} \leq \kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\frac{2^{\frac{h+2}{2}}B_k^{\frac{h}{2}}}{N^{\frac{h}{2}}}. \tag{38}$$

The following result is also key for the proof of Proposition 6.

Lemma 8. *Let D be a sample drawn independently according to ρ and $\{x_{i,j}\}_{j=1}^N$ be a sample drawn independently according to x_i for $i=1,\cdots,l$ respectively, and the mapping $K_{(\cdot)}: X_\mu \to \mathcal{H}_K$ is (h,L) Hölder continuous with $h\in(0,1]$ and $L>0$. Then we have*

$$\left\{E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|^2\right]\right\}^{\frac{1}{2}} \leq \sqrt{3}\left\{\frac{\Xi_D}{\lambda}+2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\left(2B_k\right)^{\frac{h}{2}}\frac{1}{\lambda^{\frac{3}{2}}N^{\frac{h}{2}}}+\frac{1}{\sqrt{\lambda}}\right\}, \tag{39}$$

where Ξ_D is a quantity given by

$$\Xi_D=\left\|\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(L_K-L_{K,D(x)}\right)\right\|.$$

Proof. First we divide $\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|$ into two terms

$$\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\| \leq \left\|L_K^{\frac{1}{2}}\left\{\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}-\left(L_K+\lambda I\right)^{-1}\right\}\right\|+\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-1}\right\|. \tag{40}$$

For the first term in (40), we have the decomposition

$$\begin{aligned}&\left\|L_K^{\frac{1}{2}}\left\{\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}-\left(L_K+\lambda I\right)^{-1}\right\}\right\|\\ =&\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-1}\left(L_K-L_{K,D(x)}+L_{K,D(x)}-L_{K,\hat{D}(x)}\right)\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|.\end{aligned}$$

By the bounds $\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-\frac{1}{2}}\right\|\leq 1$, $\left\|\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|\leq\frac{1}{\lambda}$ and $\left\|\left(L_K+\lambda I\right)^{-\frac{1}{2}}\right\|\leq\frac{1}{\sqrt{\lambda}}$, we have

$$\left\|L_K^{\frac{1}{2}}\left\{\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}-\left(L_K+\lambda I\right)^{-1}\right\}\right\| \leq \frac{1}{\lambda}\left\|\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(L_K-L_{K,D(x)}\right)\right\|+\frac{1}{\lambda^{\frac{3}{2}}}\left\|L_{K,D(x)}-L_{K,\hat{D}(x)}\right\|.$$

Putting the above estimate back into (40), and by (38), we have

$$\begin{aligned}E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|^2\right] &\leq \frac{3\Xi_D^2}{\lambda^2}+\frac{3}{\lambda^3}E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_{K,D(x)}-L_{K,\hat{D}(x)}\right\|^2\right]+3\left\|L_K^{\frac{1}{2}}\left(L_K+\lambda I\right)^{-1}\right\|^2\\ &\leq \frac{3\Xi_D^2}{\lambda^2}+12\kappa^2L^2\left(2+\sqrt{\pi}\right)2^hB_k^h\frac{1}{\lambda^3N^h}+\frac{3}{\lambda}.\end{aligned}$$

This completes the proof. □

Now we are ready to prove Proposition 6.

Proof of Proposition 6. First we use the explicit forms of $f_{\hat{D},\lambda}$ and $f_{D,\lambda}$ to express the difference as

$$\begin{aligned}f_{\hat{D},\lambda}-f_{D,\lambda} &= \left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\frac{1}{l}\hat{S}_D^*y-\left(L_{K,D(x)}+\lambda I\right)^{-1}\frac{1}{l}S_D^*y\\ &= \left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}\hat{S}_D^*y-\frac{1}{l}S_D^*y\right)+\left\{\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}-\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\}\frac{1}{l}S_D^*y\\ &= \left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}\hat{S}_D^*y-\frac{1}{l}S_D^*y\right)+\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}-L_{K,\hat{D}(x)}\right)f_{D,\lambda},\end{aligned}$$

the last equality holds due to the identities

$$\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}-\left(L_{K,D(x)}+\lambda I\right)^{-1}=\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}-L_{K,\hat{D}(x)}\right)\left(L_{K,D(x)}+\lambda I\right)^{-1}$$

and $f_{D,\lambda}=\left(L_{K,D(x)}+\lambda I\right)^{-1}\frac{1}{l}S_D^*y$.

Since for $g\in L^2_{\rho_{X_\mu}}$, we have $\|g\|_\rho=\left\|L_K^{\frac{1}{2}}g\right\|_K$, it follows that

$$\begin{aligned}\left\|f_{\hat{D},\lambda}-f_{D,\lambda}\right\|_\rho &= \left\|L_K^{\frac{1}{2}}\left(f_{\hat{D},\lambda}-f_{D,\lambda}\right)\right\|_K\\ &\leq \left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}\hat{S}_D^*y-\frac{1}{l}S_D^*y\right)\right\|_K\\ &+ \left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}-L_{K,\hat{D}(x)}\right)f_{D,\lambda}\right\|_K. \end{aligned} \tag{41}$$

Now we estimate the two terms of (41). For the first term, by using the Schwarz inequality

$$\begin{aligned}&E\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}\hat{S}_D^*y-\frac{1}{l}S_D^*y\right)\right\|_K\right]\\ &\leq\left\{E\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|^2\right]\right\}^{\frac{1}{2}}\left\{E\left[\left\|\frac{1}{l}\hat{S}_D^*y-\frac{1}{l}S_D^*y\right\|_K^2\right]\right\}^{\frac{1}{2}}.\end{aligned}$$

Further by applying Lemma 8, and (37), we have

$$\begin{aligned}&E\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(\frac{1}{l}\hat{S}_D^*y-\frac{1}{l}S_D^*y\right)\right\|_K\right]\\ \leq\ &\sqrt{3}E_{\mathbf{z}^l}\left[\frac{\Xi_D}{\lambda}+2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\left(2B_k\right)^{\frac{h}{2}}\frac{1}{\lambda^{\frac{3}{2}}N^{\frac{h}{2}}}+\frac{1}{\sqrt{\lambda}}\right]\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}LM\frac{2^{\frac{h}{2}}B_k^{\frac{h}{2}}}{N^{\frac{h}{2}}}\\ \leq\ &\sqrt{3}\left\{\frac{\kappa\sqrt{\mathcal{N}(\lambda)}}{\sqrt{l\lambda}}+2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\left(2B_k\right)^{\frac{h}{2}}\frac{1}{\lambda N^{\frac{h}{2}}}+1\right\}\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}LM\frac{2^{\frac{h}{2}}B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}}N^{\frac{h}{2}}},\end{aligned} \tag{42}$$

where we have used the Schwarz inequality $E_{\mathbf{z}^l}\left[\Xi_D\right]\leq\sqrt{E_{\mathbf{z}^l}\left[\Xi_D^2\right]}$ and the following bound from [18]

$$E_{\mathbf{z}^l}\left[\Xi_D^2\right]\leq\frac{\kappa^2\mathcal{N}(\lambda)}{l}.$$

For the second term of (41), it is easy to derive

$$\begin{aligned}&E\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}-L_{K,\hat{D}(x)}\right)f_{D,\lambda}\right\|_K\right]\\ \leq &E_{\mathbf{z}^l}\left[E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|\left\|L_{K,D(x)}-L_{K,\hat{D}(x)}\right\|\right]\|f_{D,\lambda}\|_K\right]\\ \leq &E_{\mathbf{z}^l}\left[\left\{E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\right\|^2\right]\right\}^{\frac{1}{2}}\left\{E_{\mathbf{x}^{N,l}|\mathbf{z}^l}\left[\left\|L_{K,D(x)}-L_{K,\hat{D}(x)}\right\|^2\right]\right\}^{\frac{1}{2}}\|f_{D,\lambda}\|_K\right].\end{aligned}$$

By Lemma 8, (38) and the Schwarz inequality, we get

$$\begin{aligned}&E\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}-L_{K,\hat{D}(x)}\right)f_{D,\lambda}\right\|_K\right]\\ \leq&\sqrt{3}E_{\mathbf{z}^l}\left[\left(\frac{\Xi_D}{\lambda}+2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\left(2B_k\right)^{\frac{h}{2}}\frac{1}{\lambda^{\frac{3}{2}}N^{\frac{h}{2}}}+\frac{1}{\sqrt{\lambda}}\right)\left(\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\frac{2^{\frac{h+2}{2}}B_k^{\frac{h}{2}}}{N^{\frac{h}{2}}}\right)\|f_{D,\lambda}\|_K\right]\\ \leq&\sqrt{3}\left(\left\{E_{\mathbf{z}^l}\left[\frac{\Xi_D^2}{\lambda}\right]\right\}^{\frac{1}{2}}\left\{E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K^2\right]\right\}^{\frac{1}{2}}+E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K\right]\left(2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\left(2B_k\right)^{\frac{h}{2}}\frac{1}{\lambda N^{\frac{h}{2}}}+1\right)\right)\\ &\times\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\frac{2^{\frac{h+2}{2}}B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}}N^{\frac{h}{2}}}.\end{aligned}$$

But $E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K\right] \leq \left\{E_{\mathbf{z}^l}\left[\|f_{D,\lambda}\|_K^2\right]\right\}^{\frac{1}{2}}$ and $E_{\mathbf{z}^l}\left[\Xi_D^2\right] \leq \frac{\kappa^2\mathcal{N}(\lambda)}{l}$. So we can apply Proposition 7 to get

$$\begin{aligned}
&E\left[\left\|L_K^{\frac{1}{2}}\left(L_{K,\hat{D}(x)}+\lambda I\right)^{-1}\left(L_{K,D(x)}-L_{K,\hat{D}(x)}\right)f_{D,\lambda}\right\|_K\right] \\
\leq&\sqrt{3}\left\{\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)\frac{25M}{\kappa}\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+4\lambda^{r-\frac{1}{2}}\|g_\rho\|_\rho\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2r-1}+\sqrt{3}\kappa^{r-\frac{1}{2}}\|g_\rho\|_\rho\right\} \qquad (43)\\
&\times\left(\frac{\kappa\sqrt{\mathcal{N}(\lambda)}}{\sqrt{l\lambda}}+2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}(2B_k)^{\frac{h}{2}}\frac{1}{\lambda N^{\frac{h}{2}}}+1\right)\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}\frac{2^{\frac{h+2}{2}}B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}}N^{\frac{h}{2}}}.
\end{aligned}$$

We complete the proof of Proposition 6 by combining this bound with (42). □

Now we are in a position to prove the main results. Theorem 1 follows immediately from Propositions 5 and 6 and (13). □

Proof of Theorem 2. From our choice of $N = l^{\frac{2\alpha+4\alpha r}{h(4\alpha r+1)}}$, $\lambda = l^{-\frac{2\alpha}{4\alpha r+1}}$ we have the bounds

$$\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{l\lambda}} \leq \frac{\sqrt{c}\lambda^{-\frac{1}{4\alpha}}}{\sqrt{l\lambda}} = \sqrt{c}l^{-\frac{1}{2}}\lambda^{-\frac{1}{4\alpha}-\frac{1}{2}} = \sqrt{c}l^{\frac{\alpha(1-2r)}{4\alpha r+1}} \leq \sqrt{c},$$

$$\frac{1}{\lambda N^{\frac{h}{2}}} = l^{\frac{2\alpha}{4\alpha r+1}}l^{-\frac{\alpha(1+2r)}{4\alpha r+1}} \leq 1,$$

and

$$\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1 = \frac{2\kappa}{\sqrt{l\lambda}}\left(\frac{\kappa}{\sqrt{l\lambda}}+\sqrt{\mathcal{N}(\lambda)}\right)+1 \leq 2\kappa\left(\frac{\kappa}{l\lambda}+\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{l\lambda}}\right)+1 \leq 2\kappa\left(\kappa+\sqrt{c}\right)+1.$$

Putting the above bounds back into (7) in Theorem 1, we get

$$E\left[\left\|f_{\hat{D},\lambda}-f_\rho\right\|_\rho\right] \leq Cl^{-\frac{2\alpha r}{4\alpha r+1}}, \qquad (44)$$

where C is the constant independent of l or N given by

$$\begin{aligned}
C =&\sqrt{3}(2+\sqrt{\pi})^{\frac{1}{2}}L(2B_k)^{\frac{h}{2}}\left\{\kappa\sqrt{c}+2\kappa L\left(2+\sqrt{\pi}\right)^{\frac{1}{2}}(2B_k)^{\frac{h}{2}}+1\right\}\\
&\times\left\{M+100M\left(2\kappa\left(\kappa+\sqrt{c}\right)+1\right)\kappa\left(\kappa+\sqrt{c}\right)+8\kappa\|g_\rho\|_\rho\left(2\kappa\left(\kappa+\sqrt{c}\right)+1\right)^{2r-1}+2\sqrt{3}\kappa^{r+\frac{1}{2}}\|g_\rho\|_\rho\right\}\\
&+\left(60M\left(\kappa+\sqrt{c}\right)+(4+\log 2)\|g_\rho\|_\rho\right)\left(2\kappa\left(\kappa+\sqrt{c}\right)+1\right)^2.
\end{aligned}$$

□

Acknowledgements. The work described in this paper is partially supported by the Research Grants Council of Hong Kong [Project No. CityU 11338616] and by National Natural Science Foundation of China [Projects No. 11531013 and No. 11671307].

Appendix

Proof of Lemma 4. We first prove (20). By Proposition 4.4 in [16] (see the same bound with an additional factor 2 in [15]), we know that for $\lambda > 0$ and $0 < \delta < 1$

$$\mathbb{P}\left(\left\|(L_K+\lambda I)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\| \leq \left(\frac{\mathcal{B}_{l,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}+1\right)^2\right) \geq 1-\delta, \qquad (45)$$

which implies for $0 < \delta < 2$

$$\mathbb{P}\left(\left\|(L_K+\lambda I)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\| \leq \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^2\log^2\frac{4}{\delta}\right) \geq 1-\frac{\delta}{2}.$$

Define a random variable $\xi = \left\|(L_K+\lambda I)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\|^d$ and make a variable change in the above probability bound $t = \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}\log^{2d}\frac{4}{\delta}$ with $d > 0$, we know that for $t > \left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}\log^{2d}2$

$$\mathbb{P}\left(\xi > t\right) = \mathbb{P}\left(\xi^{\frac{1}{d}} > t^{\frac{1}{d}}\right) \leq \frac{\delta}{2} = 2\exp\left\{-\frac{t^{\frac{1}{2d}}}{\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1}\right\}.$$

Now we apply the formula

$$E\left[\xi\right]=\int_0^\infty \mathbb{P}\left(\xi>t\right)dt$$

to estimate the expectation of $\left\|\left(L_K+\lambda I\right)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\|^d$ as

$$\begin{aligned}E\left[\xi\right]&\leq\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}\log^{2d}2+\int_{\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}\log^{2d}2}^{\infty}2\exp\left\{-\frac{t^{\frac{1}{2d}}}{\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1}\right\}dt\\&\leq\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}\log^{2d}2+4d\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}\int_{\log 2}^{\infty}e^{-x}x^{2d-1}dx\leq\left(2\Gamma(2d+1)+\log^{2d}2\right)\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}.\end{aligned}$$

Thus we know that

$$E\left[\left\|\left(L_K+\lambda I\right)\left(L_{K,D(x)}+\lambda I\right)^{-1}\right\|^d\right]\leq\left(2\Gamma(2d+1)+\log^{2d}2\right)\left(\frac{\mathcal{B}_{l,\lambda}}{\sqrt{\lambda}}+1\right)^{2d}.$$

This proves (20). □

To see (21), we apply Lemma 3 in [15] which asserts that for $0<\delta<1$

$$\mathbb{P}\left(\left\|\left(L_K+\lambda I\right)^{-\frac{1}{2}}\left(\frac{1}{l}S_D^*y-L_{K,D(x)}f_\rho\right)\right\|\leq\frac{2M}{\kappa}\mathcal{B}_{l,\lambda}\log\frac{2}{\delta}\right)\geq 1-\delta.$$

Then we can prove (21) by the similar technique as in the proof of (20).

References

[1] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. Journal of Complexity, 23:52–72, 2007.

[2] Alain Berlinet and Christine Thomas-Agnan. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer, 2004.

[3] G. Blanchard and N. Krämer, Convergence rates of kernel conjugate gradient for random design regression, Analysis and Applications:14(6):763–794, 2016.

[4] Andrea Caponnetto and Ernesto De Vito, Optimal rates for the regularized least-squares algorithm, Foundations of Computational Mathematics, 7(3):331-368, 2007.

[5] Nello Cristianini and John Shawe-Taylor. An Introduction to Support Vector Machines, Cambridge University Press, 2000.

[6] Felipe Cucker and Ding-Xuan Zhou. Learning Theory: An Approximation Theory Viewpoint. Cambridge University Press, 2007.

[7] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 89(1997): 31-71.

[8] Daniel R. Dooly, Qi Zhang, Sally A. Goldman, and Robert A. Amar. Multiple-instance learning of real-valued data. Journal of Machine Learning Research, 3(2002): 651–678.

[9] Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. Advances in computational mathematics, 13(1): 1–50, 2000.

[10] Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. Journal of Machine Learning Research, 5(2004):73–99.

[11] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. Advances in Neural Information Processing Systems 20, pages 489-496, 2008.

[12] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two sample problem. Advances in Neural Information Processing Systems 19, pages 513-520. MIT Press, 2007.

[13] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. Journal of Machine Learning Research, 13(2012):723–773.

[14] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. Advances in Neural Information Processing Systems 20, pages 585-592. MIT Press, 2008.

[15] Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou, Learning theory of distributed spectral algorithms, Inverse Problems, 33(7), 074009, 2017.

[16] Zheng-Chu Guo, Lei Shi and Qiang Wu, Learning theory of distributed regression with bias corrected regularization kernel network, Journal of Machine Learning Research, 18(2017): 4237–4261.

[17] Zheng-Chu Guo, Dao-Hong Xiang, Xin Guo, and Ding-Xuan Zhou, Thresholded spectral algorithms for sparse approximations, Analysis and Applications, 15(3):433-455, 2017.

[18] Shaobo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. Journal of Machine Learning Research 18(2017): 3202–3232.

[19] Barnab*ás* P*ó*czos, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Distribution-free distribution regression. International Conference on Artificial Intelligence and Statistics (AISTATS), 31:507C515, 2013.

[20] Soumya Ray and David Page. Multiple instance regression. In International Conference on Machine Learning (ICML), 425-432, 2001.

[21] Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. in Proceedings of the 22nd Annual Conference on Learning Theory (S. Dasgupta and A. Klivans, eds.), pp. 79-93, 2009.

[22] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. Constructive approximation, 26(2): 153-172 , 2007.

[23] Zoltán Szabó, Arthur Gretton, Barnabas Poczos, and Bharath K. Sriperumbudur. Two-stage sampled learning theory on distributions. In International Conference on Artificial Intelligence and Statistics (AISTATS), pages 948-957, San Diego, California, USA, 9-12 May 2015.

[24] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabas Poczos, and Arthur Gretton, Learning theory for distribution regression, Journal of Machine Learning Research, 17(2016):1–40.

[25] Ding-Xuan Zhou, Deep distributed convolutional neural networks: universality, Analysis and Applications. 16(2018): 895–919.

[26] Ding-Xuan Zhou, Universality of deep convolutional neural networks, Applied and Computational Harmonic Analysis, in press.