Robust Kernel-based Distribution Regression

Zhan Yu, Daniel W. C. Ho

Department of Mathematics, City University of Hong Kong
Kowloon, Hong Kong, Email: zhanyu2-c@my.cityu.edu.hk; madaniel@cityu.edu.hk

Zhongjie Shi

School of Data Science, City University of Hong Kong
Kowloon, Hong Kong, Email: zhongjshi2-c@my.cityu.edu.hk

Ding-Xuan Zhou

School of Data Science, Department of Mathematics, Liu Bie Ju Centre for Mathematical Sciences
City University of Hong Kong, Kowloon, Hong Kong, Email: mazhou@cityu.edu.hk

Abstract

Regularization schemes for regression have been widely studied in learning theory and inverse problems. In this paper, we study regularized distribution regression (DR) which involves two stages of sampling, and aims at regressing from probability measures to real-valued responses by regularization over a reproducing kernel Hilbert space (RKHS). Many important tasks in statistical learning and inverse problems can be treated in this framework. Examples include multi-instance learning and point estimation for problems without analytical solutions. Recently, theoretical analysis on DR has been carried out via kernel ridge regression and several interesting learning behaviors have been observed. However, the topic has not been explored and understood beyond the least squares based DR. By introducing a robust loss function l_σ for two-stage sampling problems, we present a novel robust distribution regression (RDR) scheme. With a windowing function V and a scaling parameter σ which can be appropriately chosen, l_σ can include a wide range of commonly used loss functions that enrich the theme of DR. Moreover, the loss l_σ is not necessarily convex, which enlarges the regression class (least squares) in the literature of DR. Learning rates in different regularity ranges of the regression function are comprehensively studied and derived via integral operator techniques. The scaling parameter σ is shown to be crucial in providing robustness and satisfactory learning rates of RDR.

Keywords: learning theory, distribution regression, robust regression, integral operator, learning rate

1 Introduction

Data from many practical applications often appear in forms of functionals or matrices. Such types of data impose difficulty in applying classical regression methods used for dealing with vector-valued data. Hence, developing suitable regression schemes for solving the corresponding problems becomes desirable. Recently, *distribution regression* (DR) was introduced to handle complicated data from

some Banach spaces ([28, 30, 31, 32, 12]). Many important tasks in machine learning, statistics, and inverse problems can be analysed in the framework of DR. One example is the multi-instance learning problem ([8, 29]) in which each instance is generated from a probability distribution in an independent identically distributed manner. In statistics, some tasks might be stated as point estimation problems for probability distributions without analytical expressions ([31]).

In DR, the input data are (probability) distributions on a compact metric space $\bar{X}$. In the first stage, we have a data set $\bar{D} = \{(x_i, y_i)\}_{i=1}^{|\bar{D}|} \subset X \times Y$, in which $|\bar{D}|$ is the cardinality of $\bar{D}$ and each pair (x_i, y_i) is *i.i.d.* sampled from a meta distribution over $X \times Y$, X is the input space of probability distributions on $\bar{X}$, and $Y = \mathbb{R}$ is the output space equipped with the standard Euclidean metric. Generally, the distributions $\{x_i\}$ cannot be observed directly. On the way of learning the regressor from X to Y, we can observe a second-stage sample drawn from the probability measures. This is done in the second stage of DR where the samples in the sample set $\hat{D} = \left\{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\right\}_{i=1}^{|\bar{D}|}$ are obtained by drawing a sample $\{x_{i,s}\}_{s=1}^{d_i} \subset \bar{X}$ according to the probability distribution x_i on $\bar{X}$.

We borrow an example on medical applications of DR from [31] to illustrate ideas of the above two-stage sampling. Here, the set X is treated as a pool of patients identified with a set of probability distributions on $\bar{X} = [0, 1]$. The ith patient x_i in the sample set $\{x_i\}_{i=1}^{|\bar{D}|}$ can be periodically assessed by blood tests $\{x_{i,s}\}_{s=1}^{d_i}$ which are made at moments $\{j/d_i\}_{j=1}^{d_i}$. Then $\{x_{i,s}\}_{s=1}^{d_i}$ is exactly the second stage sample set associated with x_i, and $\{y_i\}_{i=1}^{|\bar{D}|}$ are the values of some health indicator of the patients. The goal of DR is to learn a mapping from the set of blood tests to the health indicator values by observations on a group of patients. From the perspective of learning, we hope that by observing a large number of patients and making enough tests (with large d_i), the learned mapping can be precise enough.

The work in this paper is based on a kernel mean embedding ridge regression method for DR ([2]). Let $(\mathcal{H}_k, \|\cdot\|_k)$ be a reproducing kernel Hilbert space (RKHS) with the associated reproducing kernel $k : \bar{X} \times \bar{X} \to \mathbb{R}$. Let $(\bar{X}, \mathcal{F})$ be a measurable space with $\mathcal{F}$ being a Borel σ-algebra on $\bar{X}$. Denote the set of Borel probability measures on $(\bar{X}, \mathcal{F})$ by $\mathcal{M}_1(\mathcal{F})$. Then the *kernel mean embedding* of a distribution $x \in \mathcal{M}_1(\mathcal{F})$ to an element μ_x of RKHS $\mathcal{H}_k$ is given by

$$\mu_x = \int_{\bar{X}} k(\cdot, \eta) dx(\eta).$$

Via the kernel mean embedding, kernel methods for handling vector-valued data can be extended to those with values of probability distributions. The kernel mean embedding transformation $x \mapsto \mu_x$ is injective when k is a characteristic kernel ([18, 12]). The injectivity is shown to be useful in statistical applications (e.g. [13, 19]). Denote the set of the mean embeddings by $X_\mu = \{\mu_x : x \in \mathcal{M}_1(\mathcal{F})\} \subseteq \mathcal{H}_k$. Then the mean embeddings of $\bar{D}$ to X_μ can be represented by

$$D = \{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}.$$

Let ρ be a probability measure on the product space $Z = X_\mu \times Y$. The aim of DR is to predict the conditional mean for given X_μ by learning the regression function $f_\rho : X_\mu \to Y$ defined by

$$f_\rho(\mu_x) = \int_Y y d\rho(y|\mu_x), \qquad \mu_x \in X_\mu,$$

where $\rho(\cdot|\mu_x)$ is the conditional probability measure of ρ at $\mu_x \in X_\mu$. Note that f_ρ is just the minimizer of the least squares generalization error

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

Generally, the measure ρ is unknown, and learning f_ρ is carried out in a non-parametric setting by implementing some learning algorithms over the sample D obtained in a *one-stage sampling* process. In DR, the first stage sample $\{x_i\}_{i=1}^{|D|}$ of probability distributions is still unobservable. Instead, each probability distribution x_i is approximately available via a second stage random sample $\{x_{i,s}\}_{s=1}^{d_i} \subset \bar{X}$. So the goal of DR is to learn the regression function f_ρ from the sample

$$\hat{D} = \big\{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\big\}_{i=1}^{|D|}$$

obtained in a *two-stage sampling* process. We study a kernel-based method for DR. Consider a reproducing kernel Hilbert space $(\mathcal{H}_K, \|\cdot\|_K)$ associated with a Mercer kernel $K : X_\mu \times X_\mu \to \mathbb{R}$. As an extension of the classical kernel ridge regression scheme [1, 3, 6, 20, 33, 34, 35, 37], the regularized least squares DR scheme takes a Tikhonov regularization form [1, 12, 31] as

$$f_{\hat{D},\lambda}^{ls} = \arg\min_{f\in\mathcal{H}_K} \Big\{ \frac{1}{|D|}\sum_{i=1}^{|D|} \big(f(\mu_{\hat{x}_i}) - y_i\big)^2 + \lambda\|f\|_K^2 \Big\}, \tag{1.1}$$

in which $\hat{x}_i = \frac{1}{d_i}\sum_{s=1}^{d_i}\delta_{x_{i,s}}$ serves as the empirical distribution determined by the observable set $\hat{D} = \big\{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\big\}_{i=1}^{|D|}$,

$$\mu_{\hat{x}_i} = \frac{1}{d_i}\sum_{s=1}^{d_i} k(\cdot, x_{i,s})$$

is the corresponding kernel mean embedding, and $\lambda > 0$ is a regularization parameter. The least squares minimization problem (1.1) can be regarded ([1, 9]) as a Tikhonov regularization solution to an ill-posed inverse problem with noisy data $\hat{D} = \big\{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\big\}_{i=1}^{|D|}$.

In this paper, we investigate a more general framework of two-stage distribution regression by considering a novel regularized robust DR (RDR) scheme

$$f_{\hat{D},\lambda}^{\sigma} = \arg\min_{f\in\mathcal{H}_K} \Big\{ \frac{\sigma^2}{|D|}\sum_{i=1}^{|D|} V\Big(\frac{[f(\mu_{\hat{x}_i}) - y_i]^2}{\sigma^2}\Big) + \lambda\|f\|_K^2 \Big\}, \tag{1.2}$$

where $V : \mathbb{R}_+ \to \mathbb{R}$ is a windowing function and $\sigma > 0$ is a scaling parameter. The algorithm can also be written in the form

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

with a loss function $l_\sigma : \mathbb{R} \to \mathbb{R}$ given by $l_\sigma(u) = \sigma^2 V(\frac{u^2}{\sigma^2})$. It can be witnessed that, in regression strategy (1.2), to enhance robustness of DR, we have replaced the least squares loss by a more general robust alternative generated by the windowing function V and scaling parameter σ. By selecting appropriate windowing function V and scaling parameter σ, the loss function yields a wide range of important RDR classes, which is new in the literature of DR. For example, Welsch loss $l_\sigma(u) = \sigma^2\big[1-\exp(-\frac{u^2}{2\sigma^2})\big]$ has been shown to be powerful in various settings with one-stage sampling such as signal processing, data clustering, pattern recognition, and non-parameter regression. From a perspective of information-theoretic learning, Welsch loss can be induced by the well-known correntropy loss, which was first introduced in [26] based on entropies. The correntropy between two scaler random variables U and V is defined as $\mathbb{E}K_\sigma(U,V)$ with K_σ a Gaussian kernel given by $K_\sigma(u,v) = \exp\{-(u-v)^2/\sigma^2\}$ with the scaler $\sigma > 0$. Entropy-based losses mainly include the loss induced by maximum correntropy criterion [14] and that by minimum error entropy criterion

[24, 39]. In addition, the range of l_σ also includes many commonly used loss functions such as the Huber loss [17] and pinball loss [36]. Recall that the traditional least squares DR scheme is the most popular DR method in the literature. It relies only on the mean squared error and belongs to the second-order statistics. Also recall that the least squares regression is optimal for Gaussian noise but suboptimal for non-Gaussian noise. In practice, samples are often contaminated by non-Gaussian noise or outliers. Moreover, least squares estimators for regression models are highly sensitive to outliers, and when the noise is not Gaussian, they often have poor performances. Unfortunately, in the existing literature of two-stage DR, approaches and theoretical studies are still limited to the least squares scheme, no other mainstream regression methods have been proposed in non-Gaussian settings yet. These facts motivate us to consider the proposed RDR scheme in (1.2) to fill the gap when tackling two-stage DR. Because of the robustness to non-Gaussian noise or outliers, the proposed RDR is expected to be applicable in practice. Some numerical experiments are conducted in Section ???.

The goal of this paper is to investigate RDR in a framework of learning theory. To derive learning rates of the estimator $f^\sigma_{\hat{D},\lambda}$ when approximating f_ρ and investigate the related robustness, we use a kernel based integral operator technique as a main tool. Via kernel mean embedding, we learn the regression function f_ρ with algorithm (1.2) from the given training sample $\hat{D} = \left\{(\{x_{i,j}\}_{j=1}^{d_i}, y_i)\right\}_{i=1}^{|D|}$ with $\{x_{i,1}, x_{i,2}, \ldots, x_{i,d_i}\}$ drawn independently from x_i. Novel theoretical results on robust estimator $f^\sigma_{\hat{D},\lambda}$ are derived. Note that, in the proposed RDR, the loss function l_σ may involve non-convex functions (e.g. Welsch loss), hence the theoretical study on RDR is essentially different from those on existing DR methods.

We summarize our main contributions of the work as follows.

- We propose a novel RDR method for two-stage sampling DR. Learning theory analysis is carried out for the estimator $f^\sigma_{\hat{D},\lambda}$ given by (1.2). Novel error bounds are derived with integral operator techniques. With the introduction of the flexibly chosen windowing function V and scaling parameter σ that leads to a wide range of commonly used robust losses, the existing analysis and algorithms in the literature of DR (least squares) have been largely improved.
- The learning behaviors of RDR are comprehensively explored for regularity index r (introduced below) in the whole range of $(0, \infty)$. Accordingly, satisfactory convergence rates in terms of the sample size $|D|$ are derived. We also show that the optimal mini-max learning rates can be achieved by RDR under appropriate conditions.
- The significance of σ in providing robustness and fast learning rates of RDR are shown in our analysis and main results.

2 Main Results

We assume throughout the paper that there exists a constant $M > 0$ such that $|y| \leq M$ almost surely, and k and K are bounded Mercer (symmetric, continuous, positive semidefinite) kernels with bounds B_k and B_K:

$$B_k = \sup_{v \in \bar{X}} k(v, v) < \infty, \ B_K = \sup_{\mu_u \in X_\mu} K(\mu_u, \mu_u) < \infty. \tag{2.1}$$

Denote the Banach space of bounded linear operators from space $Y = \mathbb{R}$ to $\mathcal{H}_K$ by $\mathcal{L}(Y, \mathcal{H}_K)$. Set $K_{\mu_x} = K(\mu_x, \cdot)$ for $\mu_x \in X_\mu$. We treat K_{μ_x} as an element of $\mathcal{L}(Y, \mathcal{H}_K)$ by defining the linear mapping

$$K_{\mu_x}(y) = yK_{\mu_x}, \qquad y \in Y.$$

The mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is assumed to be (α, L)-Hölder continuous for some $\alpha \in (0, 1]$ and $L > 0$ in the sense that

$$\big\|K_{\mu_x} - K_{\mu_y}\big\|_{\mathcal{L}(Y,\mathcal{H}_K)} \leq L\big\|\mu_x - \mu_y\big\|_k^\alpha, \qquad \forall(\mu_x, \mu_y) \in X_\mu \times X_\mu. \tag{2.2}$$

According to [31], the set X_μ of mean embeddings is a separable compact set of continuous functions on $\bar{X}$. Denote the marginal distribution of ρ on X_μ by ρ_{X_μ}. Let $L^2_{\rho_{X_\mu}}$ be the Hilbert space of square-integrable functions on X_μ with norm $\|\cdot\|_{L^2_{\rho_{X_\mu}}}$ given by

$$\big\|f\big\|_{L^2_{\rho_{X_\mu}}} = \big\langle f, f\big\rangle^{1/2}_{\rho_{X_\mu}} = \left(\int_{X_\mu} \big|f(\mu_x)\big|^2 d\rho_{X_\mu}\right)^{1/2}.$$

Define an integral operator L_K on $L^2_{\rho_{X_\mu}}$ associated with the Mercer kernel $K : X_\mu \times X_\mu \to \mathbb{R}$ by

$$L_K(f) = \int_{X_\mu} K_{\mu_x} f(\mu_x) d\rho_{X_\mu}, \qquad f \in L^2_{\rho_{X_\mu}}. \tag{2.3}$$

Since the set X_μ is compact and K is a Mercer kernel, L_K is a positive compact operator on $L^2_{\rho_{X_\mu}}$. Then for any $r > 0$, its r-th power L^r_K is well defined according to the spectral theorem in functional calculus.

Throughout the paper, we assume a *regularity condition* for the regression function f_ρ as

$$f_\rho = L^r_K(g_\rho) \text{ for some } g_\rho \in L^2_{\rho_{X_\mu}} \text{ and } r > 0. \tag{2.4}$$

The assumption means that the regression function lies in the range of operator L^r_K. The special case $r = 1/2$ corresponds to $f_\rho \in \mathcal{H}_K$. According to [11], the operator $L^{1/2}_K : \overline{\mathcal{H}_K} \to \mathcal{H}_K$ is an isomorphism, in which $\overline{\mathcal{H}_K}$ denotes the closure of $\mathcal{H}_K$ in $L^2_{\rho_{X_\mu}}$. Namely, for any $f \in \overline{\mathcal{H}_K}$, $L^{1/2}_K f \in \mathcal{H}_K$ and $\|f\|_{L^2_{\rho_{X_\mu}}} = \|L^{1/2}_K f\|_K$.

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

$$\mathcal{N}(\lambda) = \mathrm{Tr}((\lambda I + L_K)^{-1} L_K), \qquad \lambda > 0.$$

For the effective dimension $\mathcal{N}(\lambda)$, we need a *capacity condition* which focuses on rates of increment of $\mathcal{N}(\lambda)$ and is stated for some $\beta \in (0, 1]$ and $\mathcal{C}_0 > 0$ as

$$\mathcal{N}(\lambda) \leq \mathcal{C}_0 \lambda^{-\beta}, \qquad \forall \lambda > 0. \tag{2.5}$$

Throughout the paper, we assume that the sample $D = \{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}$ is drawn independently according to the Borel probability measure ρ, and $\{x_{i,s}\}_{s=1}^{d_i}$ is drawn independently according to the probability distribution x_i for $i = 1, 2, \ldots, |D|$. The windowing function $V : \mathbb{R}_+ \to \mathbb{R}$ is assumed to be differentiable with $V'_+(0) = 1$ (w.l.o.g. by scaling) but not necessarily convex. It is assumed that there exist some $p > 0$ and $c_p > 0$ such that

$$|V'(s) - V'_+(0)| \leq c_p s^p, \qquad \forall s > 0, \tag{2.6}$$

and

$$C_V = \sup_{s\in(0,\infty)} |V'(s)| < \infty \text{ with } V'(s) > 0 \text{ for } s > 0. \tag{2.7}$$

These assumptions are satisfied by the windowing functions for many classical loss functions $l_\sigma(s) = \sigma^2 V(\frac{s^2}{\sigma^2})$ including Welsch loss: $l_\sigma(s) = \sigma^2[1 - \exp(-\frac{s^2}{2\sigma^2})]$, Cauchy loss: $l_\sigma(s) = \sigma^2 \log(1 + \frac{s^2}{2\sigma^2})$, and Fair loss: $l_\sigma(s) = \sigma^2\big[\frac{|s|}{\sigma} - \log(1 + \frac{|s|}{\sigma})\big]$. Such loss functions are well studied in the literature of robust regression dealing with vector-valued data. They are proposed here in the DR setting to improve the robustness to non-Gaussian noise and outliers when dealing with distribution-valued data. It can be witnessed that the Welsch loss and Cauchy loss are non-convex but satisfy the so-called redescending property meaning that the derivative $l'_\sigma(s)$ increases near the origin but decreases to 0 when s is far away from the origin.

Our first main result, to be proved in Section 4, describes explicit learning rates for the error $\left\|f^\sigma_{\hat{D},\lambda} - f_\rho\right\|_{L^2_{\rho_{X_\mu}}}$ of RDR in terms of the sample size $|D|$ of the data sets D, $\hat{D}$, and robust scaling parameter σ. The expectations are taken with respect to D and $\hat{D}$.

Theorem 1. *Suppose that the regularity condition* (2.4) *holds for some* $r > 0$ *and* $|y| \leq M$ *almost surely. Assume the capacity condition* (2.5) *for some* $\beta \in (0, 1]$*, smoothness conditions (2.6), (2.7) with* $p > 0$*, and that the mapping* $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ *is* (α, L)*-Hölder continuous for some* $\alpha \in (0, 1]$ *and* $L > 0$*. If the sample size in the second stage sampling satisfies* $d_1 = d_2 = \cdots = d_{|D|} = d$*, then by choosing*

$$\lambda \;=\; \begin{cases} |D|^{-\frac{1}{1+\beta}}, & \text{when } r \in (0, 1/2), \\ |D|^{-\frac{1}{2r+\beta}}, & \text{when } r \in [1/2, 1], \\ |D|^{-\frac{1}{2+\beta}}, & \text{when } r \in (1, \infty), \end{cases} \tag{2.8}$$

and

$$d \;=\; \begin{cases} |D|^{\frac{2}{\alpha(1+\beta)}}, & \text{when } r \in (0, 1/2), \\ |D|^{\frac{1+2r}{\alpha(2r+\beta)}}, & \text{when } r \in [1/2, 1], \\ |D|^{\frac{1}{\alpha}(\frac{3}{2+\beta})}, & \text{when } r \in (1, \infty), \end{cases} \tag{2.9}$$

we have

$$\mathbb{E}\left[\left\|f^\sigma_{\hat{D},\lambda} - f_\rho\right\|_{L^2_{\rho_{X_\mu}}}\right] \;=\; \begin{cases} \mathcal{O}\Big(\max\Big\{|D|^{-\frac{r}{1+\beta}}, \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}\Big\}\Big), & \text{when } r \in (0, 1/2), \\ \mathcal{O}\Big(\max\Big\{|D|^{-\frac{r}{2r+\beta}}, \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}}\Big\}\Big), & \text{when } r \in [1/2, 1], \\ \mathcal{O}\Big(\max\Big\{|D|^{-\frac{1}{2+\beta}}, \frac{|D|^{\frac{p+1}{2+\beta}}}{\sigma^{2p}}\Big\}\Big), & \text{when } r \in (1, \infty). \end{cases} \tag{2.10}$$

Learning rates for DR provided in the existing literature are those for least squares DR schemes in [31, 12, 27]. Reference [31] is the first work presenting learning rates for the least squares regressor $f^{ls}_{\hat{D},\lambda}$ defined in (1.1). It derived optimal learning rates under the regularity condition (2.4) with $r \in (1/2, 1]$ and suboptimal rate with $r = 1/2$. The suboptimal rate in the case $r = 1/2$ was improved to the optimal one in [12] via a novel integral operator method that is based on a second order decomposition technique for inverses of operators in Banach spaces [20]. Reference [27] proposes a kernel based stochastic gradient method in a DR setting where mini-batching is used for selection of data points in each iteration. In Theorem 1, we provide learning rates for any value of the regularity index r in the whole range $(0, \infty)$, in contrast to [31, 12] that carried out analysis only for $r \in [1/2, 1]$. Hence, the analysis for DR has been enriched. Moreover, in the explicit bounds of Theorem 1, the participation of the scaling parameter σ introduced for the windowing function V indicates differences of our work from the aforementioned results. One difference is that RDR

possesses a flexibility in selecting σ for robustness, in contrast to the current DR methods without taking robustness into consideration.

There have been many studies on robust learning algorithms in different aspects. For example, references [16] and [17] consider some robust empirical risk minimization schemes for regression. Inspired by convex risk minimization over infinite-dimensional Hilbert spaces, robustness of support vector machines is extensively investigated in [4, 5, 7, 37]. The maximum correntropy criterion induced loss is considered in [14] for regression over compact hypothesis spaces. Modal regression with robust kernels is studied in [15]. References [23, 10, 24, 39] investigate learning behaviors of minimum error entropy algorithms. Based on gradient descent iterations, reference [21] presents an efficient kernel based robust gradient descent algorithm for regression. Error analysis in these studies is carried out with standard covering number arguments for data obtained from one-stage sampling. In contrast to these works, under the capacity assumption on the effective dimension $\mathcal{N}(\lambda)$, we derive error bounds and learning rates with integral operator techniques. For the purpose of dealing with samples of probability distributions, we develop a robust regression method for DR with data obtained from two-stage sampling and provide some analysis for selecting the regularization parameter λ and the second stage sample size d.

The following corollary is a direct consequence of Theorem 1. It shows that the RDR has nice learning performances when the scaling parameter σ is chosen to be large enough.

Corollary 1. *Under the same assumption of Theorem 1, if the scaling parameter σ is chosen as*

$$\sigma \geq \begin{cases} |D|^{\frac{p+1+r}{2p(1+\beta)}}, & \text{when } r \in (0, 1/2), \\ |D|^{\frac{p+1+r}{2p(2r+\beta)}}, & \text{when } r \in [1/2, 1], \\ |D|^{\frac{p+1+r}{2p(2+\beta)}}, & \text{when } r \in (1, \infty), \end{cases} \tag{2.11}$$

then we have

$$\mathbb{E}\left[\left\|f^{\sigma}_{\hat{D},\lambda} - f_\rho\right\|^2_{L^2_{\rho_{X_\mu}}}\right] = \begin{cases} \mathcal{O}\left(|D|^{-\frac{r}{1+\beta}}\right), & \text{when } r \in (0, 1/2), \\ \mathcal{O}\left(|D|^{-\frac{r}{2r+\beta}}\right), & \text{when } r \in [1/2, 1], \\ \mathcal{O}\left(|D|^{-\frac{1}{2+\beta}}\right), & \text{when } r \in (1, \infty). \end{cases} \tag{2.12}$$

When the regularity index r lies in the lower regularity range $(0, 1/2)$, the regression function f_ρ does not belong to $\mathcal{H}_K$ in general. The learning rates of order $\mathcal{O}(|D|^{-\frac{r}{1+\beta}})$ provided in Corollary 1 for this range are sub-optimal. In the setting of regression with one-stage sampling, optimal rates $\mathcal{O}(|D|^{-\frac{r}{2r+\beta}})$ have been achieved by some learning algorithms such as point kernel-based regression [38] and semi-supervised distributed learning [3], but not by any robust regression methods. It would be interesting to improve the learning rates for the proposed RDR scheme when $r \in (0, 1/2)$. When $r \in [1/2, 1]$, the learning rates of order $\mathcal{O}(|D|^{-\frac{r}{2r+\beta}})$ in Corollary 1 become optimal. However, when the regularity index r exceeds 1, the learning rates in Corollary 1 remain the same order and the higher regularity does not help. This is the well-known saturation phenomenon in kernel based learning (e.g. [20]). It would be interesting to overcome the saturation phenomenon by developing some novel methods for the proposed regularized RDR.

The role of a large scaling parameter σ in learning performances of minimum error entropy algorithms with one-stage sampling was analyzed in [23]. Corollary 1 shows the same role of large σ but is demonstrated for RDR with two-stage sampling.

In a framework of regularized regression, our second main result provides a novel quantitative description on robustness of RDR by considering the expected error between the RDR estimator $f^{\sigma}_{\hat{D},\lambda}$ (in which the robustness is induced by the scaling parameter σ) and the classical least squares DR estimator $f^{ls}_{\hat{D},\lambda}$ (without robustness).

Theorem 2. *Let the sample set $D = \{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}$ be drawn independently according to probability measure ρ. Let $f_{\hat{D},\lambda}^{ls}$ denote the classical least square DR estimator in (1.1). Suppose that the sample size in the second stage sampling satisfies $d_1 = d_2 = \cdots = d$. Then for any given sample size $|D|$, d and regularization parameter $\lambda > 0$, there holds*

$$\mathbb{E}\left[\left\|f_{\hat{D},\lambda}^{\sigma} - f_{\hat{D},\lambda}^{ls}\right\|_{L^2_{\rho_{X_\mu}}}\right] \leq \widetilde{C}\frac{(\lambda^{-(p+\frac{1}{2})}+1)(\lambda^{-\frac{3}{2}}d^{-\frac{\alpha}{2}}\hat{\mathcal{A}}^2_{|D|,\lambda} + \lambda^{-\frac{1}{2}}\hat{\mathcal{A}}_{|D|,\lambda})}{\sigma^{2p}}. \tag{2.13}$$

$\widetilde{C}$ is a constant independent of D, d, λ, σ and the explicit form will be given in the proof. $\hat{\mathcal{A}}_{|D|,\lambda}$ is defined by $\hat{\mathcal{A}}_{|D|,\lambda} = \frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} + 1$ in which $\mathcal{A}_{|D|,\lambda} = \frac{2\kappa}{\sqrt{|D|}}(\frac{\kappa}{\sqrt{|D|\lambda}} + \sqrt{\mathcal{N}(\lambda)})$.

When the sample size $|D|$ and d are large enough, our last main result is a quantitative description on the robust $L^2_{\rho_{X_\mu}}$-gap between RDR estimator $f_{\hat{D},\lambda}^{\sigma}$ and least square DR estimator $f_{\hat{D},\lambda}^{ls}$.

Corollary 2. *Under same conditions of Theorem 2, for any given regularization parameter $\lambda > 0$, there holds*

$$\overline{\lim}_{\substack{|D|\to\infty \\ d\to\infty}} \mathbb{E}\left[\left\|f_{\hat{D},\lambda}^{\sigma} - f_{\hat{D},\lambda}^{ls}\right\|_{L^2_{\rho_{X_\mu}}}\right] \leq \widetilde{C}\frac{(\lambda^{-(p+1)} + \lambda^{-\frac{1}{2}})}{\sigma^{2p}}, \tag{2.14}$$

where $\widetilde{C}$ is a constant to be given explicitly in the proof of Theorem 2.

Recall that $f_{\hat{D},\lambda}^{\sigma}$ is generated by the introduction of the scaling parameter σ that delivers the robustness to the DR scheme. Since the classical least square DR estimator $f_{\hat{D},\lambda}^{ls}$ does not possess robustness, we know that, when the $L^2_{\rho_{X_\mu}}$-distance between $f_{\hat{D},\lambda}^{\sigma}$ and $f_{\hat{D},\lambda}^{ls}$ gets smaller, there will be less robustness of the RDR scheme induced by l_σ. In nonparametric regression problems, to enhance the robustness of RDR, one may choose appropriately small σ for use. Actually, in practice, for different purposes, the scaling parameter σ may be chosen to be large or small. This idea also matches the work in [14] which handles maximum correntropy criterion. Their work also reveals that too small σ would influence the convergence of the regressor $f_{\hat{D},\lambda}^{\sigma}$ to f_ρ. Also, the small σ case has been interpreted as modal regression in [15]. From above analysis and recent works [14, 21, 25], we know that, in practice, a moderate scaling parameter σ should be chosen appropriately to balance robustness and convergence of RDR.

[10]

3 Key Analysis and Error Decompositions

In this section, we present the key analysis and error decompositions for RDR. We first introduce the following robust regression scheme associated with the sample $D = \{(\mu_{x_i}, y_i)\}_{i=1}^{|D|}$ obtained from the one-stage sampling

$$f_{D,\lambda}^{\sigma} = \arg\min_{f\in\mathcal{H}_K}\left\{\frac{\sigma^2}{|D|}\sum_{i=1}^{|D|} V\Big(\frac{[f(\mu_{x_i}) - y_i]^2}{\sigma^2}\Big) + \lambda\|f\|_K^2\right\}. \tag{3.1}$$

It plays the role of a stepping stone in in our analysis for learning with the two-stage sampling. Then we need a data-free minimizer f_λ for the regularized least squares regression as

$$f_\lambda = \arg\min_{f\in\mathcal{H}_K}\left\{\|f - f_\rho\|^2_{L^2_{\rho_{X_\mu}}} + \lambda\|f\|_K^2\right\}. \tag{3.2}$$

Now we can make a decomposition for the error $f^{\sigma}_{\hat{D},\lambda} - f_\rho$ as

$$f^{\sigma}_{\hat{D},\lambda} - f_\rho = \left(f^{\sigma}_{\hat{D},\lambda} - f^{\sigma}_{D,\lambda}\right) + \left(f^{\sigma}_{D,\lambda} - f_\lambda\right) + \left(f_\lambda - f_\rho\right). \tag{3.3}$$

The above error decomposition will be used to estimate the error norm $\left\|f^{\sigma}_{\hat{D},\lambda} - f_\rho\right\|_{L^2_{\rho_{X_\mu}}}$ in our sampling operator approach. The last term can be easily estimated from the regularity condition (2.4) as

$$\|f_\lambda - f_\rho\|_{L^2_{\rho_{X_\mu}}} \le \|g_\rho\|_{L^2_{\rho_{X_\mu}}} \lambda^{\min\{r,1\}} \tag{3.4}$$

by results [33] on the well-studied regularization scheme (3.2). The second term $f^{\sigma}_{D,\lambda} - f_\lambda$ reflects the error caused by the robust regression scheme and will be bounded in terms of the scaling parameter σ. The first term $f^{\sigma}_{\hat{D},\lambda} - f^{\sigma}_{D,\lambda}$ reflects the error incurred by the second-stage sampling and will be bounded in terms of the size d of the second-stage sample. This section includes statements of the error bounds which will be proved in the appendix.

Let us introduce two sampling operators in our sampling operator approach for the two-stage sampling process. The sampling operator $S_D : \mathcal{H}_K \to \mathbb{R}^{|D|}$ corresponding to the first-stage sampling is defined as

$$S_D f = (f(\mu_{x_i}))_{i=1}^{|D|}, \qquad f \in \mathcal{H}_K,$$

and a scaled adjoint operator $S_D^T : \mathbb{R}^{|D|} \to \mathcal{H}_K$ is given by

$$S_D^T \mathbf{c} = \frac{1}{|D|} \sum_{i=1}^{|D|} c_i K_{\mu_{x_i}}, \qquad \mathbf{c} = (c_i)_{i=1}^{|D|} \in \mathbb{R}^{|D|}.$$

The first stage empirical integral operator $L_{K,D}$ is then defined by

$$L_{K,D}(f) = S_D^T S_D(f) = \frac{1}{|D|} \sum_{i=1}^{|D|} f(\mu_{x_i}) K_{\mu_{x_i}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \langle K_{\mu_{x_i}}, f\rangle_K K_{\mu_{x_i}}, \qquad f \in \mathcal{H}_K.$$

The sampling operator $\hat{S}_D : \mathcal{H}_K \to \mathbb{R}^{|D|}$ corresponding to the second-stage sampling is given by

$$S_D f = (f(\mu_{\hat{x}_i}))_{i=1}^{|D|}, \qquad f \in \mathcal{H}_K,$$

with its scaled adjoint operator $\hat{S}_D^T : \mathbb{R}^{|D|} \to \mathcal{H}_K$ by $S_D^T \mathbf{c} = \frac{1}{|D|} \sum_{i=1}^{|D|} c_i K_{\mu_{\hat{x}_i}}$. The corresponding empirical integral operator $L_{K,\hat{D}}$ is then defined by

$$L_{K,\hat{D}}(f) = \hat{S}_D^T \hat{S}_D(f) = \frac{1}{|D|} \sum_{i=1}^{|D|} f(\mu_{\hat{x}_i}) K_{\mu_{\hat{x}_i}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \langle K_{\mu_{\hat{x}_i}}, f\rangle_K K_{\mu_{\hat{x}_i}}, \qquad f \in \mathcal{H}_K. \tag{3.5}$$

The empirical integral operators $L_{K,D}$ and $L_{K,\hat{D}}$ can be used to represent $f^{\sigma}_{\hat{D},\lambda}$ and $f^{\sigma}_{D,\lambda}$ as follows. In the following, we denote the output vector by $y = (y_i)_{i=1}^{|D|}$.

Lemma 1. *Let $f^{\sigma}_{\hat{D},\lambda}$ and $f^{\sigma}_{D,\lambda}$ be defined by* (1.2) *and* (3.1). *Then*

$$f^{\sigma}_{\hat{D},\lambda} = (\lambda I + L_{K,\hat{D}})^{-1} \hat{S}_D^T y - (\lambda I + L_{K,\hat{D}})^{-1} E_{\hat{D},\lambda,\sigma} \tag{3.6}$$

and

$$f^{\sigma}_{D,\lambda} = (\lambda I + L_{K,D})^{-1} S_D^T y - (\lambda I + L_{K,D})^{-1} E_{D,\lambda,\sigma}, \tag{3.7}$$

where $E_{\hat{D},\lambda,\sigma}$ and $E_{D,\lambda,\sigma}$ are quantities depending on the scaling parameter σ given by

$$E_{\hat{D},\lambda,\sigma} = \frac{1}{|D|}\sum_{i=1}^{|D|}\Big[V'\Big(\frac{[f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i]^2}{\sigma^2}\Big)-V'(0)\Big](f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i)K_{\mu_{\hat{x}_i}}, \tag{3.8}$$

$$E_{D,\lambda,\sigma} = \frac{1}{|D|}\sum_{i=1}^{|D|}\Big[V'\Big(\frac{[f^{\sigma}_{D,\lambda}(\mu_{x_i})-y_i]^2}{\sigma^2}\Big)-V'(0)\Big](f^{\sigma}_{D,\lambda}(\mu_{x_i})-y_i)K_{\mu_{x_i}}. \tag{3.9}$$

For the second term of the error decomposition (3.3), we recall a representation of f_λ found in [33] as

$$f_\lambda = (\lambda I + L_K)^{-1}L_K f_\rho \tag{3.10}$$

which implies $f_\lambda = (\lambda I + L_K)^{-1}L_K f_\rho$. Combining this with the representation (3.7) of $f^{\sigma}_{D,\lambda}$ in Lemma 1, we have the following decomposition

$$\begin{aligned} f^{\sigma}_{D,\lambda} - f_\lambda &= (\lambda I + L_{K,D})^{-1}S_D^T y - (\lambda I + L_K)^{-1}L_K f_\rho - (\lambda I + L_{K,D})^{-1}E_{D,\lambda,\sigma} \\ &= (\lambda I + L_{K,D})^{-1}(S_D^T y - L_K f_\rho) + [(\lambda I + L_{K,D})^{-1} - (\lambda I + L_K)^{-1}]L_K f_\rho - (\lambda I + L_{K,D})^{-1}E_{D,\lambda,\sigma}. \end{aligned}$$

Applying the formula $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ for operator inverses to the operators $A = \lambda I + L_{K,D}, B = \lambda I + L_K$ on $\mathcal{H}_K$ and using (3.10) yield the following representation

$$f^{\sigma}_{D,\lambda} - f_\lambda = (\lambda I + L_{K,D})^{-1}\left\{(S_D^T y - L_K f_\rho) + (L_K - L_{K,D})f_\lambda\right\} - (\lambda I + L_{K,D})^{-1}E_{D,\lambda,\sigma}. \tag{3.11}$$

This expression allows us to estimate the error term $\|f^{\sigma}_{D,\lambda} - f_\lambda\|_{L^2_{\rho_{X_\mu}}}$ by bounding $S_D^T y - L_K f_\rho, L_K - L_{K,D}$ and the quantity $E_{D,\lambda,\sigma}$.

Proposition 1. *Assume $|y| \le M$ almost surely. Under the smoothness condition (2.6) for V, we have*

$$\begin{aligned} \mathbb{E}\big[\|f^{\sigma}_{D,\lambda} - f_\lambda\|_{L^2_{\rho_{X_\mu}}}\big] \le\ & C_{p,\kappa,C_V,M}\Big(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1\Big)^2\Big(\mathcal{A}'_{|D|,\lambda} + \mathcal{A}_{|D|,\lambda}\|f_\lambda\|_K\Big) \\ & + C_{p,\kappa,C_V,M}\Big(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1\Big)\frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})}+1), \end{aligned}$$

where

$$\mathcal{A}'_{|D|,\lambda} = \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{|D|}}$$

and $C_{p,\kappa,C_V,M}$ is a constant independent of $|D|, \sigma$ or λ.

For the first term of the error decomposition (3.3), we can combine the representations (3.6), (3.7) for $f^{\sigma}_{\hat{D},\lambda}, f^{\sigma}_{D,\lambda}$ with some estimates involving integral operators and the robust quantities $E_{\hat{D},\lambda,\sigma}$, $E_{D,\lambda,\sigma}$, and get the following bound.

Proposition 2. *Suppose that the regularity condition (2.4) holds with $r > 0$, $|y| \le M$ almost surely, and that the mapping $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ is (α, L)-Hölder continuous with $\alpha \in (0,1]$ and $L > 0$. Then we have*

$$\begin{aligned} \mathbb{E}\Big[\big\|f^{\sigma}_{\hat{D},\lambda} - f^{\sigma}_{D,\lambda}\big\|_{L^2_{\rho_{X_\mu}}}\Big] \le \bar{C}\frac{1}{\lambda^{\frac{1}{2}}d^{\frac{\alpha}{2}}}\left(\frac{1}{\lambda d^{\frac{\alpha}{2}}}+1\right)(\hat{\mathcal{A}}^2_{|D|,\lambda}+\hat{\mathcal{A}}_{|D|,\lambda})\Big[\hat{\mathcal{A}}^2_{|D|,\lambda}\cdot\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \\ +\hat{\mathcal{A}}^2_{|D|,\lambda}(\hat{\mathcal{A}}_{|D|,\lambda}-1)\|f_\lambda\|_K + \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1}) + \|f_\lambda\|_K\Big] + \bar{C}\hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+1)}+\lambda^{-\frac{1}{2}}), \end{aligned}$$

where $\bar{C}$ is a constant independent of $|D|$, d, and σ.

Combining Propositions 1 and 2 with (3.4) and applying the triangle inequality to the error decomposition (3.3), we obtain immediately the following theorem with a general error bound without decaying restrictions on the effective dimension $\mathcal{N}(\lambda)$. The result is crucial in deriving our learning rates.

Theorem 3. *Suppose that the regularity condition* (2.4) *holds with* $r > 0$ *and* $|y| \leq M$ *almost surely. If the mapping* $K_{(\cdot)} : X_\mu \to \mathcal{L}(Y, \mathcal{H}_K)$ *is* (α, L)*-Hölder continuous with* $\alpha \in (0, 1]$ *and* $L > 0$*, then we have*

$$\begin{aligned}
&\mathbb{E}\left[\left\|f^{\sigma}_{D,\lambda} - f_\rho\right\|^2_{L^2_{\rho_{X_\mu}}}\right] \leq \bar{C}\frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\left(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1\right)(\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \\
&\left[\hat{\mathcal{A}}^2_{|D|,\lambda} \cdot \frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} + \hat{\mathcal{A}}^2_{|D|,\lambda}(\hat{\mathcal{A}}_{|D|,\lambda} - 1)\|f_\lambda\|_K + \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})} + \lambda^{-1}) + \|f_\lambda\|_K\right] \\
&+\bar{C}\hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+1)} + \lambda^{-\frac{1}{2}}) + C_{p,\kappa,C_V,M}\left(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right)^2\left(\mathcal{A}'_{|D|,\lambda} + \mathcal{A}_{|D|,\lambda}\|f_\lambda\|_K\right) \\
&+C_{p,\kappa,C_V,M}\left(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right)\frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})} + 1) + \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{\min\{r,1\}}.
\end{aligned}$$

4 Proofs of Main Results

We estimate the learning rates of RDR in this section. In the following, for convenience of analysis on learning rates, we use the convention that $A_{|D|} \lesssim B_{|D|}$ ($A_{|D|} = \mathcal{O}(B_{|D|})$) denotes that there exist some constant $C > 0$ independent of the cardinality $|D|$, d and σ such that $A_{|D|} \lesssim C \cdot B_{|D|}$ for any $|D|$ for some functions $A_{|D|}$, $B_{|D|}$ which may depend on $|D|$. Also, we use $A_{|D|} \lesssim 1$ to denote that there is a constant $C > 0$ independent of $|D|$, d and σ such that $A_{|D|} \leq C$. We need to estimate the right hand side of (**??**) in Theorem 3 in different regularity range of r when the regularization parameter λ and second stage sample size d take different orders of $|D|$, the cardinality of data set D. According to Smale and Zhou [33]

$$\|f_\lambda\|_K \quad \leq \quad \begin{cases} \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{r-\frac{1}{2}}, & r \in (0, 1/2); \\ \kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_{X_\mu}}}, & r \in [1/2, \infty); \end{cases} \tag{4.1}$$

this estimate will be used in the following in different regularity range of r. For convenience, we denote the main terms of right hand side of (**??**) of Theorem 3 by

$$\begin{aligned}
\mathcal{T}_{1,|D|,\lambda} &= \left(\frac{1}{\lambda d^{\frac{\alpha}{2}}}\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}\right)\frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}, \\
\mathcal{T}_{2,|D|,\lambda} &= \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\left(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1\right)(\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda})\Big[\hat{\mathcal{A}}^2_{|D|,\lambda}\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} + \hat{\mathcal{A}}_{|D|,\lambda}(\hat{\mathcal{A}}_{|D|,\lambda} - 1)\|f_\lambda\|_K \\
&\quad +\hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})} + \lambda^{-1}) + \|f_\lambda\|_K\Big], \\
\mathcal{T}_{3,|D|,\lambda} &= \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\left(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1\right)\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})} + \lambda^{-1})(\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}), \\
\mathcal{T}_{4,|D|,\lambda} &= \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+1)} + \lambda^{-1/2}), \\
\mathcal{T}_{5,|D|,\lambda} &= \hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}'_{|D|,\lambda} + \hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\|f_\lambda\|_K + \hat{\mathcal{A}}_{|D|,\lambda} \cdot \frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})} + 1), \\
\mathcal{T}_{6,|D|,\lambda} &= \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{\min\{r,1\}}.
\end{aligned}$$

In the following, the goal is to estimate the main terms $\mathcal{T}_{1,|D|,\lambda} \sim \mathcal{T}_{6,|D|,\lambda}$ for different regularity range of r when λ and d take different orders of $|D|$.

4.1 Learning rates for $r \in (0, 1/2)$

When $r \in (0, 1/2)$, (4.1) implies $\|f_\lambda\|_K \le \|g_\rho\|_{L^2_{\rho_{X_\mu}}} \lambda^{r-\frac{1}{2}}$, $r \in (0, 1/2)$. After taking $\lambda = |D|^{-\frac{1}{1+\beta}}$, $d = |D|^{\frac{2}{\alpha(1+\beta)}}$, the following basic estimates hold for $r \in (0, 1/2)$:

$$\frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} = |D|^{\frac{1}{2(1+\beta)}} |D|^{-\frac{1}{1+\beta}} = |D|^{-\frac{1}{2(1+\beta)}},$$

$$\frac{1}{\lambda d^{\frac{\alpha}{2}}} = |D|^{\frac{1}{1+\beta}} |D|^{-\frac{1}{1+\beta}} = 1.$$

According to the condition $\mathcal{N}(\lambda) \le \mathcal{C}_0 \lambda^{-\beta}$, $\beta \in (0, 1]$, we have

$$\mathcal{A}_{|D|,\lambda} = \frac{2\kappa}{\sqrt{|D|}} \Big(\frac{\kappa}{\sqrt{|D|\lambda}} + \sqrt{\mathcal{N}(\lambda)} \Big) \le 2\kappa(\kappa + \sqrt{\mathcal{C}_0}) |D|^{-\frac{r}{1+\beta}} \lesssim |D|^{-\frac{r}{1+\beta}}.$$

and

$$\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} \le 2\kappa(\kappa + \sqrt{\mathcal{C}_0}), \ \hat{\mathcal{A}}_{|D|,\lambda} = \frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} + 1 \le 2\kappa(\kappa + \sqrt{\mathcal{C}_0}) + 1.$$

Same procedure with above inequalities implies

$$\mathcal{A}'_{|D|,\lambda} \lesssim |D|^{-\frac{r}{1+\beta}} \text{ and } \frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \le 1 + \sqrt{\mathcal{C}_0}.$$

Then, since $\hat{\mathcal{A}}_{|D|,\lambda} \lesssim 1$ and $\hat{\mathcal{A}}_{|D|,\lambda} - 1 \lesssim 1$, it follows that

$$\mathcal{T}_{1,|D|,\lambda} = \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} \hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda} \Big) \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \lesssim |D|^{-\frac{1}{2(1+\beta)}} \lesssim |D|^{-\frac{r}{1+\beta}} \text{ (since } r \in (0, 1/2)).$$

Now turn to $\mathcal{T}_{2,|D|,\lambda}$. We spit $\mathcal{T}_{2,|D|,\lambda}$ into four parts and estimate them each other.

$$(i) \quad \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \hat{\mathcal{A}}^2_{|D|,\lambda} \frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \lesssim |D|^{-\frac{1}{2(1+\beta)}} \lesssim |D|^{-\frac{r}{1+\beta}}.$$

When $\lambda = |D|^{-\frac{1}{1+\beta}}$, $d = |D|^{\frac{2}{\alpha(1+\beta)}}$,

$$\begin{aligned} (ii) \quad \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \hat{\mathcal{A}}^2_{|D|,\lambda} (\hat{\mathcal{A}}_{|D|,\lambda} - 1) \|f_\lambda\|_K &\lesssim \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \cdot \lambda^{r-\frac{1}{2}} \\ &\lesssim |D|^{-\frac{1}{2(1+\beta)}} |D|^{-\frac{r-\frac{1}{2}}{1+\beta}} \lesssim |D|^{-\frac{r}{1+\beta}}. \end{aligned}$$

Same way with (ii) implies

$$(iii) \quad \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \|f_\lambda\|_K \lesssim |D|^{-\frac{r}{1+\beta}}.$$

For the forth term,

$$\begin{aligned} (iv) \quad & \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \hat{\mathcal{A}}_{|D|,\lambda} \sigma^{-2p} (\lambda^{-(p+\frac{3}{2})} + \lambda^{-1}) \\ & \lesssim |D|^{-\frac{1}{2(1+\beta)}} \cdot 1 \cdot \sigma^{-2p} \Big(|D|^{\frac{p+\frac{3}{2}}{1+\beta}} + |D|^{\frac{1}{1+\beta}} \Big) \\ & \lesssim \sigma^{-2p} \Big(|D|^{\frac{p+1}{1+\beta}} + |D|^{\frac{1}{2(1+\beta)}} \Big) \lesssim \max \Big\{ \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}, \frac{|D|^{\frac{1}{2(1+\beta)}}}{\sigma^{2p}} \Big\} \lesssim \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}. \end{aligned}$$

Combining $(i) \sim (iv)$, we obtain that

$$\mathcal{T}_{2,|D|,\lambda} \lesssim \max\Big\{|D|^{-\frac{r}{1+\beta}}, \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}\Big\}.$$

Then the same way with above (iv) implies

$$\mathcal{T}_{3,|D|,\lambda} = \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}}+1\Big)\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1})(\hat{\mathcal{A}}^2_{|D|,\lambda}+\hat{\mathcal{A}}_{|D|,\lambda}) \lesssim \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}.$$

For $\mathcal{T}_{4,|D|,\lambda}$, it follows that

$$\mathcal{T}_{4,|D|,\lambda} = \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+1)}+\lambda^{-1/2}) \le \sigma^{-2p}\Big(|D|^{\frac{p+1}{1+\beta}}+|D|^{\frac{1}{2(1+\beta)}}\Big) \lesssim \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}} \ (\text{since } p+1>\frac{1}{2}).$$

For $\mathcal{T}_{5,|D|,\lambda}$, since

$$\hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}'_{|D|,\lambda} \lesssim |D|^{-\frac{1}{2(1+\beta)}} \lesssim |D|^{-\frac{r}{1+\beta}},$$

$$\hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\|f_\lambda\|_K \le \hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{r-\frac{1}{2}} \lesssim |D|^{-\frac{1}{2(1+\beta)}}|D|^{\frac{r-\frac{1}{2}}{1+\beta}} \lesssim |D|^{-\frac{r}{1+\beta}},$$

$$\hat{\mathcal{A}}_{|D|,\lambda}\cdot\frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})}+1) \lesssim \sigma^{-2p}\Big(|D|^{\frac{1+p}{1+\beta}}+|D|^{\frac{1}{2(1+\beta)}}\Big) \lesssim \frac{|D|^{\frac{1+p}{1+\beta}}}{\sigma^{2p}},$$

it follows that

$$\mathcal{T}_{5,|D|,\lambda} \lesssim \max\Big\{|D|^{-\frac{r}{1+\beta}}, \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}\Big\}.$$

Then, for $\mathcal{T}_{6,|D|,\lambda}$,

$$\mathcal{T}_{6,|D|,\lambda} = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{\min\{r,1\}} = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}|D|^{-\frac{r}{1+\beta}} \lesssim |D|^{-\frac{r}{1+\beta}}.$$

Finally, combining above estimates for $\mathcal{T}_{1,|D|,\lambda} \sim \mathcal{T}_{6,|D|,\lambda}$ yields

$$\mathbb{E}\Big[\big\|f^\sigma_{D,\lambda}-f_\rho\big\|_{L^2_{\rho_{X_\mu}}}\Big] = \mathcal{O}\Big(\max\Big\{|D|^{-\frac{r}{1+\beta}}, \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}\Big\}\Big),\ r\in(0,1/2).$$

4.2 Learning rates for $r \in [1/2, 1]$

When $r \in [1/2, 1]$, (4.1) implies $\|f_\lambda\|_K \le \kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_{X_\mu}}}$, $r \in [1/2, 1]$. Before estimating the terms in (**??**), when $\lambda = |D|^{-\frac{1}{2r+\beta}}$, $d = |D|^{\frac{1+2r}{\alpha(2r+\beta)}}$, we derive the following basic estimates at first. After substituting λ and d, we have

$$\frac{1}{\lambda^{\frac{1}{2}}d^{\frac{\alpha}{2}}} = |D|^{-\frac{r}{2r+\beta}},\ \frac{1}{\lambda d^{\frac{\alpha}{2}}} = |D|^{\frac{1}{2r+\beta}}|D|^{-\frac{1+2r}{2(2r+\beta)}} = |D|^{\frac{1-2r}{2(2r+\beta)}} \le 1.$$

Use the condition $\mathcal{N}(\lambda) \le \mathcal{C}_0\lambda^{-\beta}$, $\beta \in (0, 1]$, it follows that

$$\mathcal{A}_{|D|,\lambda} = \frac{2\kappa}{\sqrt{|D|}}\Big(\frac{\kappa}{\sqrt{|D|\lambda}}+\sqrt{\mathcal{N}(\lambda)}\Big) \le 2\kappa(\kappa+\sqrt{\mathcal{C}_0})|D|^{-\frac{r}{2r+\beta}} \lesssim |D|^{-\frac{r}{2r+\beta}}.$$

and

$$\hat{\mathcal{A}}_{|D|,\lambda} = \frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1 = \frac{2\kappa}{\sqrt{|D|\lambda}}\Big(\frac{\kappa}{\sqrt{|D|\lambda}}+\sqrt{\mathcal{N}(\lambda)}\Big)+1 \le 2\kappa(\kappa+\sqrt{\mathcal{C}_0})+1.$$

In a same way with estimate of above $\mathcal{A}_{|D|,\lambda}$, $\hat{\mathcal{A}}_{|D|,\lambda}$, we have

$$\mathcal{A}'_{|D|,\lambda} \lesssim |D|^{-\frac{r}{2r+\beta}} \quad \text{and} \quad \frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \leq \frac{1}{|D|\lambda} + \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{|D|\lambda}} \leq 1 + \sqrt{\mathcal{C}_0}.$$

With above basic estimates, we can estimate main terms in (**??**) in Theorem 3 to derive the learning rates of RDR when $r \in [1/2, 1]$. Since $\hat{\mathcal{A}}_{|D|,\lambda} \lesssim 1$, $\hat{\mathcal{A}}_{|D|,\lambda} - 1 \lesssim 1$ and $\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \lesssim 1$, it follows that

$$\mathcal{T}_{1,|D|,\lambda} = \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}}\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}\Big)\frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \lesssim |D|^{-\frac{r}{2r+\beta}}.$$

Also,

$$\begin{aligned}\mathcal{T}_{2,|D|,\lambda} = \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\frac{1}{\lambda d^{\frac{\alpha}{2}}}(\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda})\Big[\hat{\mathcal{A}}^2_{|D|,\lambda}\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} + \hat{\mathcal{A}}_{|D|,\lambda}(\hat{\mathcal{A}}_{|D|,\lambda} - 1)\|f_\lambda\|_K \\ + \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})} + \lambda^{-1}) + \|f_\lambda\|_K\Big] \lesssim |D|^{-\frac{r}{2r+\beta}} + \frac{|D|^{\frac{(\frac{3}{2}+p)-r}{2r+\beta}}}{\sigma^{2p}} + \frac{|D|^{\frac{1-r}{2r+\beta}}}{\sigma^{2p}} \\ \lesssim \max\Big\{|D|^{-\frac{r}{2r+\beta}}, \frac{|D|^{(\frac{3}{2}-r)+p}}{\sigma^{2p}}\Big\},\end{aligned}$$

in which the last inequality follows from the fact $\frac{3}{2} + p > 1$. For $\mathcal{T}_{3,|D|,\lambda}$, we have

$$\mathcal{T}_{3,|D|,\lambda} = \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}(\frac{1}{\lambda d^{\frac{\alpha}{2}}}+1)\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1})(\hat{\mathcal{A}}^2_{|D|,\lambda}+\hat{\mathcal{A}}_{|D|,\lambda}) \lesssim \max\Big\{\frac{|D|^{\frac{\frac{3}{2}+p-r}{2r+\beta}}}{\sigma^{2p}}, \frac{|D|^{\frac{1-r}{2r+\beta}}}{\sigma^{2p}}\Big\} \lesssim \frac{|D|^{\frac{\frac{3}{2}+p-r}{2r+\beta}}}{\sigma^{2p}}.$$

Also, for $\mathcal{T}_{4,|D|,\lambda}$, we have

$$\mathcal{T}_{4,|D|,\lambda} = \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+1)}+\lambda^{-1/2}) \lesssim \sigma^{-2p}\Big(|D|^{\frac{p+1}{2r+\beta}}+|D|^{\frac{1}{2}(\frac{1}{2r+\beta})}\Big) \lesssim \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}} \ (\text{since } p+1 > \frac{1}{2}).$$

On the other hand, since

$$\begin{aligned}\hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}'_{|D|,\lambda} \lesssim |D|^{-\frac{r}{2r+\beta}}, \\ \hat{\mathcal{A}}_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\|f_\lambda\|_K \leq \hat{\mathcal{A}}_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_{X_\mu}}} \lesssim |D|^{-\frac{r}{2r+\beta}}, \\ \hat{\mathcal{A}}_{|D|,\lambda}\frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})}+1) \lesssim \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}} + \frac{|D|^{\frac{1}{2}(\frac{1}{2r+\beta})}}{\sigma^{2p}} \lesssim \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}},\end{aligned}$$

Therefore, we have

$$\mathcal{T}_{5,|D|,\lambda} \lesssim |D|^{-\frac{r}{2r+\beta}} + \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}} \lesssim \max\{|D|^{-\frac{r}{2r+\beta}}, \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}}\}.$$

Also, $\mathcal{T}_{6,|D|,\lambda}$ is estimated as follow

$$\mathcal{T}_{6,|D|,\lambda} = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{\min\{r,1\}} = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^r \lesssim |D|^{-\frac{r}{2r+\beta}}.$$

Now combining above estimates for the main terms $\mathcal{T}_{1,|D|,\lambda} \sim \mathcal{T}_{6,|D|,\lambda}$, noting the fact that when $r \in [1/2, 1]$, $\frac{3}{2} - r \leq 1$ and

$$|D|^{\frac{\frac{3}{2}+p-r}{2r+\beta}}/\sigma^{2p} \leq |D|^{\frac{1+p}{2r+\beta}}/\sigma^{2p},$$

we finally have

$$\mathbb{E}\Big[\big\|f^\sigma_{\hat{D},\lambda} - f_\rho\big\|^2_{L^2_{\rho_{X_\mu}}}\Big] = \mathcal{O}\Big(\max\Big\{|D|^{-\frac{r}{2r+\beta}}, \frac{|D|^{\frac{1+p}{2r+\beta}}}{\sigma^{2p}}\Big\}\Big), \ r \in [1/2, 1].$$

4.3 Learning rates for $r \in (1, \infty)$

When $r \in (0, \infty)$, (4.1) implies $\|f_\lambda\|_K \leq \kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_{X_\mu}}}$, $r > 1$. After taking $\lambda = |D|^{-\frac{1}{2+\beta}}$, $d = |D|^{\frac{1}{\alpha}(\frac{3}{2+\beta})}$, we start with following basic estimates:

$$\frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \leq |D|^{\frac{1}{2(2+\beta)}} |D|^{-\frac{3}{2(2+\beta)}} = |D|^{-\frac{1}{2+\beta}},$$

$$\frac{1}{\lambda d^{\frac{\alpha}{2}}} \leq |D|^{\frac{1}{2+\beta}} |D|^{-\frac{3}{2(2+\beta)}} = |D|^{\frac{2}{2(2+\beta)} - \frac{3}{2(2+\beta)}} = |D|^{-\frac{1}{2(2+\beta)}} < 1.$$

Since $\mathcal{N}(\lambda) \leq \mathcal{C}_0 \lambda^{-\beta}$, $\beta \in (0, 1]$, it follows that

$$\mathcal{A}'_{|D|,\lambda} = \frac{1}{|D|\sqrt{\lambda}} + \frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{|D|}} \leq (1 + \sqrt{\mathcal{C}_0}) \frac{\lambda^{-\frac{\beta}{2}}}{\sqrt{|D|}} \Big(\frac{1}{\sqrt{|D|}} \lambda^{\frac{\beta-1}{2}} + 1 \Big).$$

When $\lambda = |D|^{-\frac{1}{2+\beta}}$, $\frac{1}{\sqrt{|D|}} \lambda^{\frac{\beta-1}{2}} = |D|^{-\frac{1}{2}} |D|^{\frac{1-\beta}{2} \frac{1}{2+\beta}} = |D|^{-\frac{2\beta+1}{2(2+\beta)}} < 1$ and $\frac{\lambda^{-\frac{\beta}{2}}}{\sqrt{|D|}} = |D|^{\frac{\beta}{2(2+\beta)} - \frac{1}{2}} = |D|^{\frac{\beta-2-\beta}{2(2+\beta)}} = |D|^{-\frac{1}{2+\beta}}$, hence we have

$$\mathcal{A}'_{|D|,\lambda} \lesssim |D|^{-\frac{1}{2+\beta}}, \quad \text{same way implies} \quad \mathcal{A}_{|D|,\lambda} \lesssim |D|^{-\frac{1}{2+\beta}}.$$

Also,

$$\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \leq (1 + \sqrt{\mathcal{C}_0}) \frac{|D|^{-\frac{1}{2+\beta}}}{|D|^{-\frac{1}{2(2+\beta)}}} \leq (1 + \sqrt{\mathcal{C}_0}) |D|^{-\frac{1}{2(2+\beta)}} \leq 1 + \sqrt{\mathcal{C}_0}.$$

Same way implies

$$\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} \leq 1 + \sqrt{\mathcal{C}_0}.$$

Now based on above estimates, note that $\hat{\mathcal{A}}_{|D|,\lambda} \lesssim 1$, $\hat{\mathcal{A}}_{|D|,\lambda} - 1 \lesssim 1$ and $\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \lesssim 1$, we can estimate $\mathcal{T}_{1,|D|,\lambda} \sim \mathcal{T}_{6,|D|,\lambda}$ as follows,

$$\mathcal{T}_{1,|D|,\lambda} = \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} \hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}^2_{|D|,\lambda} \Big) \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \lesssim |D|^{-\frac{1}{2+\beta}}.$$

For $\mathcal{T}_{2,|D|,\lambda}$, since

$$(i) \qquad \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \hat{\mathcal{A}}^2_{|D|,\lambda} \frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}} \lesssim |D|^{-\frac{1}{2+\beta}},$$

$$(ii) \qquad \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \Big[\hat{\mathcal{A}}_{|D|,\lambda} (\hat{\mathcal{A}}_{|D|,\lambda} - 1) \|f_\lambda\|_K + \|f_\lambda\|_K \Big] \lesssim |D|^{-\frac{1}{2+\beta}}.$$

$$\begin{aligned} (iii) \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}} \Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}} + 1 \Big) (\hat{\mathcal{A}}^2_{|D|,\lambda} + \hat{\mathcal{A}}_{|D|,\lambda}) \hat{\mathcal{A}}_{|D|,\lambda} \sigma^{-2p} (\lambda^{-(p+\frac{3}{2})} + \lambda^{-1}) \\ \lesssim \sigma^{-2p} \Big(|D|^{\frac{p+\frac{3}{2}-1}{2+\beta}} + 1 \Big) \lesssim \frac{|D|^{\frac{p+\frac{1}{2}}{2+\beta}}}{\sigma^{2p}}, \end{aligned}$$

it follows that

$$\mathcal{T}_{2,|D|,\lambda} \lesssim \max \Big\{ |D|^{-\frac{1}{2+\beta}}, \frac{|D|^{\frac{p+\frac{1}{2}}{2+\beta}}}{\sigma^{2p}} \Big\}.$$

Same reason with (iii) implies

$$\mathcal{T}_{3,|D|,\lambda} = \frac{1}{\lambda^{\frac{1}{2}} d^{\frac{\alpha}{2}}}\Big(\frac{1}{\lambda d^{\frac{\alpha}{2}}}+1\Big)\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1})(\hat{\mathcal{A}}^2_{|D|,\lambda}+\hat{\mathcal{A}}_{|D|,\lambda}) \lesssim \frac{|D|^{\frac{p+\frac{1}{2}}{2+\beta}}}{\sigma^{2p}}.$$

For $\mathcal{T}_{4,|D|,\lambda}$, we have

$$\mathcal{T}_{4,|D|,\lambda} = \hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+1)}+\lambda^{-1/2}) \lesssim \sigma^{-2p}\Big(|D|^{\frac{p+1}{2+\beta}}+|D|^{\frac{1}{2(2+\beta)}}\Big) \lesssim \frac{|D|^{\frac{p+1}{2+\beta}}}{\sigma^{2p}}.$$

Also, since

$$\hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}'_{|D|,\lambda} \lesssim |D|^{-\frac{1}{2+\beta}},$$

$$\hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\|f_\lambda\|_K \leq \hat{\mathcal{A}}^2_{|D|,\lambda}\mathcal{A}_{|D|,\lambda}\kappa^{2r-1}\|g_\rho\|_{L^2_{\rho_{X_\mu}}} \lesssim \mathcal{A}_{|D|,\lambda} \lesssim |D|^{-\frac{1}{2+\beta}},$$

$$\hat{\mathcal{A}}_{|D|,\lambda}\frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})}+1) \lesssim \frac{1}{\sqrt{\lambda}}\sigma^{-2p}(\lambda^{-(p+\frac{1}{2})}+1) \lesssim \sigma^{-2p}(|D|^{\frac{p+1}{2+\beta}}+|D|^{\frac{1}{2(2+\beta)}}) \lesssim \frac{|D|^{\frac{p+1}{2+\beta}}}{\sigma^{2p}}, \quad (4.2)$$

it follows that

$$\mathcal{T}_{5,|D|,\lambda} \lesssim \max\Big\{|D|^{-\frac{1}{2+\beta}}, \frac{|D|^{\frac{p+1}{2+\beta}}}{\sigma^{2p}}\Big\}.$$

Finally, since $r > 1$, we have

$$\mathcal{T}_{6,|D|,\lambda} = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^{\min\{r,1\}} = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\lambda^1 = \|g_\rho\|_{L^2_{\rho_{X_\mu}}}\cdot|D|^{-\frac{1}{2+\beta}} \lesssim |D|^{-\frac{1}{2+\beta}}.$$

Combining above estimates for $\mathcal{T}_{1,|D|,\lambda} \sim \mathcal{T}_{6,|D|,\lambda}$, we obtain

$$\mathbb{E}\Big[\big\|f^\sigma_{\hat{D},\lambda}-f_\rho\big\|_{L^2_{\rho_{X_\mu}}}\Big] = \mathcal{O}\Big(\max\Big\{|D|^{-\frac{1}{2+\beta}}, \frac{|D|^{\frac{p+1}{2+\beta}}}{\sigma^{2p}}\Big\}\Big). \quad (4.3)$$

Now it is ready to provide proof for Corollary 1.

4.4 Proof of Corollary 1

Proof. The proof is obvious after using Theorem 1 and noting the fact that, if

$$\sigma \quad \geq \quad \begin{cases} |D|^{\frac{p+1+r}{2p(1+\beta)}}, & r \in (0,1/2); \\ |D|^{\frac{p+1+r}{2p(2r+\beta)}}, & r \in [1/2,1]; \\ |D|^{\frac{p+1+r}{2p(2+\beta)}}, & r \in (1,\infty), \end{cases} \quad (4.4)$$

then

$$|D|^{-\frac{r}{1+\beta}} \geq \frac{|D|^{\frac{p+1}{1+\beta}}}{\sigma^{2p}}, \ r \in (0,1/2);$$

$$|D|^{-\frac{r}{2r+\beta}} \geq \frac{|D|^{\frac{p+1}{2r+\beta}}}{\sigma^{2p}}, \ r \in [1/2,1];$$

$$|D|^{-\frac{r}{2+\beta}} \geq \frac{|D|^{\frac{p+1}{2+\beta}}}{\sigma^{2p}}, \ r \in (1,\infty).$$

4.5 Proof of Theorem 2

Proof. It follows from Fang et al. [12] that the least square distribution regressor has the form

$$f^{ls}_{\hat{D},\lambda} = (\lambda I + L_{K,\hat{D}})^{-1}\hat{S}_D^T y.$$

With the representation of $f^{\sigma}_{\hat{D},\lambda}$ in Lemma 1, we have

$$\begin{aligned}
&\|f^{\sigma}_{\hat{D},\lambda} - f^{ls}_{\hat{D},\lambda}\|_{L^2_{\rho_{X_\mu}}} = \|(\lambda I + L_{K,\hat{D}})^{-1}E_{\hat{D},\lambda,\sigma}\|_{L^2_{\rho_{X_\mu}}} = \|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}E_{\hat{D},\lambda,\sigma}\|_K \\
&\leq \|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|\|E_{\hat{D},\lambda,\sigma}\|_K.
\end{aligned}$$

Use Lemma 4 and take expectation on both sides of above inequality, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\left\|f^{\sigma}_{\hat{D},\lambda} - f^{ls}_{\hat{D},\lambda}\right\|_{L^2_{\rho_{X_\mu}}}\right] \\
&\leq 2^{2p}c_p\kappa\sigma^{-2p}\left[\kappa^{2p+1}(\sqrt{C_V}M)^{2p+1}\lambda^{-(p+\frac{1}{2})} + M^{2p+1}\right]\mathbb{E}_{\mathbf{z}^{|D|}}\left[\left\{\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}[\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|^2]\right\}^{1/2}\right] \\
&\leq 2^{2p}c_p\kappa\sigma^{-2p}\left[\kappa^{2p+1}(\sqrt{C_V}M)^{2p+1}\lambda^{-(p+\frac{1}{2})} + M^{2p+1}\right] \\
&\quad\cdot\left[\left(\sqrt{2}\lambda^{-\frac{3}{2}}\kappa(2+\sqrt{\pi})^{\frac{1}{2}}L\frac{2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}\right)\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}] + \sqrt{2}\lambda^{-1/2}\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}^{1/2}_{|D|,\lambda}]\right],
\end{aligned}$$

in which the first inequality follows from the basic fact that $\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}[\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|]\Big\} \leq \left\{\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}[\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|^2]\right\}^{1/2}$, the second inequality follows from Lemma 7. After using Lemma 6 to $\mathcal{C}_{|D|,\lambda}$ and taking out corresponding coefficients by setting

$$\widetilde{C} = 2^{2p}c_p\kappa[\kappa^{2p+1}(\sqrt{C_V}M)^{2p+1} + M^{2p+1}]\Big\{\sqrt{2}(2+\sqrt{\pi})^{\frac{1}{2}}L2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}(2\Gamma(3)+\log^2 2) + \sqrt{2}(2\Gamma(2)+\log 2)\Big\}, \tag{4.5}$$

we arrive at

$$\mathbb{E}\left[\left\|f^{\sigma}_{\hat{D},\lambda} - f^{ls}_{\hat{D},\lambda}\right\|_{L^2_{\rho_{X_\mu}}}\right] \leq \widetilde{C}\frac{(\lambda^{-(p+\frac{1}{2})}+1)(\lambda^{-\frac{3}{2}}d^{-\frac{\alpha}{2}}\hat{\mathcal{A}}^2_{|D|,\lambda} + \lambda^{-\frac{1}{2}}\hat{\mathcal{A}}_{|D|,\lambda})}{\sigma^{2p}},$$

which completes the proof.

4.6 Proof of Corollary 2

Proof. For any given $\lambda > 0$, note that

$$\lim_{|D|\to\infty}\mathcal{A}_{|D|,\lambda} = \frac{2\kappa}{\sqrt{|D|}}\left(\frac{\kappa}{\sqrt{|D|\lambda}} + \sqrt{\mathcal{N}(\lambda)}\right) = 0,$$

then it follows that

$$\lim_{|D|\to\infty}\hat{\mathcal{A}}_{|D|,\lambda} = \lim_{|D|\to\infty}\left(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}} + 1\right) = 1, \tag{4.6}$$

and

$$\overline{\lim}_{\substack{|D|\to\infty \\ d\to\infty}} d^{-\frac{\alpha}{2}}\hat{\mathcal{A}}^2_{|D|,\lambda} = 0. \tag{4.7}$$

From Theorem 2, we have known that

$$\mathbb{E}\left[\left\|f^{\sigma}_{\hat{D},\lambda}-f^{ls}_{\hat{D},\lambda}\right\|^2_{L^2_{\rho_{X_\mu}}}\right]\leq \widetilde{C}\frac{(\lambda^{-(p+\frac{1}{2})}+1)(\lambda^{-\frac{3}{2}}d^{-\frac{\alpha}{2}}\hat{\mathcal{A}}^2_{|D|,\lambda}+\lambda^{-\frac{1}{2}}\hat{\mathcal{A}}_{|D|,\lambda})}{\sigma^{2p}}.$$

By taking upper limit with respect to $\overset{|D|\to\infty}{d\to\infty}$ on above inequality and using (4.6) and (4.7), we obtain

$$\overline{\lim}_{\overset{|D|\to\infty}{d\to\infty}}\mathbb{E}\left[\left\|f^{\sigma}_{\hat{D},\lambda}-f^{ls}_{\hat{D},\lambda}\right\|^2_{L^2_{\rho_{X_\mu}}}\right]\leq \widetilde{C}\frac{(\lambda^{-(p+1)}+\lambda^{-\frac{1}{2}})}{\sigma^{2p}},$$

which completes the proof.

5 Numerical Experiments

In this section, we implement numerical experiments for our proposed regularized RDR scheme with some robust loss functions realized by appropriately chosen window functions V, such as Cauchy loss, Huber loss and Fair loss, and compare them with the regularized least square DR scheme to demonstrate the robustness of RDR scheme to outliers. In this experiment, we utilize the iteratively re-weighted least squares algorithm to solve the optimization problem and concatenated cross validation method to select best hayperparameters for a certain loss function.

5.1 Experiment Setup

We compare our regularized RDR algorithm (1.2) with the regularized least square DR algorithm (1.1) on a benchmark problem from [30], which aims at learning the entropy of Gaussian distributions. We first choose a random matrix $M\in\mathbb{R}^{2\times 2}$, where each entry $M_{i,j}$ is uniformly distributed on $[0,1]$ ($M_{i,j}\sim U[0,1]$). For the first stage, we generate 200 sample sets from $\{N(0,\Sigma_s)\}_{s=1}^{200}$, where the covariance matrix $\Sigma_s=R(\alpha_s)MM^TR(\alpha_s)^T$, and $R(\alpha_s)$ is the 2d rotation matrix with angle $\alpha_s\sim U[0,\pi]$. For the second stage, we sample 100 2d points i.i.d. from each $N(0,\Sigma_s)$. The target function is the entropy of the first marginal distribution: $Y=\frac{1}{2}\ln(2\pi e(\Sigma_s)_{1,1})$. Then we use 100 sample sets for training and another 100 sample sets for the test.

5.2 Algorithms

According to the representor theorem, the solution of (1.2) can be represented by

$$f^{\sigma}_{\hat{D},\lambda}(x)=\sum_{i=1}^{|D|}\alpha_{\hat{D},i}K(\mu_x,\mu_{\hat{x}_i})+b_{\hat{D}},\ x\in M_1(\mathcal{F}),$$

where $\boldsymbol{\alpha}_{\hat{D}}=(\alpha_{\hat{D},1},\alpha_{\hat{D},2},\ldots,\alpha_{\hat{D},|D|})^T\in\mathbb{R}^{|D|}$ and $b_{\hat{D}}\in\mathbb{R}$ are learned parameters from data $\hat{D}$. In this experiment, we choose K to be the Gaussian kernel with the bandwidth parameter $h>0$.

Since the regularized RDR scheme (1.2) is essentially a regularized M-estimation problem, we use the iteratively re-weighted least square (IRLS) algorithm to solve it, the pseudo code of which is shown in Algorithm 1.

Then we utilize the Concatenated Cross Validation (CCV) algorithm for the model selection problem of our proposed regularized RDR estimator of a specific robust loss function, which means how to select the hyperparameters in our model, i.e., the scaling parameter σ of the loss, the regularization parameter λ and the Gaussian kernel bandwidth h.

In the process of implementing cross-validation method, we need to specify a proper error criterion for the certain loss function l_σ. Since we focus on learning robust estimators, the least

Algorithm 1 IRLS Algorithm for Solving (1.2)

Input: data $\hat{D} = \{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\}_{i=1}^{|D|}$, regularization parameter $\lambda > 0$, Gaussian kernel bandwidth $h > 0$, scale parameter $\sigma > 0$, and the initial guess $\boldsymbol{\alpha}_0 = \mathbf{0} \in \mathbb{R}^{|D|}$, $b_0 = 0 \in \mathbb{R}$.
Output: the learned coefficients $\boldsymbol{\alpha}$ and b.

Calculate the gram matrix $K = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{|D|\times|D|}$, and set initial weight $w_i = 1$,
$(\boldsymbol{\alpha}_1, b_1) = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{|D|}, b\in\mathbb{R}} \sum_{i=1}^{|D|} w_i(y_i - K_i^T\boldsymbol{\alpha} - b)^2 + \lambda\boldsymbol{\alpha}^T\boldsymbol{\alpha}$,
while $|(\boldsymbol{\alpha}_1, b_1) - (\boldsymbol{\alpha}_0, b_0)| > err$ **do**
  $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_1$, $b_0 = b_1$,
  Set weight $w_i = \frac{|\nabla l_\sigma(y_i - K_i^T\boldsymbol{\alpha}_0 - b_0)|}{\sigma^2|y_i - K_i^T\boldsymbol{\alpha}_0 - b_0|}$,
  $(\boldsymbol{\alpha}_1, b_1) = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{|D|}, b\in\mathbb{R}} \sum_{i=1}^{|D|} w_i(y_i - K_i^T\boldsymbol{\alpha} - b)^2 + \lambda\boldsymbol{\alpha}^T\boldsymbol{\alpha}$,
return $\boldsymbol{\alpha}_1, b_1$

square loss is not suitable for our algorithm. Notice that the Empirical Risk Minimization (ERM) algorithm of the loss function l_σ is actually

$$\arg\min_{f\in\mathcal{H}_K} \frac{1}{|D|}\sum_{i=1}^{|D|} l_\sigma(f(\mu_{\hat{x}_i}) - y_i),$$

where K is the Gaussian kernel in this experiment. The error criterion we choose in our CCV algorithm is exactly the loss function in the above ERM approach. Specifically, denote $\{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\}_{i=1}^{m}$ as the validation set, and $\{\hat{y}_{i,\sigma,\lambda,h}\}_{i=1}^{m}$ as the estimated values calculated by the IRLS algorithm for the specific scaling parameter σ, regularization parameter λ and Gaussian kernel bandwidth parameter h, the CCV algorithm for selecting the best hyperparameters is shown in Algorithm 2 which consists of three steps for the selection of the scaling parameter σ, in each step, the best σ value is selected according to the ERM approach based on the best σ value selected in the last step.

Algorithm 2 CCV algorithm for selecting the best hyperparameters

Input: data $\hat{D} = \{(\{x_{i,s}\}_{s=1}^{d_i}, y_i)\}_{i=1}^{|D|}$, initial scaling parameter $\sigma_0 > 0$.
Output: the best hyperparameters σ, λ and h.

Step 1: $(\sigma_1, \lambda_1, h_1) = \arg\min_{\sigma,\lambda,h} \frac{1}{m}\sum_{i=1}^{m} l_{\sigma_0}(y_i - \hat{y}_{i,\sigma,\lambda,h})$,
Step 2: $(\sigma_2, \lambda_2, h_2) = \arg\min_{\sigma,\lambda,h} \frac{1}{m}\sum_{i=1}^{m} l_{\sigma_1}(y_i - \hat{y}_{i,\sigma,\lambda,h})$,
Step 3: $(\sigma_3, \lambda_3, h_3) = \arg\min_{\sigma,\lambda,h} \frac{1}{m}\sum_{i=1}^{m} l_{\sigma_2}(y_i - \hat{y}_{i,\sigma,\lambda,h})$,
return σ_3, λ_3, h_3

5.3 Numerical Results

In this section, we demonstrate the numerical results of our experiment. We implement the algorithms in section 5.2 to learn the best estimators for Cauchy loss, Huber loss, Fair loss and Least Square loss respectively on two datasets, the first dataset is exactly generated according to section 5.1, the second dataset is just based on the first dataset except replacing several points in the training set by the random outliers. We can observe from Figure 1 that the learned estimators of all of four losses learn quite well for the first dataset which doesn't include noise, and the learned estimator of Least Square loss can even learn somehow better; In the case of the second dataset which includes outliers, the learned estimator of Least Square loss is highly influenced by the outliers and doesn't learn well comparing with the true function, whereas the learned estimators of

square loss (without noise).pdf square loss (without noise).png square loss (without noise).jpg square loss (without no

square loss (with outliers).pdf square loss (with outliers).png square loss (with outliers).jpg square loss (with outliers)

loss (without noise).pdf loss (without noise).png loss (without noise).jpg loss (without noise).mps loss (without noise)

loss (with outliers).pdf loss (with outliers).png loss (with outliers).jpg loss (with outliers).mps loss (with outliers).jpeg

loss (without noise).pdf loss (without noise).png loss (without noise).jpg loss (without noise).mps loss (without noise)

loss (with outliers).pdf loss (with outliers).png loss (with outliers).jpg loss (with outliers).mps loss (with outliers).jpeg

loss (without noise).pdf loss (without noise).png loss (without noise).jpg loss (without noise).mps loss (without noise)

loss (with outliers).pdf loss (with outliers).png loss (with outliers).jpg loss (with outliers).mps loss (with outliers).jpeg

Figure 1: The learned entropy of the first marginal distribution of a rotated 2d Gaussian w.r.t. the rotation angle for Least Square loss, Cauchy loss, Huber loss and Fair loss on noise-free data (left) and data with outliers (right).

the robust losses, i.e. Cauchy loss, Huber loss and Fair loss still learn as well as that learned from noise-free data ignoring the influence of the outliers. Such numerical results actually demonstrate the robustness of our regularized RDR scheme with respect to the outliers.

Acknowledgements

The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 11202819], the CityU Strategic Grant 7005511. The last author is supported partially by the Research Grants Council of Hong Kong [Project # CityU 11307319], Hong Kong Institute for Data Science, and National Science Foundation of China [Project No. 12061160462]. This paper was written when the last author visited SAMSI/Duke during his sabbatical leave. He would like to express his gratitude to their hospitality and financial support.

Appendix

This appendix provides detailed proofs of the representations and estimates stated in Section 3.

A Representations and norm estimates with integral operators

This part derives representations of the estimators in terms of integral operators and provides related lemmas concerning their norms. We first prove Lemma 1.

Proof of Lemma 1. Taking the Fréchet derivative of the regularized functional in (1.2) yields

$$\frac{1}{|D|}\sum_{i=1}^{|D|} V'\Big(\frac{[f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i]^2}{\sigma^2}\Big)(f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i)K_{\mu_{\hat{x}_i}}+\lambda f^{\sigma}_{\hat{D},\lambda}=0. \tag{A.1}$$

It follows that

$$\frac{1}{|D|}\sum_{i=1}^{|D|}\Big[V'\Big(\frac{[f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i]^2}{\sigma^2}\Big)-V'(0)\Big](f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i)K_{\mu_{\hat{x}_i}}$$
$$+V'(0)\frac{1}{|D|}\sum_{i=1}^{|D|}(f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i})-y_i)K_{\mu_{\hat{x}_i}}+\lambda f^{\sigma}_{\hat{D},\lambda}=0.$$

Substituting the representation of $E_{\hat{D},\lambda,\sigma}$ and using the definitions of $L_{K,\hat{D}}$ and $\hat{S}^T_D y$ together with the normalization condition $V'(0)=1$, we find

$$E_{\hat{D},\lambda,\sigma}+[L_{K,\hat{D}}f^{\sigma}_{\hat{D},\lambda}-\hat{S}^T_D y]+\lambda f^{\sigma}_{\hat{D},\lambda}=0.$$

Namely,

$$(\lambda I+L_{K,\hat{D}})f^{\sigma}_{\hat{D},\lambda}-\hat{S}^T_D y+E_{\hat{D},\lambda,\sigma}=0.$$

Hence we have the first representation (3.6). The second one follows immediately after replacing $\hat{D}$ by D in the above procedure. ■

In the following, we use $\mathbb{E}_{\mathbf{z}^{|D|}}[\cdot]$ to denote the expectation w.r.t. $\mathbf{z}^{|D|}=\{z_i=(\mu_{x_i},y_i)\}_{i=1}^{|D|}$. Use $\mathbb{E}_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}$ to denote the conditional expectation w.r.t. sample $\big\{\{x_{i,s}\}_{s=1}^{d_i}\big\}_{i=1}^{|D|}$ conditioned on $\{z_1,z_2,...,z_{|D|}\}$. Namely

$$\mathbb{E}_{\mathbf{z}^{|D|}}[\cdot]:=\mathbb{E}_{\{(\mu_{x_i},y_i)\}_{i=1}^{|D|}}[\cdot],\qquad \mathbb{E}_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}[\cdot]:=\mathbb{E}_{\{\{x_{i,s}\}_{s=1}^{d_i}\}_{i=1}^{|D|}\big|\{z_i\}_{i=1}^{|D|}}[\cdot].$$

The following lemma found in [12] will be used for dealing with approximations of integral operators.

Lemma 2. *Suppose the boundedness condition* (2.1) *of kernels k and K and (α,L)-Hölder continuity condition* (2.2) *holds for K. If $d_1=d_2=\cdots=d_{|D|}=d$, then*

$$\Big\{\mathbb{E}_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}\Big[\big\|\hat{S}^T_D y-S^T_D y\big\|_K^2\Big]\Big\}^{\frac{1}{2}}\le(2+\sqrt{\pi})^{\frac{1}{2}}ML\frac{2^{\frac{\alpha}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}},$$
$$\Big\{\mathbb{E}_{\mathbf{x}^{\mathbf{d}},|D||\mathbf{z}^{|D|}}\Big[\big\|L_{K,\hat{D}}-L_{K,D}\big\|^2\Big]\Big\}^{\frac{1}{2}}\le B_K^{\frac{1}{2}}L(2+\sqrt{\pi})^{\frac{1}{2}}\frac{2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}.$$

The RKHS norms of $f^{\sigma}_{\hat{D},\lambda}$ and $f^{\sigma}_{D,\lambda}$ can be bounded as follows.

Lemma 3. *Assume the smoothness condition (2.7) for V and $|y|\le M$ almost surely, then*

$$\|f^{\sigma}_{\hat{D},\lambda}\|_K\le\sqrt{C_V}M\lambda^{-1/2}, \tag{A.2}$$

and

$$\|f^{\sigma}_{D,\lambda}\|_K\le\sqrt{C_V}M\lambda^{-1/2}. \tag{A.3}$$

Proof. Denote

$$\mathcal{E}_{\hat{D},\sigma}(f) = \frac{\sigma^2}{|D|}\sum_{i=1}^{|D|} V\Big(\frac{[f(\mu_{\hat{x}_i}) - y_i]^2}{\sigma^2}\Big). \tag{A.4}$$

According to the definition of $f^{\sigma}_{\hat{D},\lambda}$ in (1.2), we have

$$\mathcal{E}_{\hat{D},\sigma}(f^{\sigma}_{\hat{D},\lambda}) + \lambda\|f^{\sigma}_{\hat{D},\lambda}\|_K^2 \le \mathcal{E}_{\hat{D},\sigma}(0). \tag{A.5}$$

Note that $V'(s) > 0$ for $s > 0$. Then $V(s) \ge V(0)$ for $s \ge 0$ which implies $\mathcal{E}_{\hat{D},\sigma}(f^{\sigma}_{\hat{D},\lambda}) \ge \sigma^2 V(0)$. It follows that

$$\begin{aligned}\lambda\|f^{\sigma}_{\hat{D},\lambda}\|_K^2 &\le \mathcal{E}_{\hat{D},\sigma}(0) - \mathcal{E}_{\hat{D},\sigma}(f^{\sigma}_{\hat{D},\lambda}) \le \frac{\sigma^2}{|D|}\sum_{i=1}^{|D|} V(\frac{|y|_i^2}{\sigma^2}) - \sigma^2 V(0)\\ &\le \frac{\sigma^2}{|D|}\sum_{i=1}^{|D|}\Big[V(\frac{|y_i|^2}{\sigma^2}) - V(0)\Big] \le \frac{C_V\sigma^2}{|D|}\sum_{i=1}^{|D|}\frac{|y_i|^2}{\sigma^2} \le C_V M^2.\end{aligned}$$

Hence, we have $\|f^{\sigma}_{\hat{D},\lambda}\|_K \le \sqrt{C_V}M\lambda^{-1/2}$. The same procedure with $\hat{D}$ replaced by D and $\mu_{\hat{x}_i}$ replaced by μ_{x_i} implies $\|f^{\sigma}_{D,\lambda}\|_K \le \sqrt{C_V}M\lambda^{-1/2}$. ■

The following lemma provides upper bounds for the RKHS norms of the quantities $E_{D,\lambda,\sigma}$ and $E_{\hat{D},\lambda,\sigma}$.

Lemma 4. *Assume $|y| \le M$ almost surely. Under the smoothness condition (2.6) for V, the norms of $E_{D,\lambda,\sigma}$ and $E_{\hat{D},\lambda,\sigma}$ can be bounded as*

$$\|E_{\hat{D},\lambda,\sigma}\|_K,\ \|E_{D,\lambda,\sigma}\|_K \le 2^{2p}c_p\sqrt{B_K}\sigma^{-2p}\Big[B_K^{p+1/2}(\sqrt{C_V}M)^{2p+1}\lambda^{-(p+\frac{1}{2})} + M^{2p+1}\Big]. \tag{A.6}$$

Proof. According to the smoothness condition (2.6) for V, we have

$$\begin{aligned}\|E_{\hat{D},\lambda,\sigma}\|_K &\le \frac{1}{|D|}\sum_{i=1}^{|D|}\Big|V'\Big(\frac{[f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i}) - y_i]^2}{\sigma^2}\Big) - V'(0)\Big||f^{\sigma}_{\hat{D},\lambda}(\mu_{\hat{x}_i}) - y_i|\cdot\|K_{\mu_{\hat{x}_i}}\|_K\\ &\le \frac{\kappa c_p\sigma^{-2p}}{|D|}\sum_{i=1}^{|D|}\Big(\|f^{\sigma}_{\hat{D},\lambda}\|_\infty + |y_i|\Big)^{2p+1}.\end{aligned}$$

But $\|f^{\sigma}_{\hat{D},\lambda}\|_\infty \le \sqrt{B_K}\|f^{\sigma}_{\hat{D},\lambda}\|_K$ by the reproducing property of the Mercer kernel K. Then we can apply Lemma 3 and see that

$$\begin{aligned}\|E_{\hat{D},\lambda,\sigma}\|_K &\le c_p\sqrt{B_K}\sigma^{-2p}\Big(\sqrt{B_K}\|f^{\sigma}_{\hat{D},\lambda}\|_K + M\Big)^{2p+1}\\ &\le 2^{2p}c_p\sqrt{B_K}\sigma^{-2p}[B_K^{p+1/2}(\sqrt{C_V}M)^{2p+1}\lambda^{-(p+\frac{1}{2})} + M^{2p+1}].\end{aligned}$$

The proof for the bound of $\|E_{D,\lambda,\sigma}\|_K$ is the same. ■

The following lemma with a standard proof is needed in our analysis.

Lemma 5. *Let a sample D be drawn independently according to a Borel probability measure ρ and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be a positive continuous function. If for some $m \in \mathbb{N}_+$, a positive random variable $X_{|D|,\lambda} \geq 0$ satisfies $X_{|D|,\lambda} \leq \Phi(\mathcal{A}_{|D|,\lambda}) \log^m \frac{2}{\delta}$ with probability at least $1-\delta$ for any $\delta \in (0,1)$, then*

$$\mathbb{E}\Big[X^s_{|D|,\lambda}\Big] \leq \big(2\Gamma(ms+1) + (\log 2)^{ms}\big)\Phi(\mathcal{A}_{|D|,\lambda})^s, \qquad \forall s \geq 1.$$

The same result holds when $\mathcal{A}_{|D|,\lambda}$ is replaced by $\mathcal{A}'_{|D|,\lambda}$.

Proof. The condition implies that, for $0 < \delta < 2$,

$$\text{Prob}\Big\{X_{|D|,\lambda} \leq \Phi(\mathcal{A}_{|D|,\lambda}) \log^m \frac{4}{\delta}\Big\} \geq 1 - \frac{\delta}{2}. \tag{A.7}$$

Make a variable change $\gamma = \Phi(\mathcal{A}_{|D|,\lambda})^s \log^{ms} \frac{4}{\delta}$. It follows that $\gamma^{\frac{1}{s}} = \Phi(\mathcal{A}_{|D|,\lambda}) \big(\log \frac{4}{\delta}\big)^m$ and $\frac{\delta}{2} = 2\exp\{-\gamma^{\frac{1}{ms}}/\Phi(\mathcal{A}_{|D|,\lambda})^{1/m}\}$. Note that for $\gamma > \Phi(\mathcal{A}_{|D|,\lambda})^s \log^{ms} 2$, the random variable $\xi = X^s_{|D|,\lambda}$ satisfies

$$\text{Prob}\Big\{\xi > \gamma\Big\} = \text{Prob}\Big\{\xi^{1/s} > \gamma^{1/s}\Big\} \leq \frac{\delta}{2} = 2\exp\Big\{-\frac{\gamma^{1/ms}}{\Phi(\mathcal{A}_{|D|,\lambda})^{1/m}}\Big\}.$$

Then by using the formula $\mathbb{E}[\xi] = \int_0^\infty \text{Prob}(\xi > \gamma)d\gamma$, we have

$$\begin{aligned}\mathbb{E}\Big[\xi\Big] &= \int_0^{\Phi(\mathcal{A}_{|D|,\lambda})^s \log^{ms} 2} \text{Prob}(\xi > \gamma)d\gamma + \int_{\Phi(\mathcal{A}_{|D|,\lambda})^s \log^{ms} 2}^\infty \text{Prob}(\xi > \gamma)d\gamma \\ &\leq \Phi(\mathcal{A}_{|D|,\lambda})^s \log^{ms} 2 + \int_{\Phi(\mathcal{A}_{|D|,\lambda})^s \log^{ms} 2}^\infty 2\exp\Big\{-\frac{\gamma^{1/ms}}{\Phi(\mathcal{A}_{|D|,\lambda})^{1/m}}\Big\}d\gamma.\end{aligned}$$

With a simple variable change $\gamma = \Phi(\mathcal{A}_{|D|,\lambda})^s x^{ms}$, we see that the integral in above second term equals

$$2ms\Phi(\mathcal{A}_{|D|,\lambda})^s \int_{\log 2}^\infty x^{ms-1}e^{-x}dx \leq 2\Gamma(ms+1)\Phi(\mathcal{A}_{|D|,\lambda})^s,$$

which completes the proof for the bound in terms of $\mathcal{A}_{|D|,\lambda}$. Using the same procedures with $\mathcal{A}_{|D|,\lambda}$ replaced by $\mathcal{A}'_{|D|,\lambda}$, we known the inequality holds with $\mathcal{A}_{|D|,\lambda}$ replaced by $\mathcal{A}'_{|D|,\lambda}$. ■

In our error analysis with an integral operator approach, we need

$$\mathcal{B}_{|D|,\lambda} = \|(\lambda I + L_K)^{-\frac{1}{2}}(S_D^T y - L_K f_\rho)\|_K, \tag{A.8}$$

$$\mathcal{C}_{|D|,\lambda} = \|(\lambda I + L_K)(\lambda I + L_{K,D})^{-1}\|, \tag{A.9}$$

$$\mathcal{D}_{|D|,\lambda} = \|(\lambda I + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})\|. \tag{A.10}$$

Probabilistic bounds for these quantities can be found in [3, 20, 22] which together with Lemma 5 applied to the functions $\Phi_1(x) = \frac{2M(\kappa+1)}{\kappa}x$, $\Phi_2(x) = \Big(\frac{x}{\sqrt{\lambda}} + 1\Big)^2$, and $\Phi_3(x) = 2x$ give the following bounds in expectations.

Lemma 6. *For any $s \geq 0$, the quantities $\mathcal{B}_{|D|,\lambda}$, $\mathcal{C}_{|D|,\lambda}$ and $\mathcal{D}_{|D|,\lambda}$ defined by* (A.8), (A.9) *and* (A.10) *satisfy*

$$\begin{aligned}&\mathbb{E}[\mathcal{B}^s_{|D|,\lambda}] \leq (2\Gamma(s+1) + \log^s 2)\Big(\frac{2M(\kappa+1)}{\kappa}\mathcal{A}'_{|D|,\lambda}\Big)^s, \\ &\mathbb{E}[\mathcal{C}^s_{|D|,\lambda}] \leq (2\Gamma(2s+1) + \log^{2s} 2)\hat{\mathcal{A}}^{2s}_{|D|,\lambda}, \\ &\mathbb{E}[\mathcal{D}^s_{|D|,\lambda}] \leq (2\Gamma(s+1) + \log^s 2)\Big(2\mathcal{A}_{|D|,\lambda}\Big)^s.\end{aligned}$$

B Bounding the error involving the scaling parameter

With the above preparation, we can now prove Proposition 1 for estimating the error term $\mathbb{E}\big[\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_{L^2_{\rho_{X_\mu}}}\big]$.

Proof of Proposition 1. Note that $\|g\|_{L^2_{\rho_{X_\mu}}} = \|L_K^{1/2} g\|_K$ for any $g \in L^2_{\rho_{X_\mu}}$. Then we know that

$$\left\|(\lambda I + L_K)^{1/2} g\right\|_K \geq \left\|L_K^{1/2} g\right\|_K = \|g\|_{L^2_{\rho_{X_\mu}}}$$

and for $g \in \mathcal{H}_K$,

$$\left\|(\lambda I + L_K)^{1/2} g\right\|_K \geq \lambda^{1/2} \|g\|_K .$$

In particular,

$$\max\{\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_{L^2_{\rho_{X_\mu}}}, \sqrt{\lambda}\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_K\} \leq \|(\lambda I + L_K)^{1/2} \left(f^{\sigma}_{D,\lambda} - f_{\lambda}\right)\|_K.$$

Applying the expression (3.11) for $f^{\sigma}_{D,\lambda} - f_{\lambda}$ to the above bound and denoting $Q := (S_D^T y - L_K f_\rho) + (L_K - L_{K,D}) f_\lambda$, we know that $\max\{\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_{L^2_{\rho_{X_\mu}}}, \sqrt{\lambda}\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_K\}$ is bounded by

$$\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}(\lambda I + L_{K,D})^{-1/2} Q\|_K + \|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1} E_{D,\lambda,\sigma}\|_K.$$

Inserting $(\lambda I + L_K)^{1/2}(\lambda I + L_K)^{-1/2}$ in the middle, we see that the first term can be further bounded by

$$\begin{aligned}&\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\| \, \|(\lambda I + L_{K,D})^{-1/2}(\lambda I + L_K)^{1/2}\| \, \|(\lambda I + L_K)^{-1/2} Q\|_K \\ &+\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\|\|(\lambda I + L_{K,D})^{-1/2} E_{D,\lambda,\sigma}\|_K.\end{aligned}$$

The first two operator norms above can be bounded by $\mathcal{C}^{1/2}_{|D|,\lambda}$ due to the identity $\|T_1^s T_2^s\| \leq \|T_1 T_2\|^s$ valid for any $s \in (0,1]$ and positive self-adjoint operators T_1, T_2. Thus, in terms of the notations $\mathcal{B}_{|D|,\lambda}, \mathcal{D}_{|D|,\lambda}$ defined in (A.8) and (A.10) for the norms of $(\lambda I + L_K)^{-\frac{1}{2}}(S_D^T y - L_K f_\rho)$ and $(\lambda I + L_K)^{-\frac{1}{2}}(L_K - L_{K,D})$, noting the fact $\|(\lambda I + L_{K,D})^{-1/2}\| \leq \lambda^{-1/2}$, we have

$$\begin{aligned}&\max\{\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_{L^2_{\rho_{X_\mu}}}, \sqrt{\lambda}\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_K\} \\ &\leq \mathcal{C}_{|D|,\lambda}\mathcal{B}_{|D|,\lambda} + \mathcal{C}_{|D|,\lambda}\mathcal{D}_{|D|,\lambda}\|f_\lambda\|_K + \mathcal{C}^{1/2}_{|D|,\lambda}\frac{1}{\sqrt{\lambda}}\|E_{D,\lambda,\sigma}\|_K. \qquad \text{(B.1)}\end{aligned}$$

Taking expectations and using the Schwarz inequality, we know that $\mathbb{E}\big[\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_{L^2_{\rho_{X_\mu}}}\big]$ can be bounded by

$$\left\{\mathbb{E}[\mathcal{C}^2_{|D|,\lambda}]\right\}^{\frac{1}{2}}\left\{\mathbb{E}[\mathcal{B}^2_{|D|,\lambda}]\right\}^{\frac{1}{2}} + \left\{\mathbb{E}[\mathcal{C}^2_{|D|,\lambda}]\right\}^{\frac{1}{2}}\left\{\mathbb{E}[\mathcal{D}^2_{|D|,\lambda}]\right\}^{\frac{1}{2}}\|f_\lambda\|_K + \left\{\mathbb{E}[\mathcal{C}_{|D|,\lambda}]\right\}^{\frac{1}{2}}\frac{1}{\sqrt{\lambda}}\left\{\mathbb{E}[\|E_{D,\lambda,\sigma}\|^2_K]\right\}^{\frac{1}{2}}.$$

At the end, we apply Lemmas 4 and 6 and know that the desired bound for $\mathbb{E}\big[\|f^{\sigma}_{D,\lambda} - f_{\lambda}\|_{L^2_{\rho_{X_\mu}}}\big]$ holds true. The proof of Proposition 1 is complete. ■

C Bounding the error involving the second-stage sampling

This part aims at proving Proposition 2 for estimating the norm $\|f^{\sigma}_{\hat{D},\lambda} - f^{\sigma}_{D,\lambda}\|_{L^2_{\rho_{X_\mu}}}$ of the error term involving the second-stage sampling. To this end, we need basic estimates for the norm $\mathbb{E}_{\mathbf{z}^{|D|}}\big[\|f^{\sigma}_{D,\lambda}\|_K^2\big]$ and the operator norm $\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|$.

Proposition 3. *Assume $|y| \le M$ almost surely. Under the smoothness condition (2.6) for V, there holds*

$$\begin{aligned}\Big\{\mathbb{E}\big[\|f^{\sigma}_{D,\lambda}\|_K^2\big]\Big\}^{1/2} \quad &\le \quad 2C_{p,\kappa,C_V,M}\Big(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1\Big)^2\left(\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}}+\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}\|f_\lambda\|_K\right)\\ &\quad +2C_{p,\kappa,C_V,M}\Big(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1\Big)\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1})+2\|f_\lambda\|_K,\end{aligned}$$

where $C_{p,\kappa,C_V,M}$ is the constant given in the statement of Proposition 1

Proof. By (B.1), we have

$$\|f^{\sigma}_{D,\lambda}-f_\lambda\|_K \le \frac{1}{\sqrt{\lambda}}\mathcal{C}_{|D|,\lambda}\mathcal{B}_{|D|,\lambda}+\frac{1}{\sqrt{\lambda}}\mathcal{C}_{|D|,\lambda}\mathcal{D}_{|D|,\lambda}\|f_\lambda\|_K+\frac{1}{\sqrt{\lambda}}\mathcal{C}^{1/2}_{|D|,\lambda}\frac{1}{\sqrt{\lambda}}\|E_{D,\lambda,\sigma}\|_K.$$

Combining this with the triangle inequality $\|f^{\sigma}_{D,\lambda}\|_K \le \|f^{\sigma}_{D,\lambda}-f_\lambda\|_K+\|f_\lambda\|_K$ yields

$$\|f^{\sigma}_{D,\lambda}\|_K \le \frac{1}{\sqrt{\lambda}}\mathcal{C}_{|D|,\lambda}\mathcal{B}_{|D|,\lambda}+\frac{1}{\sqrt{\lambda}}\mathcal{C}_{|D|,\lambda}\mathcal{D}_{|D|,\lambda}\|f_\lambda\|_K+\frac{1}{\sqrt{\lambda}}\mathcal{C}^{1/2}_{|D|,\lambda}\frac{1}{\sqrt{\lambda}}e(\lambda)+\|f_\lambda\|_K,$$

where we have denoted the right-hand side of the error bound (A.6) for $\|E_{D,\lambda,\sigma}\|_K$ by $e(\lambda)$. It follows that

$$\|f^{\sigma}_{D,\lambda}\|_K^2 \le \frac{4}{\lambda}\Big\{\mathcal{C}^2_{|D|,\lambda}\mathcal{B}^2_{|D|,\lambda}+\mathcal{C}^2_{|D|,\lambda}\mathcal{D}^2_{|D|,\lambda}\|f_\lambda\|_K^2+\mathcal{C}_{|D|,\lambda}\frac{1}{\lambda}e(\lambda)^2\Big\}+4\|f_\lambda\|_K^2.$$

Taking expectations tells us that $\mathbb{E}\Big[\|f^{\sigma}_{D,\lambda}\|_K^2\Big]$ can be bounded by

$$\begin{aligned}&\frac{4}{\lambda}\Bigg(\Big\{\mathbb{E}\big[\mathcal{C}^4_{|D|,\lambda}\big]\Big\}^{1/2}\Big\{\mathbb{E}\big[\mathcal{B}^4_{|D|,\lambda}\big]\Big\}^{1/2}+\Big\{\mathbb{E}\big[\mathcal{C}^4_{|D|,\lambda}\big]\Big\}^{1/2}\Big\{\mathbb{E}\big[\mathcal{D}^4_{|D|,\lambda}\big]\Big\}^{1/2}\|f_\lambda\|_K^2\\ &+\Big\{\mathbb{E}\big[\mathcal{C}_{|D|,\lambda}\big]\Big\}\frac{1}{\lambda}e(\lambda)^2\Bigg)+4\|f_\lambda\|_K^2.\end{aligned}$$

Hence

$$\begin{aligned}\Big\{\mathbb{E}\big[\|f^{\sigma}_{D,\lambda}\|_K^2\big]\Big\}^{1/2} \quad &\le \quad \frac{2}{\sqrt{\lambda}}\Big\{\mathbb{E}\big[\mathcal{C}^4_{|D|,\lambda}\big]\Big\}^{1/4}\Big\{\mathbb{E}\big[\mathcal{B}^4_{|D|,\lambda}\big]\Big\}^{1/4}\\ &\quad +\frac{2}{\sqrt{\lambda}}\Big\{\mathbb{E}\big[\mathcal{C}^4_{|D|,\lambda}\big]\Big\}^{1/4}\Big\{\mathbb{E}\big[\mathcal{D}^4_{|D|,\lambda}\big]\Big\}^{1/4}\|f_\lambda\|_K\\ &\quad +\frac{2}{\sqrt{\lambda}}\Big\{\mathbb{E}\big[\mathcal{C}_{|D|,\lambda}\big]\Big\}^{1/2}\frac{1}{\sqrt{\lambda}}e(\lambda)+2\|f_\lambda\|_K.\end{aligned}$$

Then the desired bound for $\mathbb{E}\big[\|f^{\sigma}_{D,\lambda}\|_K^2\big]\Big\}^{1/2}$ follows from Lemma 6 with the constant $C_{p,\kappa,C_V,M}$ given in Proposition 1. The proof of the proposition is complete. ■

Lemma 7. *Suppose the boundedness condition (2.1) of kernels k and K and (α, L)-Hölder continuity condition (2.2) holds for K. If $d_1 = d_2 = \cdots = d_{|D|} = d$, then*

$$\left\{\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}[\|L_K^{1/2}(\lambda I+L_{K,\hat{D}})^{-1}\|^2]\right\}^{1/2} \leq \left(\sqrt{2}\lambda^{-\frac{3}{2}}\sqrt{B_K}(2+\sqrt{\pi})^{\frac{1}{2}}L\frac{2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}\right)\mathcal{C}_{|D|,\lambda}+\sqrt{2}\lambda^{-1/2}\mathcal{C}_{|D|,\lambda}^{1/2}.$$

Proof. Write $(\lambda I + L_{K,\hat{D}})^{-1}$ as $\left\{(\lambda I + L_{K,\hat{D}})^{-1} - (\lambda I + L_{K,D})^{-1}\right\} + (\lambda I + L_{K,D})^{-1}$ and identify $\left\{(\lambda I + L_{K,\hat{D}})^{-1} - (\lambda I + L_{K,D})^{-1}\right\}$ with $(\lambda I+L_{K,D})^{-1}\left\{L_{K,D} - L_{K,\hat{D}}\right\}(\lambda I+L_{K,\hat{D}})^{-1}$. We follow the same procedure as in the proof of Proposition 1 and know that $\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|$ can be bounded by

$$\begin{aligned}
& \|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\|\ \|(\lambda I + L_{K,D})^{-1/2}(\lambda I + L_K)^{1/2}\| \\
& \|(\lambda I + L_K)^{-1/2}(L_{K,D} - L_{K,\hat{D}})(\lambda I + L_{K,\hat{D}})^{-1}\| \\
& +\|(\lambda I + L_K)^{1/2}(\lambda I + L_{K,D})^{-1/2}\|\ \|(\lambda I + L_{K,D})^{-1/2}\| \\
\leq\ & \mathcal{C}_{|D|,\lambda}\lambda^{-1/2}\lambda^{-1}\|L_{K,D} - L_{K,\hat{D}}\| + \lambda^{-1/2}\mathcal{C}_{|D|,\lambda}^{1/2},
\end{aligned}$$

It follows that

$$\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|^2 \leq 2\lambda^{-3}\mathcal{C}_{|D|,\lambda}^2\|L_{K,D} - L_{K,\hat{D}}\|^2 + 2\lambda^{-1}\mathcal{C}_{|D|,\lambda}.$$

Taking expectations and applying Lemma 2 verifies the desired bound. This completes the proof. ∎

We are in a position to prove Proposition 2.

Proof of Proposition 2. Based on the representations (3.6) and (3.7), we can decompose the difference $f_{\hat{D},\lambda}^{\sigma} - f_{D,\lambda}^{\sigma}$ as $I_1 - I_2$ where

$$I_1 = (\lambda I + L_{K,\hat{D}})^{-1}\hat{S}_D^T y - (\lambda I + L_{K,D})^{-1}S_D^T y, \tag{C.1}$$

$$I_2 = (\lambda I + L_{K,\hat{D}})^{-1}E_{\hat{D},\lambda,\sigma} - (\lambda I + L_{K,D})^{-1}E_{D,\lambda,\sigma}. \tag{C.2}$$

Then $\mathbb{E}[\|f_{\hat{D},\lambda}^{\sigma} - f_{D,\lambda}^{\sigma}\|_{L^2_{\rho_{X_\mu}}}] \leq \mathbb{E}[\|I_1\|_{L^2_{\rho_{X_\mu}}}] + \mathbb{E}[\|I_2\|_{L^2_{\rho_{X_\mu}}}]$ and we estimate the two terms in the following.

To estimate $\mathbb{E}[\|I_1\|_{L^2_{\rho_{X_\mu}}}] = \mathbb{E}[\|L_K^{1/2}I_1\|_K^2]$, we decompose $L_K^{1/2}I_1$ further as

$$\begin{aligned}
L_K^{1/2}I_1 &= (\lambda I + L_{K,\hat{D}})^{-1}(\hat{S}_D^T y - S_D^T y) + \left[(\lambda I + L_{K,\hat{D}})^{-1} - (\lambda I + L_{K,D})^{-1}\right]S_D^T y \\
&= L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}(\hat{S}_D^T y - S_D^T y) \\
&\quad +L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}(L_{K,D} - L_{K,\hat{D}})(\lambda I + L_{K,D})^{-1}S_D^T y.
\end{aligned} \tag{C.3}$$

Then we follow the same procedure as in the proof of Proposition 1 and apply Lemmas 2, 7 and 6 to obtain a bound for the first term in (C.3) as

$$\begin{aligned}
& \mathbb{E}\left[\left\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}(\hat{S}_D^T y - S_D^T y)\right\|_K\right] \\
\leq\ & \mathbb{E}_{\mathbf{z}^{|D|}}\left[\left\{\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}[\|L_K^{1/2}(\lambda I + L_{K,\hat{D}})^{-1}\|^2]\right\}^{1/2}\left\{E_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}\left[\left\|\hat{S}_D^T y - S_D y\right\|_K^2\right]\right\}^{\frac{1}{2}}\right]
\end{aligned}$$

$$\leq \left[\left(\sqrt{2}\lambda^{-3/2}(2+\sqrt{\pi})L\frac{2^{\frac{2+\alpha}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}\right)\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}]+\sqrt{2}\lambda^{-1/2}\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}^{1/2}]\right]\cdot(2+\sqrt{\pi})^{1/2}LM\frac{2^{\frac{\alpha}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}$$

$$\leq c_1 d^{-\frac{\alpha}{2}}\left[\frac{1}{\lambda^{\frac{3}{2}}d^{\frac{\alpha}{2}}}\left(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1\right)^2+\lambda^{-\frac{1}{2}}\left(\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}+1\right)\right],$$

where c_1 is a constant independent of $|D|, d, \lambda$, or σ.

The second term of (C.3) can be seen from (3.7) of Lemma 1 to be equal to

$$L_K^{1/2}(\lambda I+L_{K,\hat{D}})^{-1}(L_{K,D}-L_{K,\hat{D}})\Big[f_{D,\lambda}^{\sigma}+(\lambda I+L_{K,D})^{-1}E_{D,\lambda,\sigma})\Big].$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_K^{1/2}(\lambda I+L_{K,\hat{D}})^{-1}(L_{K,D}-L_{K,\hat{D}})(\lambda I+L_{K,D})^{-1}S_D^T y\right\|_K\right]\\
\leq\ &\mathbb{E}_{\mathbf{z}^{|D|}}\left[\left\{\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}[\|L_K^{1/2}(\lambda I+L_{K,\hat{D}})^{-1}\|^2]\right\}^{1/2}\left\{\mathbb{E}_{\mathbf{x}^{\mathbf{d},|D|}|\mathbf{z}^{|D|}}\left[\left\|L_{K,\hat{D}}-L_{K,D}\right\|_K^2\right]\right\}^{\frac{1}{2}}\right.\\
&\left.\left(\|f_{D,\lambda}^{\sigma}\|_K+\|(\lambda I+L_{K,D})^{-1}E_{D,\lambda,\sigma}\|_K\right)\right]\\
&\leq\sqrt{B_K}L(2+\pi)^{\frac{1}{2}}\frac{2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}\left[\sqrt{2}(2+\sqrt{\pi})^{\frac{1}{2}}L2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}\frac{1}{\lambda^{\frac{3}{2}}d^{\frac{\alpha}{2}}}+\sqrt{2}\lambda^{-\frac{1}{2}}\right]\\
&\cdot\left(\{\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}^2]\}^{1/2}\{\mathbb{E}_{\mathbf{z}^{|D|}}[\|f_{D,\lambda}^{\sigma}\|_K^2]\}^{1/2}+\{\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}]\}^{1/2}\{\mathbb{E}_{\mathbf{z}^{|D|}}[\|f_{D,\lambda}^{\sigma}\|_K^2]\}^{1/2}\right)\\
&+\sqrt{B_K}L(2+\pi)^{\frac{1}{2}}\frac{2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}}{d^{\frac{\alpha}{2}}}\left[\sqrt{2}(2+\sqrt{\pi})^{\frac{1}{2}}L2^{\frac{\alpha+2}{2}}B_k^{\frac{\alpha}{2}}\frac{1}{\lambda^{\frac{3}{2}}d^{\frac{\alpha}{2}}}+\sqrt{2}\lambda^{-\frac{1}{2}}\right]\\
&\cdot\left(\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}]+\mathbb{E}_{\mathbf{z}^{|D|}}[\mathcal{C}_{|D|,\lambda}^{1/2}]\right)\cdot 2^{2p}c_p\sqrt{B_K}\sigma^{-2p}[B_K^{p+1/2}(\sqrt{C_V}M)^{2p+1}\lambda^{-(p+\frac{3}{2})}+M^{2p+1}\lambda^{-1}],
\end{aligned}$$

where Lemma 4 has been used. Applying Lemma 6 to $\mathcal{C}_{|D|,\lambda}$ and Proposition 3, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_K^{1/2}(\lambda I+L_{K,\hat{D}})^{-1}(L_{K,D}-L_{K,\hat{D}})(\lambda I+L_{K,D})^{-1}S_D^T y\right\|_K\right]\\
\leq\ &c_2 d^{-\frac{\alpha}{2}}\left[\frac{1}{\lambda^{\frac{3}{2}}d^{\frac{\alpha}{2}}}+\lambda^{-\frac{1}{2}}\right]\cdot\left[\hat{\mathcal{A}}_{|D|,\lambda}^2+\hat{\mathcal{A}}_{|D|,\lambda}\right]\\
&\left\{\hat{\mathcal{A}}_{|D|,\lambda}^2\frac{\mathcal{A}'_{|D|,\lambda}}{\sqrt{\lambda}}+\hat{\mathcal{A}}_{|D|,\lambda}^2\frac{\mathcal{A}_{|D|,\lambda}}{\sqrt{\lambda}}\|f_\lambda\|_K+\hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1})+\|f_\lambda\|_K\right\},
\end{aligned}$$

where c_2 is a constant independent of $|D|, d, \lambda$, or σ.

Now we estimate the norm of I_2 in (C.2) which can be expressed as

$$I_2=[(\lambda I+L_{K,\hat{D}})^{-1}-(\lambda I+L_{K,D})^{-1}]E_{\hat{D},\lambda,\sigma}+(\lambda I+L_{K,D})^{-1}(E_{\hat{D},\lambda,\sigma}-E_{D,\lambda,\sigma}).$$

Applying the identity $(\lambda I+L_{K,\hat{D}})^{-1}-(\lambda I+L_{K,D})^{-1}=(\lambda I+L_{K,\hat{D}})^{-1}(L_{K,D}-L_{K,\hat{D}})(\lambda I+L_{K,D})^{-1}$ again, the first term above can be bounded as

$$\begin{aligned}
&\left\|L_K^{1/2}[(\lambda I+L_{K,\hat{D}})^{-1}-(\lambda I+L_{K,D})^{-1}]E_{\hat{D},\lambda,\sigma}\right\|_K\\
&\leq\|L_K^{1/2}[(\lambda I+L_{K,\hat{D}})^{-1}\|\|L_{K,D}-L_{K,\hat{D}}\|\|(\lambda I+L_{K,D})^{-1}E_{\hat{D},\lambda,\sigma}\|_K.
\end{aligned}$$

Then we apply Lemma 4 and the same procedure we use for estimating the second term of I_1 to get

$$\begin{aligned}&\mathbb{E}\left[\left\|L_K^{1/2}[(\lambda I+L_{K,\hat{D}})^{-1}-(\lambda I+L_{K,D})^{-1}]E_{\hat{D},\lambda,\sigma}\right\|_K\right]\\&\leq c_3 d^{-\frac{\alpha}{2}}\left[\frac{1}{\lambda^{\frac{3}{2}}d^{\frac{\alpha}{2}}}+\lambda^{-\frac{1}{2}}\right]\cdot\left[\hat{\mathcal{A}}^2_{|D|,\lambda}+\hat{\mathcal{A}}_{|D|,\lambda}\right]\sigma^{-2p}(\lambda^{-(p+\frac{3}{2})}+\lambda^{-1}),\end{aligned}$$

where c_3 is a constant independent of $|D|, d, \lambda$, or σ.

In the same way, the second term of I_2 can be bounded as

$$\begin{aligned}&\left\|L_K^{1/2}(\lambda I+L_{K,D})^{-1}(E_{\hat{D},\lambda,\sigma}-E_{D,\lambda,\sigma})\right\|_K\\&\leq\left\|L_K^{1/2}(\lambda I+L_{K,D})^{-1}\right\|\left(\|E_{\hat{D},\lambda,\sigma}\|_K+\|E_{D,\lambda,\sigma}\|_K\right)\\&\leq\left\|(\lambda I+L_K)^{1/2}(\lambda I+L_{K,D})^{-1/2}\right\|\ \left\|(\lambda I+L_{K,D})^{-1/2}\right\|\left(\|E_{\hat{D},\lambda,\sigma}\|_K+\|E_{D,\lambda,\sigma}\|_K\right).\end{aligned}$$

Then we apply Lemmas 4 and 6 and estimate the expected value of the norm as

$$\mathbb{E}\left[\left\|L_K^{1/2}(\lambda I+L_{K,D})^{-1}(E_{\hat{D},\lambda,\sigma}-E_{D,\lambda,\sigma})\right\|_K\right]\leq c_4\hat{\mathcal{A}}_{|D|,\lambda}\sigma^{-2p}\left[\lambda^{-(p+1)}+\lambda^{-1/2}\right],$$

where c_4 is a constant independent of $|D|, d, \lambda$, or σ.

Combining all the above bounds for the two terms of I_1 and two terms of I_2, we know that with $\bar{C}=\max\{c_1,c_2,c_3,c_4\}$, the desired bound holds true. The proof of Proposition 2 is complete. ■

References

[1] Frank Bauer, Sergei Pereverzev, Lorenzo Rosasco. On regularization algorithms in learning theory. Journal of complexity, 23.1, 52-72, 2017.

[2] Alain Berlinet, Christine Thomas-Agnan. Reproducing kernel Hilbert spaces in probability and statistics. Springer Science, Business Media, 2011.

[3] Xiangyu Chang, Shao-Bo Lin, Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. Journal of Machine Learning Research, 18.1, 1493-1514, 2017.

[4] Andreas Christmann, Ingo Steinwart. Consistency and robustness of kernel-based regression in convex risk minimization. Bernoulli 13.3, 799-819, 2007.

[5] Andreas Christmann, Arnout Van Messem. Bouligand derivatives and robustness of support vector machines for regression. Journal of Machine Learning Research 9, 915-936, 2018.

[6] Ernesto De Vito, Sergei Pereverzyev, Lorenzo Rosasco. Adaptive kernel methods using the balancing principle. Foundations of Computational Mathematics, 10.4, 455-479, 2010.

[7] Florian Dumpert, Andreas Christmann. Universal consistency and robustness of localized support vector machines. Neurocomputing, 315, 96-106, 2018.

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

[9] Heinz Werner Engl, Martin Hanke, Andreas Neubauer. Regularization of inverse problems. Vol. 375. Springer Science, Business Media, 1996.

[10] J. Fan, T. Hu, Q. Wu and D. X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmonic Anal. **41** (2016), 164–189.

[11] Cucker Felipe, Ding-Xuan Zhou. Learning theory: an approximation theory viewpoint. Vol. 24. Cambridge University Press, 2007.

[12] Zhiying Fang, Zheng-Chu Guo, Ding-Xuan Zhou. Optimal learning rates for distribution regression. Journal of Complexity, 56, 101426, 2020.

[13] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, Bernhard Scholkopf. Kernel measures of conditional dependence. Advances in Neural Information Processing Systems, 20, 489-496, 2007.

[14] Yunlong Feng, Xiaolin Huang, Lei Shi, Yuning Yang, Johan A. K. Suykens. Learning with the maximum correntropy criterion induced losses for regression. Journal of Machine Learning Research, 16, 993-1034, 2015.

[15] Yunlong Feng, Jun Fan, Johan AK Suykens. A Statistical Learning Approach to Modal Regression. Journal of Machine Learning Research, 21.2, 1-35, 2020.

[16] Yunlong Feng, Qiang Wu. Learning under $(1+\epsilon)$-moment conditions. Applied and Computational Harmonic Analysis, 49.2, 495-520, 2020.

[17] Yunlong Feng, Qiang Wu. A Statistical Learning Assessment of Huber Regression. arXiv preprint arXiv:2009.12755, 2020.

[18] Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Scholkopf, Alexander J. Smola. A kernel method for the two-sample-problem. Advances in Neural Information Processing Systems, 19, 513-520, 2006.

[19] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Scholkopf, Alexander J. Smola. A kernel statistical test of independence. Advances in Neural Information Processing Systems, 20, 585-592, 2007.

[20] Zheng-Chu Guo, Shao-Bo Lin, Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. Inverse Problems, 33.7, 074009, 2017.

[21] Zheng-Chu Guo, Ting Hu, Lei Shi. Gradient descent for robust kernel-based regression. Inverse Problems 34.6, 065009, 2018.

[22] Zheng-Chu Guo, Lei Shi, Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. Journal of Machine Learning Research 18.1, 4237-4261, 2017.

[23] T. Hu, J. Fan, Q. Wu and D. X. Zhou, Learning theory approach to minimum error entropy criterion, J. Machine Learning Research **14** (2013), 377–397.

[24] Ting Hu, Qiang Wu, and Ding-Xuan Zhou. Distributed kernel gradient descent algorithm for minimum error entropy principle. Applied and Computational Harmonic Analysis 49.1, 229-256, 2020.

[25] Fusheng Lv, Jun Fan. Optimal learning with Gaussians and correntropy loss. Analysis and Applications, 19.1, 107-124, 2019.

[26] Weifeng Liu, Puskal P. Pokharel, and Jose C. Príncipe. Correntropy: properties and applications in non-Gaussian signal processing. IEEE Transactions on Signal Processing 55.11, 5286-5298, 2017.

[27] Nicole. Mücke. Stochastic gradient descent meets distribution regression. arXiv preprint arXiv:2010.12842 (2020).

[28] Barnabás Póczos, Alessamdro Rinaldo, Aarti Singh, Larry Wasserman. Distribution-free distribution regression. Artificial Intelligence and Statistics PMLR, 507-515, 2013.

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

[30] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, Bharath K. Sriperumbudur. Two-stage sampled learning theory on distributions. In Artificial Intelligence and Statistics, 948-957, 2015.

[31] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, Arthur Gretton. Learning theory for distribution regression. Journal of Machine Learning Research, 17.1, 5272-5311, 2016.

[32] Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, Arthur Gretton. Minimax-optimal distribution regression. In International Society for NonParametric Statistics (ISNPS) Conference, 2016.

[33] Steve Smale, Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. Constructive approximation, 26.2, 153-172, 2007.

[34] Steve Smale, Ding-Xuan Zhou. Shannon sampling II: Connections to learning theory. Applied and Computational Harmonic Analysis 19.3, 285-302, 2005.

[35] Steve Smale, Ding-Xuan Zhou. Shannon sampling and function reconstruction from point values. Bulletin of the American Mathematical Society 41.3: 279-305, 2004.

[36] Ingo Steinwart, Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. Bernoulli 17.1, 211-225, 2011.

[37] Ingo Steinwart, Andreas Christmann. Support vector machines. Springer Science and Business Media, 2008.

[38] Steinwart, Ingo, Don R. Hush, and Clint Scovel. Optimal Rates for Regularized Least Squares Regression. COLT. 79-93, 2009.

[39] Cheng Wang, Ting Hu. Online minimum error entropy algorithm with unbounded sampling. Analysis and Applications 17.02, 293-322, 2019.