Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 34(1): 291–317.
DOI: 10.3934/era.2026014
Received: 02 November 2025
Revised: 05 December 2025
Accepted: 17 December 2025
Published: 09 January 2026

Research article

Coefficient-based regularized distribution regression under the moment conditions

Qin Guo[*]**and Shuli Liu**

School of Science, Shandong Jianzhu University, Jinan 250101, China

* **Correspondence:** Email: guoqin_1985@163.com.

Abstract: In this paper, we investigated the coefficient-based regularized distribution regression for data generated by unbounded sampling processes. The algorithm adopts a two-stage sampling framework: the first-stage sample consists of probability distributions, from which the second-stage sample is drawn. A rigorous capacity-dependent convergence analysis was conducted under more general conditions, and its performance was comparable to that of one-stage sampling learning. Regularization was imposed on the coefficients and the kernel K was permitted to be indefinite. The important feature of this algorithm is that it can improve the saturation effect suffered by classical kernel ridge regression (KRR). Notably, the output sample values were assumed to satisfy a moment condition (rather than the stricter uniform boundedness constraint common in related works). We derived the convergence error bounds via the novel integral operator techniques, and further established the minimax optimal learning rates of the algorithm, which were comparable to those achieved under bounded sampling settings.

Keywords: coefficient-based regularized distribution regression; moment hypothesis; integral operator; learning rates

1. Introduction and main results

1.1. Introduction

In the study, we investigate the learning problem of distribution regression (DR) under a two-stage sampling framework, where the aim is to predict real-valued responses from probability distributions [1–4]. We start by providing a brief introduction to the concept of distribution regression. Let $\widetilde{D} = \{(x_i, y_i)\}_{i=1}^m$ be a dataset, where x_i represents a probability distribution and each corresponding label y_i takes values in $\mathbb{R}^d$. The pair (x_i, y_i) is independent and identically distributed (i.i.d.) from a meta distribution $\mathcal{X}$. However, it is not feasible to directly observe the distributions x_i. Instead, we obtain samples $x_{i,1}, \ldots, x_{i,N_i}$, independently and identically sampled from x_i. Consequently, we can write $\{(\{x_{i,n}\}_{n=1}^{N_i}, y_i)\}_{i=1}^{m}$. Since $x_{i,j}$ is sampled from x_i and each x_i itself comes from $\mathcal{X}$, the hierarchical sampling procedure can be referred to as two-stage sampling. The objective is to predict y_{m+1} using observations $x_{m+1,1}, \ldots, x_{m+1,N_{m+1}}$ generated from a previously unknown distribution x_{m+1}.

By using distribution regression methods, we can efficiently address problems that traditional regression methods struggle with, such as point estimation tasks lacking analytical solutions. The approach not only expands the scope of machine learning applications but also enhances the ability to model complex datasets. Additionally, by mean embedding [5], distributions are embedded into the reproducing kernel Hilbert space (RKHS), enabling us to take advantage of established machine learning algorithms to analyze and learn from these distributions, thus overcoming the limitations of traditional regression analysis.

The Mercer kernel K is defined on $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and the induced RKHS is $(\mathcal{H}_K, \|\cdot\|_K)$, with the corresponding embedding given by

$$\mu_x = \int_X K(\cdot, s) dx(s) \in \mathcal{H}_K. \tag{1.1}$$

Let the embedding set of the distributions be denoted as

$$X_\mu = \{\mu_x, x \in X\} \subset \mathcal{H}_K,$$

which forms a separable compact set consisting of continuous functions defined on $\mathcal{X}$. In particular, when K is a characteristic kernel, for distributions x and x', $\|\mu_x - \mu_{x'}\|_{\mathcal{H}_K} = 0$ if and only if $x = x'$. This injective property ensures that the mean embedding represents the underlying probability distribution within the RKHS, thereby enabling a rigorous functional analysis framework for addressing distribution regression problems.

In the least squares regression setting, X_μ and ρ are, respectively, the input space and the Borel probability measure on $Z = X_\mu \times Y$. $\rho(\cdot|\mu_x)$ is the conditional distribution given μ_x. We provide the expected risk for a function $f : X_\mu \to Y$:

$$\mathcal{E}(f) = \int_Z (f(\mu_x) - y)^2 d\rho. \tag{1.2}$$

The regression function $f_\rho : X_\mu \to Y$ is given by

$$f_\rho(\mu_x) = \int_Y y d\rho(y|\mu_x), \quad \mu_x \in X_\mu, \tag{1.3}$$

which has been proven to be the global minimizer of the expected risk. Hence, the primary goal of DR is to construct an efficient approximation of f_ρ using $\hat{D} = \left\{\left(\{x_{i,j}\}_{j=1}^{N}, y_i\right)\right\}_{i=1}^{m}$ obtained by a two-stage i.i.d. sampling procedure.

Next, let us review the classical kernel ridge regression (KRR) method [6–8] for distribution learning. The KRR scheme for distribution regression [9, 10], which is set in a RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ induced by a Mercer kernel K, is defined as follows:

$$f_{\hat{D}}^K = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(\mu_{\hat{x}_i}) - y_i)^2 + \lambda \|f\|_K^2 \right\}, \tag{1.4}$$

where $\lambda > 0$ is a regularization parameter. For the given samples $\{x_{i,j}\}_{j=1}^N$, the empirical distribution $\hat{x}_i$ is defined as $\hat{x}_i = \frac{1}{N}\sum_{j=1}^N \delta_{x_{i,j}}$, and its mean embedding is given by $\mu_{\hat{x}_i} = \frac{1}{N}\sum_{j=1}^N K(\cdot, x_{i,j})$. Let the sample-dependent hypothesis space be given by

$$\mathcal{H}_{K,\hat{D}} = \left\{ f_\alpha \in \mathcal{H}_K \mid f_\alpha = \sum_{i=1}^m \alpha_i K(\cdot, \mu_{\hat{x}_i}), \alpha_1, \cdots, \alpha_m \in \mathbb{R} \right\}.$$

Based on representer theorem, we can uniquely determine the estimator $f_{\hat{D}}^K$ as

$$f_{\hat{D}}^K(\mu_x) = \sum_{i=1}^m \alpha_{\hat{D},i}^K K(\mu_x, \mu_{\hat{x}_i}) \quad \text{with} \quad \alpha_{\hat{D}}^K = (\lambda m \mathbb{I}_m + \hat{\mathbb{K}}_m)^{-1}\mathbf{y}, \quad \mu_x \in X_\mu,$$

where $\mathbb{I}_m$ represents the identity matrix on the space $\mathbb{R}^m \times \mathbb{R}^m$, the vector $\mathbf{y} = (y_1, \cdots, y_m)^T \in \mathbb{R}^m$, and $\hat{\mathbb{K}}_m = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})]_{i,j=1}^m$.

KRR relies on positive semi-definite kernels, but in real-world scenarios, there may exist asymmetric, negative, or non-metric proximities (such as in graphs, texts, etc.), where positive semi-definite kernels are not applicable. So in this paper, we consider learning by indefinite kernels $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which are only required to be continuous and bounded. This relaxation enables the use of indefinite kernels when prior knowledge is available and one aims to fit the data to a specific trend. For instance, the sigmoid kernel has been successfully applied in support vector machines [11], while fractional power polynomials have seen effective use in face recognition tasks [12]. Since there is no representation theorem for such kernels, we adopt the coefficient-based regularization scheme, that is, applying regularization to the coefficients $\{\alpha_i\}$ [13,14]. Thus, the distribution regression algorithm with coefficient-based regularization is given by

$$f_{\hat{D}} = f_{\alpha_{\hat{D}}} \quad \text{where} \quad \alpha_{\hat{D}} = \arg\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^m (f_\alpha(\mu_{\hat{x}_i}) - y_i)^2 + \lambda \Omega(\alpha) \right\}, \tag{1.5}$$

with $\Omega(\alpha) = \alpha^T \hat{\mathbb{K}}_m \alpha$.

Algorithm (1.5) has several advantages [15]. First, it directly corresponds to a finite-dimensional optimization problem and can be easily adapted to other algorithms. Second, coefficient regularization makes it possible to use indefinite kernels, which has already been successfully applied in support vector machines, face recognition, and wavelet analysis. Third, the parameter q in the ℓ_q-coefficient regularization term ($1 \le q \le 2$) can be chosen according to the properties to be estimated [16], such as smoothness, computational complexity, and sparsity. For example, ℓ_2-regularization yields smooth estimators [17, 18], whereas ℓ_1-regularization produces sparse estimators. Here, we consider $\Omega(\alpha) = m\|\alpha\|_2^2$:

$$f_{\hat{D}} = f_{\alpha_{\hat{D}}} \quad \text{where} \quad \alpha_{\hat{D}} = \arg\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{1}{m} \sum_{i=1}^m (f_\alpha(\mu_{\hat{x}_i}) - y_i)^2 + \lambda m \|\alpha\|_2^2 \right\}. \tag{1.6}$$

The explicit expression of the estimator $f_{\hat{D}}$ thus follows:

$$f_{\hat{D}}(\mu_x) = \sum_{i=1}^m \alpha_{\hat{D},i} K(\mu_x, \mu_{\hat{x}_i}) \quad \text{with} \quad \alpha_{\hat{D}} = (\lambda m \mathbb{I}_m + \hat{\mathbb{K}}_m^T \hat{\mathbb{K}}_m)^{-1} \hat{\mathbb{K}}_m^T \mathbf{y}. \tag{1.7}$$

The error performance of algorithm (1.6) was studied in [19] and [20]. In [19], the authors carried out the error analysis of algorithm (1.6) with the kernel being positive semi-definite. In [20], they first derived finite sample bounds and further presented explicit learning rates for DR with indefinite kernels. The error results in [19] and [20] were obtained under the uniform boundedness assumption of output data, that is, $|y| \leq M$ almost surely for some constant $M \geq 0$, which means $\rho(\cdot|x)$ is supported on $[-M, M]$ for almost every $x \in X$. However, such an assumption often deviates from real-world scenarios, especially in cases where the output variables follow Gaussian distributions or other unbounded distributions. In recent years, research on learning algorithms under unbounded sampling has received significant attention. For example, the authors in [21–24] derived the error bounds of the learning algorithms with the following moment hypothesis.

Assumption 1. *For some constants $C > 0$ and $M \geq 1$, it holds that*

$$\int_Y |y|^l d\rho(y|\mu_x) \leq Cl!M^l, \tag{1.8}$$

for all $\mu_x \in X_\mu$, $2 \leq l \in \mathbb{N}$.

Condition (1.8) is more general than the bounded output assumption and satisfied if the distribution $\rho(\cdot|\mu_x)$ is Gaussian or sub-Gaussian. A simple computation verifies (1.8) for Gaussian noise.

Example 1. *Let $B > 0$ and $B_0 > 0$. If for each $x \in X$, $|f_\rho(x)| \leq B$ and the condition distribution $\rho(\cdot|\mu_x)$ is a normal distribution with variance σ_x^2 bounded by B_0, then (1.8) is satisfied with $M = \max\{\sqrt{2B_0, B}\}$ and $C = 4$.*

In this paper, our main goal is to derive the learning rates of algorithm (1.6), which are based on the indefinite kernels under the assumption (1.8) and some mild assumptions on $\mathcal{H}_K$ and f_ρ presented below.

1.2. Assumptions and main results

Throughout this paper we always assume that the kernel K satisfies the boundedness, that is, $B_K = \sup_{s \in \widetilde{X}} K(s, s) < \infty$ almost surely. The boundedness for the kernel K ensures the existence of mean embeddings for all probability distributions on $\mathcal{X}$. All the constants in this paper are independent of m, N, δ, or γ.

Denote by $L^2_{\rho_{X_\mu}}$ the Hilbert space of the square integrable functions defined on X_μ with respect to the measure ρ_{X_μ}, where ρ_{X_μ} is the marginal measure of ρ on X_μ and the norm $\|\cdot\|_{\rho_{X_\mu}}$ induced by

$$\langle f, g \rangle_{\rho_{X_\mu}} = \int_{X_\mu} f(\mu_x) g(\mu_x) d\rho_{X_\mu}(\mu_x).$$

For an indefinite kernel K, the integral operator $L_K : L^2_{\rho_{X_\mu}} \to L^2_{\rho_{X_\mu}}$ is further defined as follows:

$$L_K f(\cdot) = \int_{X_\mu} K(\cdot, \mu_x) f(\mu_x) d\rho_{X_\mu}(\mu_x), \quad \forall f \in L^2_{\rho_{X_\mu}}. \tag{1.9}$$

Its adjoint operator L_K^* is given by

$$L_K^* f(\cdot) = \int_{X_\mu} K(\mu_x, \cdot) f(\mu_x) d\rho_{X_\mu}(\mu_x), \quad \forall f \in L^2_{\rho_{X_\mu}}. \tag{1.10}$$

Since X_μ is compact and K is continuous, both L_K and its adjoint L_K^* are compact operators. When K is positive definite, for given $r > 0$, the fractional power operator L_K^r is well defined. Let RKHS $\mathcal{H}_K$ associated with the Mercer kernel K be the closure of the linear span of the set of functions $\{K(\cdot, \mu_x), \mu_x \in X_\mu\}$ with the inner product

$$\langle K(\cdot, \mu_x), K(\cdot, \mu_{x'})\rangle = K(\mu_x, \mu_{x'}), \quad \forall \mu_x, \mu_{x'} \in X_\mu.$$

The reproducing property in $\mathcal{H}_K$ can be formulated as:

$$f(\mu_x) = \langle f, K(\cdot, \mu_x)\rangle_K. \tag{1.11}$$

It follows that

$$\|f\|_\infty \leq \sup_{\mu_x \in X_\mu} \sqrt{K(\mu_x, \mu_x)}\|f\|_K. \tag{1.12}$$

Thus $\mathcal{H}_K$ can be embedded into $C(X)$, equipped with the norm $\|f\|_\infty = \sup_{\mu_x \in X_\mu} |f(\mu_x)|$. In particular, $L_K^{\frac{1}{2}}$ is an isomorphism from $L^2_{\rho_{X_\mu}}$ to $\mathcal{H}_K$, i.e., $L_K^{\frac{1}{2}} f \in \mathcal{H}_K$ and $\|f\|_{\rho_{X_\mu}} = \|L_K^{\frac{1}{2}} f\|_K$ for all $f \in L^2_{\rho_{X_\mu}}$. Then by the spectral theorem, we have

$$L_K = \sum_{\ell \geq 1} \sigma_\ell \phi_\ell \otimes \psi_\ell, \qquad L_K^* = \sum_{\ell \geq 1} \sigma_\ell \psi_\ell \otimes \phi_\ell,$$

where $\{\sigma_\ell\}_{\ell \geq 1}$ denotes the singular values of L_K, and $\{\phi_\ell\}_{\ell \geq 1}$ and $\{\psi_\ell\}_{\ell \geq 1}$ are orthonormal systems of $L^2_{\rho_{X_\mu}}$ satisfying $L_K \psi_\ell = \sigma_\ell \phi_\ell$ and $L_K^* \phi_\ell = \sigma_\ell \psi_\ell$. The subsequent assumption with respect to $\{\sigma_\ell, \phi_\ell, \psi_\ell\}_{\ell \geq 1}$ is introduced in [15].

Assumption 2. *There exists a constant $\kappa \geq 1$ such that*

$$\sup_{\mu_x \in X_\mu} \sum_{\ell \geq 1} \sigma_\ell \phi_\ell^2(\mu_x) \leq \kappa^2 \quad \textit{and} \quad \sup_{\mu_x \in X_\mu} \sum_{\ell \geq 1} \sigma_\ell \psi_\ell^2(\mu_x) \leq \kappa^2. \tag{1.13}$$

The continuity of the kernel function K yields that if $\sigma_k > 0$, ϕ_ℓ and ψ_ℓ can be chosen to be continuous on X_μ. Consequently, the terms $\sum_{\ell \geq 1} \sigma_\ell \phi_\ell^2(x)$ and $\sum_{\ell \geq 1} \sigma_\ell \psi_\ell^2(x)$ in (1.13) are defined point-wise on X_μ. The hypothesis is satisfied for some specific and popular kernels. In particular, when K is a Mercer kernel, Assumption 2 holds with $\kappa = \max\left\{1, \sup_{\mu_x \in X_\mu} \sqrt{K(\mu_x, \mu_x)}\right\}$ by Mercer's theorem. Moreover, condition (1.13) holds if the kernel K admits a Kolmogorov decomposition (i.e., $K = K_+ + K_-$ for positive semi-definite K_+ and K_-). This decomposition is guaranteed for kernels that are reproducing kernels of a reproducing kernel Krein space, which includes common indefinite kernels such as linear combinations of positive semi-definite kernels and conditionally positive definite kernels, see [20] and the references therein for the details.

Under Assumption 2, two positive semi-definite kernels on $X_\mu \times X_\mu$ can be defined as follows:

$$K_0(\mu_x, \mu_{x'}) = \sum_{\ell \geq 1} \sigma_\ell \phi_\ell(\mu_x) \phi_\ell(\mu_{x'}) \quad \text{and} \quad K_1(\mu_x, \mu_{x'}) = \sum_{\ell \geq 1} \sigma_\ell \psi_\ell(\mu_x) \psi_\ell(\mu_{x'}), \tag{1.14}$$

which induce RKHSs $\mathcal{H}_{K_0}$ and $\mathcal{H}_{K_1}$. Note that the integral operators L_{K_0} and L_{K_1} have the same eigenvalues $\{\sigma_\ell\}_{\ell \geq 1}$. Also, it can be observed that

$$L_{K_0} = \sum_{\ell \geq 1} \sigma_\ell \phi_\ell \otimes \phi_\ell \quad \text{and} \quad L_{K_1} = \sum_{\ell \geq 1} \sigma_\ell \psi_\ell \otimes \psi_\ell.$$

The fractional power operator $L^r_{K_0}$ of L_{K_0} is also well-defined as

$$L^r_{K_0} = \sum_{\ell \geq 1} \sigma^r_\ell \phi_\ell \otimes \phi_\ell.$$

Its range space can be expressed as $L^r_{K_0}(L^2_{\rho_{X_\mu}}) = \left\{ f \in L^2_{\rho_{X_\mu}} : \sum_{\ell \geq 1} \frac{\langle f, \phi_\ell \rangle^2}{\sigma^{2r}_\ell} < \infty \right\}$. It yields that $L^r_{K_0}(L^2_{\rho_{X_\mu}}) \subseteq L^{r'}_{K_0}(L^2_{\rho_{X_\mu}})$, if $r > r'$, and $L^r_{K_0}(L^2_{\rho_{X_\mu}}) \subseteq \mathcal{H}_{K_0}$, if $r \geq \frac{1}{2}$. Throughout this paper, we require that f_ρ satisfies the regularity condition of order r.

Assumption 3. *Regularity condition. For $r > 0$,*

$$f_\rho = L^r_{K_0}(g_\rho), \quad \textit{for some } g_\rho \in L^2_{\rho_{X_\mu}}. \tag{1.15}$$

To measure the complexity of $\mathcal{H}_{K_0}$ and $\mathcal{H}_{K_1}$, we now provide the definition of the effective dimension $\mathcal{N}(\lambda)$ with respect to ρ_{X_μ}.

$$\mathcal{N}(\lambda) = \mathrm{Tr}[(\lambda I + L_{K_0})^{-1} L_{K_0}] = \mathrm{Tr}[(\lambda I + L_{K_1})^{-1} L_{K_1}] = \sum_{\ell \geq 1} (\sigma_\ell + \lambda)^{-1} \sigma_\ell. \tag{1.16}$$

To derive the explicit learning rates for algorithm (1.6), it is necessary to introduce the following assumption regarding the decay rate of $\{\sigma_\ell\}_{\ell \geq 1}$ to characterize $N(\lambda)$.

Assumption 4. *For constants $c_\alpha > 0$ and $\alpha > 1$, the inequality*

$$\sigma_\ell \leq c_\alpha \ell^{-\alpha}, \quad \forall \ell \geq 1 \tag{1.17}$$

holds.

Under Assumption 4, the following capacity condition

$$\mathcal{N}(\lambda) \leq \frac{\alpha c_\alpha}{\alpha - 1} \lambda^{-\frac{1}{\alpha}} \tag{1.18}$$

holds, see [25].

To estimate $\|\hat{S}^*_0 \mathbf{y} - S^*_0 \mathbf{y}\|_{K_0}$, $\|T^{\hat{\mathbf{x}}}_0 - T^{\mathbf{x}}_0\|$ and $\|\hat{T}^{\hat{\mathbf{x}}}_0 - \hat{T}^{\mathbf{x}}_0\|$ appear in the proof, and we recall the following Hölder continuity condition for the kernel K.

Assumption 5. *Hölder continuity. We say that the mappings $K_{(\cdot)} : X_\mu \to \mathcal{H}_{K_0}$ ($K_{\mu_x} = K(\cdot, \mu_x)$) and $K^*_{(\cdot)}$ ($K^*_{\mu_x} = K(\mu_x, \cdot)$) are Hölder continuous, if there exist constants $h_1, h_2 \in (0, 1]$ and $L > 0$, such that*

$$\|K_{\mu_a} - K_{\mu_b}\|_{K_0} \leq L \|\mu_a - \mu_b\|^{h_1}_{\mathcal{H}_k} \quad \textit{and} \quad \|K^*_{\mu_a} - K^*_{\mu_b}\|_{K_1} \leq L \|\mu_a - \mu_b\|^{h_2}_{\mathcal{H}_k}, \forall \mu_a, \mu_b \in X_\mu. \tag{1.19}$$

In particular, for positive semi-definite kernels, we have $h_1 = h_2$ and $K_{(\cdot)} = K^*_{(\cdot)}$. Some examples of Hölder continuous kernels are listed in Table 1; these kernels are the natural extensions to distributions of the Gaussian, exponential, Cauchy, generalized t-student, and inverse multiquadratic kernels, see [26] and the references therein. However, in the case of indefinite kernels, it has been proved in Lemma 4.1 of [14] that there exists a linear isometric operator $U \in \mathcal{B}(\mathcal{H}_{K_1}, \mathcal{H}_{K_0})$ such that

$$K^*_{\mu_x} = K(\mu_x, \cdot) = U^* K_0(\mu_x, \cdot) \in \mathcal{H}_{K_1}, \quad K_{\mu_x} = K(\cdot, \mu_x) = U K_1(\cdot, \mu_x) \in \mathcal{H}_{K_0}, \forall \mu_x \in X_\mu \tag{1.20}$$

hold, where $U^* \in \mathcal{B}(\mathcal{H}_{K_0}, \mathcal{H}_{K_1})$ denotes the adjoint operator of U and the Banach space $\mathcal{B}(\mathcal{H}, \mathcal{H}')$ consists of all bounded linear operators from Hilbert space $\mathcal{H}$ to Hilbert space $\mathcal{H}'$. Based on this fact, we can obtain the Hölder continuity condition for indefinite kernels.

Table 1. Hölder continuity kernels.

K_G	K_e	K_C	K_t	K_i
$e^{-\frac{\|\mu_a-\mu_b\|_H^2}{2\theta^2}}$	$e^{-\frac{\|\mu_a-\mu_b\|_H}{2\theta^2}}$	$\left(1+\|\mu_a-\mu_b\|_H^2/\theta^2\right)^{-1}$	$\left(1+\|\mu_a-\mu_b\|_H^\theta\right)^{-1}$	$\left(\|\mu_a-\mu_b\|_H^2+\theta^2\right)^{-\frac{1}{2}}$
$h=1$	$h=\frac{1}{2}$	$h=1$	$h=\frac{\theta}{2}\ (\theta\le 2)$	$h=1$

Now we state the error bound of algorithm (1.6) under the above assumptions.

Theorem 1. *The estimator $f_{\hat{D}}$ is given by algorithm (1.6) with the indefinite and continuous kernel K. Assume $L_{K_0}^{-r} f_\rho \in L^2_{\rho_{X_\mu}}$ for $r \geq \frac{1}{2}$ and K satisfies (1.13) and (1.19). Suppose the moment hypothesis (1.8) and the capacity condition (1.18) hold. Let $h = \min\{h_1, h_2\}$ and $h' = \max\{h_1, h_2\}$. If the regularization parameter λ satisfies $\kappa^4 c(t)\log^2(13m/\delta)m^{-2} \leq \lambda \leq \kappa^4$ with $c(t) = 4(1+t/3)^2 t^{-4}$ for $t \in (0, 1/4]$, then for any $\gamma > 0$ and $\delta \in (0,1)$, with confidence $1-\delta-e^{-\gamma}$, $\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}}$ is bounded by*

$$\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \leq \begin{cases} c_1\left(\log\frac{13}{\delta}\right)^{2r+4}\mathcal{P}_{m,\lambda}^{2r+4}\left[\mathcal{B}_{m,\lambda}+\lambda^{\frac{r}{2}}+\mathcal{D}_{m,N,\lambda}(1+m^{-\frac{1}{2}})\right] & \\ \times(1+\lambda^{-\frac{1}{4}}\mathcal{D}_{m,N,\lambda})\left(1+\|g_\rho\|_{\rho_{X_\mu}}+\lambda^{\frac{r}{2}-\frac{1}{4}}\right), & \text{if } \frac{1}{2}\leq r\leq\frac{3}{2}, \\ c_2\left(\log\frac{13}{\delta}\right)^{5}\mathcal{P}_{m,\lambda}^{5}\left[\mathcal{B}_{m,\lambda}+\lambda^{\frac{1}{4}}m^{-\frac{1}{2}}+\lambda^{\min\{1,\frac{r}{2}\}}+\mathcal{D}_{m,N,\lambda}(1+m^{-\frac{1}{2}})\right] & \\ \times(1+\lambda^{-\frac{1}{4}}\mathcal{D}_{m,N,\lambda})\left(2+\|g_\rho\|_{\rho_{X_\mu}}+\lambda^{\min\{\frac{3}{4},\frac{r}{2}-\frac{1}{4}\}}\right), & \text{if } r>\frac{3}{2}, \end{cases} \tag{1.21}$$

where

$$\mathcal{B}_{m,\lambda} = \frac{2\kappa}{\sqrt{m}}\left\{\frac{\kappa}{\sqrt{m}\lambda^{1/4}} + \sqrt{\mathcal{N}(\lambda^{1/2})}\right\}, \tag{1.22}$$

$$\mathcal{P}_{m,\lambda} = 1 + \lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}, \tag{1.23}$$

and

$$\mathcal{D}_{m,N,\lambda} = \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} (1 + \sqrt{\log m + \gamma})^{h'}. \tag{1.24}$$

Corollary 1. *Under the assumptions of Theorem 1 and Assumption 4, if we take $\lambda = \kappa^4 m^{-\beta}$ with*

$$\beta = \begin{cases} \frac{2\alpha}{2\alpha r+1}, & \text{if } \frac{1}{2}\leq r\leq 2, \\ \frac{2\alpha}{4\alpha+1}, & \text{if } r>2, \end{cases} \tag{1.25}$$

$N = m^\zeta \log m$ with

$$\zeta = \begin{cases} \frac{3\alpha+2\alpha r}{h(2\alpha r+1)}, & \text{if } \frac{1}{2}\leq r\leq 2, \\ \frac{7\alpha}{h(4\alpha+1)}, & \text{if } r>2, \end{cases} \tag{1.26}$$

and m satisfies

$$m \geq \max\left\{[4c(t)\log^2(13/\delta)]^{\frac{1}{2-\beta}}, [2^{12}e^{-4}(2-\beta)^{-4}c^2(t)]^{\frac{1}{2-\beta}}\right\},$$

then for any $\delta \in (0,1)$, $\gamma > 0$, and $t \in (0, \frac{1}{4}]$, with confidence $1-\delta-e^{-\gamma}$, we have

$$\left\|f_{\hat{D}} - f_\rho\right\|_{\rho_{X_\mu}} \leq \begin{cases} \tilde{c}_1\left(\log\frac{13}{\delta}\right)^{2r+4}(1+\sqrt{1+\gamma})^{2h'}(\log m)^{h'-h}m^{-\frac{\alpha r}{2\alpha r+1}}, & \text{if } \frac{1}{2}\leq r\leq\frac{3}{2}, \\ \tilde{c}_2\left(\log\frac{13}{\delta}\right)^{5}(1+\sqrt{1+\gamma})^{2h'}(\log m)^{h'-h}m^{-\frac{\alpha\min\{r,2\}}{2\alpha\min\{r,2\}+1}}, & \text{if } r>\frac{3}{2}. \end{cases} \tag{1.27}$$

If the positive definite kernel K is used in algorithm (1.6), then the logarithmic term $\log m$ in the learning rate (1.27) can be removed while the assumption on the sample size m is less restrictive. Furthermore, due to the properties of positive semi-definite kernels such as $h_1 = h_2$ and $K_{(\cdot)} = K^*_{(\cdot)}$ in Assumption 5, the learning rate can be improved to the optimal order $O(m^{-\frac{\alpha \min\{r,2\}}{2\alpha \min\{r,2\}+1}})$ for $r > \frac{1}{2}$ in a minimax sense. Then we provide the following error bound of algorithm (1.6) for the positive definite kernel including $0 < r < \frac{1}{2}$.

Theorem 2. *Under the same assumptions of Theorem 1 and $L_{K_0}^{-r} f_\rho \in L^2_{\rho_{X_\mu}}$ for $r > 0$, if K is positive semi-definite, then for any $\gamma > 0$ and $\delta \in (0,1)$, with confidence $1 - \delta - e^{-\gamma}$, we have*

$$\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \le \begin{cases} d_0 \mathcal{A}^3_{m,N,\lambda} \left(\log \frac{7}{\delta}\right)^3 \left[\lambda^{\frac{r}{2}} + \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} (1 + \sqrt{\gamma + \log m})^h (m^{-\frac{1}{2}} + 1)\right], & \text{if } 0 < r < \frac{1}{2}, \\ d_1 \mathcal{A}^{2\max\{1,r\}}_{m,N,\lambda} \left(\log \frac{7}{\delta}\right)^{\max\{3,2r+1\}} & \\ \times \left[\mathcal{B}_{m,\lambda} + \lambda^{\frac{r}{2}} + \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} (1 + \sqrt{\gamma + \log m})^h (m^{-\frac{1}{2}} + 1)\right], & \text{if } \frac{1}{2} \le r \le \frac{3}{2}, \\ d_2 \mathcal{A}^2_{m,N,\lambda} \left(\log \frac{9}{\delta}\right)^3 & \\ \times \left[\mathcal{B}_{m,\lambda} + \lambda^{1/4} m^{-1/2} + \lambda^{\min\{1,\frac{r}{2}\}} + \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} (1 + \sqrt{\gamma + \log m})^h (m^{-\frac{1}{2}} + 1)\right], & \text{if } r > \frac{3}{2}, \end{cases} \tag{1.28}$$

where

$$\mathcal{A}_{m,N,\lambda} = 1 + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda} + \kappa L \left(1 + \sqrt{\gamma + \log m}\right)^h \frac{2^{\frac{h+2}{2}} B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}} N^{\frac{h}{2}}}.$$

Corollary 2. *Under the assumptions of Theorem 2 and Assumption 4, if we take $\lambda = m^{-\beta}$ with*

$$\beta = \begin{cases} \frac{2\alpha}{\alpha+1}, & \text{if } 0 < r < \frac{1}{2}, \\ \frac{2\alpha}{2\alpha r+1}, & \text{if } \frac{1}{2} \le r \le 2, \\ \frac{2\alpha}{4\alpha+1}, & \text{if } r > 2, \end{cases} \tag{1.29}$$

$N = m^\zeta \log m$ with

$$\zeta = \begin{cases} \frac{3\alpha+2\alpha r}{h(\alpha+1)}, & \text{if } 0 < r < \frac{1}{2}, \\ \frac{3\alpha+2\alpha r}{h(2\alpha r+1)}, & \text{if } \frac{1}{2} \le r \le 2, \\ \frac{7\alpha}{h(4\alpha+1)}, & \text{if } r > 2, \end{cases} \tag{1.30}$$

and $m \ge 3$, then for any $\delta \in (0,1)$ and $\gamma > 0$, with confidence $1 - \delta - e^{-\gamma}$, we have

$$\left\|f_{\hat{D}} - f_\rho\right\|_{\rho_{X_\mu}} \le \begin{cases} \tilde{d}_0 \left(\log \frac{7}{\delta}\right)^3 (1 + \sqrt{1+\gamma})^{4h} m^{-\frac{\alpha r}{\alpha+1}}, & \text{if } 0 < r < \frac{1}{2}, \\ \tilde{d}_1 \left(\log \frac{7}{\delta}\right)^{\max\{3,2r+1\}} (1 + \sqrt{1+\gamma})^{h\max\{3,2r+1\}} m^{-\frac{\alpha r}{2\alpha r+1}}, & \text{if } \frac{1}{2} \le r \le \frac{3}{2}, \\ \tilde{d}_2 \left(\log \frac{9}{\delta}\right)^3 (1 + \sqrt{1+\gamma})^{3h} m^{-\frac{\alpha \min\{r,2\}}{2\alpha \min\{r,2\}+1}}, & \text{if } r > \frac{3}{2}. \end{cases} \tag{1.31}$$

Notice that the convergence rate is suboptimal when $0 < r < \frac{1}{2}$. It would be interesting to further improve the learning performance in the future. Detailed proofs for these results are presented in Sections 3 and 4.

1.3. Related work and discussions

In this section, we present a detailed discussion and explain our main contributions by comparing the learning rates of algorithm (1.6) with the previous results. A comparative summary of some of the literature is presented in Table 2.

Consider the positive semi-definite kernel case. We first recall that Theorem 5 in [9] provides a capacity dependent analysis of algorithm (1.4) under the uniform boundedness assumption of output sample values, which shows that by taking $N \geq m^{\frac{\alpha+2\alpha r}{h(2\alpha r+1)}} \log m$, $\|f_{\hat{D}}^K - f_\rho\|_{\rho_{X_\mu}}^2 \leq O(m^{-\frac{2\alpha r}{2\alpha r+1}})$ for $\frac{1}{2} < r \leq 1$. The above rate is the same as the result in Corollary 2 for $\frac{1}{2} \leq r \leq \frac{3}{2}$ but with less restriction on the second-stage sample size N. In particular, Theorem 9 in [9] obtained the capacity independent case $\|f_{\hat{D}}^K - f_\rho\|_{\rho_{X_\mu}}^2 \leq O(m^{-\frac{2r}{r+2}})$ for $0 < r \leq \frac{1}{2}$ with $N \geq m^{\frac{2(1+r)}{h(r+2)}} \log m$. For the purpose of comparison, we take $\alpha = 1$ in (1.18) on the capacity assumption condition and derive the learning rate $\|f_{\hat{D}}^K - f_\rho\|_{\rho_{X_\mu}}^2 \leq O(m^{-r})$ which is sharper than that in [9]. Note when $r > 1$, the rate in Corollary 2 keeps improving while algorithm (1.4) in [9] suffers from saturation and the rate stops improving. This explains why our rate obtained in Corollary 2 is better when $r > 1$.

As for algorithm (1.6), the capacity independent learning rate has been derived in [19] for $\frac{1}{2} < r < 2$ when $N = m^{\frac{7}{h(2r+1)}}$. Compared to the result in [19], our study can relax the restriction on the second stage sample size in distributed learning and extend to a more general setting including indefinite kernels and unbounded outputs. In addition, it should be pointed out that our convergence rate is the same as those obtained in [20] with indefinite kernels under the bounded sampling assumption without increasing the requirement for N.

Table 2. Comparison of rates and assumptions for kernel-based learning methods.

Literature comparison	Output condition	Kernel type	Range of r	Learning rate
Reference [9]	Uniform boundedness	Positive definite	$0 < r \leq \frac{1}{2}$	$O(m^{-\frac{2r}{r+2}})$
			$\frac{1}{2} < r \leq 1$	$O(m^{-\frac{2\alpha r}{2\alpha r+1}})$
			$r > 1$	$O(m^{-\frac{2\alpha}{2\alpha+1}})$
Reference [19]	Uniform boundedness	Positive definite	$\frac{1}{2} \leq r < \frac{3}{2}$	$O(m^{-\frac{4r^2}{2r+9}})$
			$\frac{3}{2} \leq r < 2$	$O(m^{-\frac{2r}{2r+1}})$
			$r \geq 2$	$O(m^{-\frac{4}{5}})$
Reference [20]	Uniform boundedness	Indefinite	$0 < r < \frac{1}{2}$	$O(m^{-\frac{2\alpha r}{\alpha+1}})$
			$\frac{1}{2} \leq r \leq \frac{3}{2}$	$O(m^{-\frac{2\alpha r}{2\alpha r+1}})$
			$r > \frac{3}{2}$	$O(m^{-\frac{2\alpha \min\{r,2\}}{2\alpha \min\{r,2\}+1}})$
Our work	Moment assumption (weaker)	Indefinite	$0 < r < \frac{1}{2}$	$O(m^{-\frac{2\alpha r}{\alpha+1}})$
			$\frac{1}{2} \leq r \leq \frac{3}{2}$	$O(m^{-\frac{2\alpha r}{2\alpha r+1}})$
			$r > \frac{3}{2}$	$O(m^{-\frac{2\alpha \min\{r,2\}}{2\alpha \min\{r,2\}+1}})$

We also present experimental results to demonstrate the generalization performance of distributed learning algorithms under varying noise levels. Specifically, we define the input space as $X = [0, 1]$ and adopt the kernel function $K(x, x') = 1 + \min(x, x')$. The target function $f(x)$ is defined as

$$f(x) = \begin{cases} x & 0 < x \leq 0.5, \\ 1 - x & 0.5 < x \leq 1. \end{cases}$$

Response variables are generated based on the following regression model:

$$y_i = f(x_i) + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ denotes Gaussian noise. To investigate the impact of noise on model performance, we compare the *noise-free case* (i.e., $\sigma_\epsilon = 0$) with three noisy scenarios: $\sigma_\epsilon = 0.01, 0.1, 0.25$. The generalization performance is evaluated using the *mean squared error (MSE)* on an independent test set $\{\tilde{x}_i\}_{i=1}^{N_t}$ (test set size $N_t = 1000$), which is defined as:

$$\mathrm{MSE} = \frac{1}{N_t} \sum_{i=1}^{N_t} (\hat{y}_i - f(\tilde{x}_i))^2 ,$$

where $\hat{y}_i$ denotes the predicted value of the model for the test input $\tilde{x}_i$, and $f(\tilde{x}_i)$ is the corresponding true value.

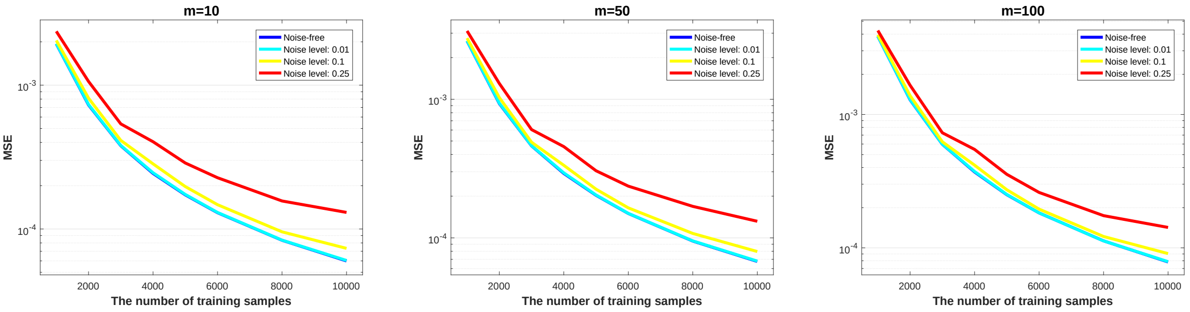

Figure 1. Generalization performance (MSE) of distributed learning algorithms under varying noise levels, numbers of local machines, and training sample sizes.

The relationship between MSE and the number of training samples, under varying noise levels and different numbers of local machines, is visualized in Figure 1. It can be observed that as the number of training samples increases, all curves exhibit a clear downward trend, indicating that enlarging the sample size effectively reduces the generalization error of the algorithm. Moreover, under low noise levels (i.e., $\sigma_\epsilon < 0.25$), the generalization performance of the algorithm is highly consistent with that in the noise-free scenario. In addition, the variation in the number of local machines m does not exert a significant impact on the generalization performance of the distributed learning algorithm under different noise levels. When m is large, the algorithm still exhibits better performance under low noise levels. This phenomenon verifies the good applicability and effectiveness of the moment-bounded assumption in distributed learning algorithms.

The remainder of the paper is organized as follows. In the second section, we will provide some preliminary results and establish the error decomposition of algorithm (1.6). The upper bounds of the terms in the error decomposition will be given in Sections 3 and 4. In the final section, we will prove the main results.

2. Preliminary results and error decomposition

To describe the properties of the integral operators mentioned in the previous section, we invoke the following lemma.

Lemma 2.1. *Under Assumption 2, the following properties hold:*

1) $\left\{\sqrt{\sigma_\ell}\phi_\ell : \sigma_\ell > 0\right\}$ *and* $\left\{\sqrt{\sigma_\ell}\psi_\ell : \sigma_\ell > 0\right\}$ *form orthonormal bases of* $\mathcal{H}_{K_0}$ *and* $\mathcal{H}_{K_1}$*, respectively.*
2) *The operator* L_K *belongs to* $\mathcal{B}(\mathcal{H}_{K_1}, \mathcal{H}_{K_0})$*, and its adjoint operator* L_K^* *belongs to* $\mathcal{B}(\mathcal{H}_{K_0}, \mathcal{H}_{K_1})$ *with* $\|L_K\| = \|L_K^*\| \leq \kappa^2$*.*
3) $L_{K_0} \in \mathcal{B}(\mathcal{H}_{K_0})$ *and* $L_{K_1} \in \mathcal{B}(\mathcal{H}_{K_1})$ *with* $\|L_{K_0}\| = \|L_{K_1}^*\| \leq \kappa^2$*.*
4) *There exists a linear isometry* $U \in \mathcal{B}(\mathcal{H}_{K_1}, \mathcal{H}_{K_0})$ *satisfying* $\phi_\ell = U\psi_\ell$ *for all* $\sigma_\ell \geq 0$*. Moreover, for any* $\mu_x \in X_\mu$*, it holds that* $K_{\mu_x} = K(\cdot, \mu_x) = UK_1(\cdot, \mu_x) \in \mathcal{H}_{K_0}$ *and* $K_{\mu_x}^* = K(\mu_x, \cdot) = U^*K_0(\mu_x, \cdot) \in \mathcal{H}_{K_1}$*, where* $U^* \in \mathcal{B}(\mathcal{H}_{K_0}, \mathcal{H}_{K_1})$ *denotes the adjoint operator of* U*.*

Before proceeding with the error decomposition, we introduce the following sampling operators and empirical integral operators. Define the sampling operator $S_q : \mathcal{H}_{K_q} \to \mathbb{R}^m$ associated to the sampling points $\mathbf{x} = \{\mu_{x_1}, \cdots, \mu_{x_m}\}$ for the first-stage sampling as follows:

$$S_q f = (f(\mu_{x_1}), \cdots, f(\mu_{x_m})), \quad \forall f \in \mathcal{H}_{K_q},$$

where $q \in \{0, 1\}$.
Let S_q^* be its scaled adjoint operator from $\mathbb{R}^m$ to $\mathcal{H}_{K_q}$ defined as:

$$S_q^*\alpha = \frac{1}{m}\sum_{i=1}^m \alpha_i K_q(\mu_{x_i}, \cdot), \quad \forall \alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}^m. \tag{2.1}$$

Similarly, the sampling operator $\hat{S}_q : \mathcal{H}_{K_q} \to \mathbb{R}^m$ for the second-stage sampling is given by

$$\hat{S}_q f = (f(\mu_{\hat{x}_1}), \cdots, f(\mu_{\hat{x}_m})), \quad \forall f \in \mathcal{H}_{K_q}.$$

Based on $\hat{\mathbf{x}} = \{\mu_{\hat{x}_1}, \cdots, \mu_{\hat{x}_m}\}$, its scaled adjoint operator $\hat{S}_q^* : \mathbb{R}^m \to \mathcal{H}_{K_q}$ is defined by

$$\hat{S}_q^*\alpha = \frac{1}{m}\sum_{i=1}^m \alpha_i K_q(\mu_{\hat{x}_i}, \cdot), \quad \forall \alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{R}^m. \tag{2.2}$$

In the setting of one-stage sampling, we can define the empirical integral operators as

$$\begin{aligned} T_0^{\mathbf{x}} &= \frac{1}{m}\sum_{i=1}^m K_0(\mu_{x_i}, \cdot) \otimes K_0(\mu_{x_i}, \cdot), \quad T_1^{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^m K_1(\mu_{x_i}, \cdot) \otimes K_1(\mu_{x_i}, \cdot), \\ T^{\mathbf{x}} &= \frac{1}{m}\sum_{i=1}^m K(\cdot, \mu_{x_i}) \otimes K_1(\mu_{x_i}, \cdot), \quad T_*^{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^m K(\mu_{x_i}, \cdot) \otimes K_0(\mu_{x_i}, \cdot). \end{aligned} \tag{2.3}$$

The empirical integral operators in the setting of two-stage sampling are similarly defined by

$$\begin{aligned} T_0^{\hat{\mathbf{x}}} &= \frac{1}{m}\sum_{i=1}^m K_0(\mu_{\hat{x}_i}, \cdot) \otimes K_0(\mu_{\hat{x}_i}, \cdot), \quad T_1^{\hat{\mathbf{x}}} = \frac{1}{m}\sum_{i=1}^m K_1(\mu_{\hat{x}_i}, \cdot) \otimes K_1(\mu_{\hat{x}_i}, \cdot), \\ T^{\hat{\mathbf{x}}} &= \frac{1}{m}\sum_{i=1}^m K(\cdot, \mu_{\hat{x}_i}) \otimes K_1(\mu_{\hat{x}_i}, \cdot), \quad T_*^{\hat{\mathbf{x}}} = \frac{1}{m}\sum_{i=1}^m K(\mu_{\hat{x}_i}, \cdot) \otimes K_0(\mu_{\hat{x}_i}, \cdot). \end{aligned} \tag{2.4}$$

From Lemma 2.1 and the definition of the aforementioned operators, it follows that

$$L_K = UL_{K_1} = L_{K_0}U, \quad L_K^* = U^*L_{K_0} = L_{K_1}U^*. \tag{2.5}$$

$$T^{\mathbf{x}} = UT_1^{\mathbf{x}} = US_1^*S_1, \quad T_*^{\mathbf{x}} = U^*T_0^{\mathbf{x}} = U^*S_0^*S_0. \tag{2.6}$$

$$T^{\hat{\mathbf{x}}} = UT_1^{\hat{\mathbf{x}}} = U\hat{S}_1^*\hat{S}_1, \quad T_*^{\hat{\mathbf{x}}} = U^*T_0^{\hat{\mathbf{x}}} = U\hat{S}_0^*\hat{S}_0. \tag{2.7}$$

Furthermore, $T_0^{\mathbf{x}}$ and $T_1^{\mathbf{x}}$ converge to the limits L_{K_0} and L_{K_1}, respectively. $T^{\mathbf{x}}$ and $T_*^{\mathbf{x}}$ can be viewed as approximations of L_K and L_K^*. As a kernel matrix, $\mathbb{K}_m = [K(\mu_{x_i}, \mu_{x_j})]_{i,j=1}^m$ satisfies $\mathbb{K}_m = mS_0US_1^*$ and $\hat{\mathbb{K}}_m = m\hat{S}_0U\hat{S}_1^*$.

In order to quantify the approximations of the integral operators and sampling operators, combining Lemma 2.1 with Assumption 5, the inequality (1.19) for all $(\mu_a, \mu_b) \in X_\mu \times X_\mu$ can be rewritten as follows:

$$\|K_1(\cdot, \mu_a) - K_1(\cdot, \mu_b)\|_{K_1} = \|K(\cdot, \mu_a) - K(\cdot, \mu_b)\|_{K_0} \le L\|\mu_a - \mu_b\|_{\mathcal{H}_k}^{h_1}, \tag{2.8}$$

$$\|K_0(\mu_a, \cdot) - K_0(\mu_b, \cdot)\|_{K_0} = \|K(\mu_a, \cdot) - K(\mu_b, \cdot)\|_{K_1} \le L\|\mu_a - \mu_b\|_{\mathcal{H}_k}^{h_2}, \tag{2.9}$$

which hold for all $(\mu_a, \mu_b) \in X_\mu \times X_\mu$.

By applying (2.8) and (2.9), we obtain the following Lemma 2.2, and its proof can be found in [9].

Lemma 2.2. *Under Assumptions 1, 2, and 5, with confidence* $1 - me^{-\theta}$*, it holds that*

$$\|T_0^{\hat{\mathbf{x}}} - T_0^{\mathbf{x}}\| \le \kappa L(1 + \sqrt{\theta})^{h_2} \frac{2^{\frac{2+h_2}{2}} B_k^{\frac{h_2}{2}}}{N^{\frac{h_2}{2}}}, \tag{2.10}$$

$$\|T_1^{\hat{\mathbf{x}}} - T_1^{\mathbf{x}}\| \le \kappa L(1 + \sqrt{\theta})^{h_1} \frac{2^{\frac{2+h_1}{2}} B_k^{\frac{h_1}{2}}}{N^{\frac{h_1}{2}}}, \tag{2.11}$$

where $\theta = \log m + \delta$.

We provide the upper estimate of $\|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0}$, which is crucial for the subsequent error analysis. To this end, we recall Lemma 2.3 from [27].

Lemma 2.3. *Under Assumption 1, for any* $\delta \in (0, 1)$*, with confidence* $1 - \delta$*, the inequality*

$$\frac{1}{m}\sum_{i=1}^m |y_i| \le CM + 4M(1 + \sqrt{2C})\frac{\log(1/\delta)}{\sqrt{m}} \tag{2.12}$$

holds.

Now we prove Proposition 2.1.

Proposition 2.1. *Under Assumptions 1, 2, and 5, for any $\delta \in (0, 1)$, with confidence $1 - \frac{\delta}{4} - me^{-\theta}$, the inequality*

$$\|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0} \le (9C + 4)(m^{-\frac{1}{2}} + 1)LM \log \frac{4}{\delta} \frac{(1 + \sqrt{\theta})^{h_2}(2B_k)^{\frac{h_2}{2}}}{N^{\frac{h_2}{2}}} \tag{2.13}$$

holds.

Proof. By the definition (2.1) and (2.2) of the two-stage sampling operator, we have

$$\begin{aligned} \|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0}^2 &\le \frac{1}{m} \left\|[(K_0(\mu_{\hat{x}_i}, \cdot) - K_0(\mu_{x_i}, \cdot)]\, y_i\right\|_{K_0}^2 \\ &\le \frac{1}{m} \sum_{i=1}^{m} \left\|K_0(\mu_{\hat{x}_i}, \cdot) - K_0(\mu_{x_i}, \cdot)\right\|_{K_0}^2 |y_i|^2, \end{aligned} \tag{2.14}$$

and by substituting (2.9) into (2.14), we have

$$\|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0}^2 \le \frac{L^2}{m} \sum_{i=1}^{m} |y_i|^2 \left\|\mu_{\hat{x}_i} - \mu_{x_i}\right\|_{\mathcal{H}_k}^{2h_2}. \tag{2.15}$$

Furthermore it is known from [26] that

$$\left\|\mu_{\hat{x}_i} - \mu_{x_i}\right\|_{\mathcal{H}_k} \le \frac{(1 + \sqrt{\theta})\sqrt{2B_k}}{\sqrt{N}}. \tag{2.16}$$

Then plugging (2.12) and (2.16) into (2.15), with confidence $1 - \frac{\delta}{4} - me^{-\theta}$, we obtain

$$\begin{aligned} \|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0}^2 &\le \left[CM + 4M(1 + \sqrt{2C})\frac{\log(4/\delta)}{\sqrt{m}}\right]^2 L^2 \frac{(1 + \sqrt{\theta})^{2h_2}(2B_k)^{h_2}}{N^{h_2}} \\ &\le \left[(9C + 4)(m^{-\frac{1}{2}} + 1)LM \log \frac{4}{\delta}\right]^2 \frac{(1 + \sqrt{\theta})^{2h_2}(2B_k)^{h_2}}{N^{h_2}}. \end{aligned} \tag{2.17}$$

Thus we get the desired result. □

To carry out the error decomposition, we recall the expression of the solution of the algorithm (1.6)

$$f_{\hat{D}} = \left(\lambda I + T^{\hat{\mathbf{x}}} T_*^{\hat{\mathbf{x}}}\right)^{-1} T^{\hat{\mathbf{x}}} U^* \hat{S}_0^* \mathbf{y}, \tag{2.18}$$

where the vector $y = (y_1, \cdots, y_m)^T \in \mathbb{R}^m$ is formed by the output data of sample $\hat{D}$.

Let f_D be the minimizer of the coefficient-based regularization on the first-stage sample $D = \{(\mu_{x_i}, y_i)\}_{i=1}^m$. The explicit expression of f_D can be derived in the same way as follows:

$$f_D = (\lambda I + T^{\mathbf{x}} T_*^{\mathbf{x}})^{-1} T^{\mathbf{x}} U^* S_0^* \mathbf{y}. \tag{2.19}$$

Then we introduce the following regularization function f_λ with a positive semi-definite kernel $\widetilde{K}(\mu_x, \mu_{x'}) = \mathbb{E}_{\mu_y}\left[K(\mu_x, \mu_y)K(\mu_{x'}, \mu_y)\right]$, i.e.,

$$f_\lambda = \arg\min_{f \in \mathcal{H}_{\widetilde{K}}} \left\{\|f - f_\rho\|_\rho^2 + \lambda \|f\|_{\widetilde{K}}^2\right\},$$

and then it follows from $L_{\widetilde{K}} = L_K L_K^* = L_{K_0} U U^* L_{K_0} = L_{K_0}^2$ that

$$\begin{aligned} f_\lambda &= (\lambda I + L_{\widetilde{K}})^{-1} L_{\widetilde{K}} f_\rho \\ &= (\lambda I + L_K L_K^*)^{-1} L_K L_K^* f_\rho \\ &= (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^2 f_\rho, \end{aligned} \tag{2.20}$$

see [28] and the references therein. Note that f_λ is the population version of f_D. Now we present the following error decomposition of the total error $f_{\hat{D}} - f_\rho$:

$$\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \leq \|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}} + \|f_D - f_\lambda\|_{\rho_{X_\mu}} + \|f_\lambda - f_\rho\|_{\rho_{X_\mu}}, \tag{2.21}$$

where the first term $\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}}$ is known as the distributed error. The second term $\|f_D - f_\lambda\|_{\rho_{X_\mu}}$ is known as the sample error. The last term $\|f_\lambda - f_\rho\|_{\rho_{X_\mu}}$ is known as the approximation error.

We proceed with further error analysis of the distributed error $\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}}$ and the sample error $\|f_D - f_\lambda\|_{\rho_{X_\mu}}$. We define $\hat{T}_0^{\mathbf{x}} = T^{\mathbf{x}} U^* = \frac{1}{m}\sum_{i=1}^m K(\cdot, \mu_{x_i}) \otimes K(\cdot, \mu_{x_i})$, which is positive semi-definite on $\mathcal{H}_{K_0}$ and provides a good approximation to L_{K_0}. Accordingly, f_D can be represented using $\hat{T}_0^{\mathbf{x}}$ as follows:

$$f_D = \left(\lambda I + \hat{T}_0^{\mathbf{x}} T_0^{\mathbf{x}}\right)^{-1} \hat{T}_0^{\mathbf{x}} S_0^* \mathbf{y}. \tag{2.22}$$

Similarly,

$$f_{\hat{D}} = \left(\lambda I + \hat{T}_0^{\hat{\mathbf{x}}} T_0^{\hat{\mathbf{x}}}\right)^{-1} \hat{T}_0^{\hat{\mathbf{x}}} \hat{S}_0^* \mathbf{y}, \tag{2.23}$$

where $\hat{T}_0^{\hat{\mathbf{x}}} = T^{\hat{\mathbf{x}}} U^* = \frac{1}{m}\sum_{i=1}^m K(\cdot, \mu_{\hat{x}_i}) \otimes K(\cdot, \mu_{\hat{x}_i})$.

For the approximation error, we provide the following proposition, which can be proved by the same method as that employed in Theorem 4 of [28].

Proposition 2.2. *Under Assumption 3, it holds that*

$$\|f_\lambda - f_\rho\|_{\rho_{X_\mu}} \leq c_r \lambda^{\min\left\{1, \frac{r}{2}\right\}}, \quad r > 0, \tag{2.24}$$

$$\|f_\lambda - f_\rho\|_{K_0} \leq c_r' \lambda^{\min\left\{1, \frac{2r-1}{4}\right\}}, \quad r \geq \frac{1}{2}, \tag{2.25}$$

where $c_r = \max\left\{1, \kappa^{2r-4}\right\} \|g_\rho\|_{\rho_{X_\mu}}$ *and* $c_r' = \max\left\{1, \kappa^{2r-5}\right\} \|g_\rho\|_{\rho_{X_\mu}}$.

3. Error analysis

In this section, to obtain the total error $\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}}$, we will analyze the sample error $\|f_D - f_\lambda\|_{\rho_{X_\mu}}$ and the distribution error $\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}}$ in turn. We first consider the indefinite kernel case.

Proposition 3.1. *(Indefinite kernel case). Let K be an indefinite kernel. Under Assumption 3 for* $g_\rho \in L^2_{\rho_{X_\mu}}$ *with* $r \geq \frac{1}{2}$, *for any* $\delta \in (0, 1)$, *with confidence* $1 - 3\delta$, *it holds that*

$$\|f_D - f_\lambda\|_{\rho_{X_\mu}} \leq 2U_1 U_2^2 U_4 + \lambda U_2 U_3 U_4 \|g_\rho\|_{\rho_{X_\mu}},$$

where

$$\begin{aligned}
U_1 &= \left\|(\sqrt{\lambda}I + L_{K_0})^{-\frac{1}{2}}(S_0^*\mathbf{y} - T_0^{\mathbf{x}} f_\lambda)\right\|_{K_0},\\
U_2 &= \left\|(\sqrt{\lambda}I + L_{K_0})^{\frac{1}{2}}(\sqrt{\lambda}I + \hat{T}_0^{\mathbf{x}})^{-\frac{1}{2}}\right\|,\\
U_3 &= \left\|(\sqrt{\lambda}I + \hat{T}_0^{\mathbf{x}})^{\frac{1}{2}}(\lambda I + \hat{T}_0^{\mathbf{x}}\hat{T}_0^{\mathbf{x}})^{-1} L_{K_0}^{\max\{0, r-\frac{1}{2}\}}\right\|,\\
U_4 &= \left\|(\sqrt{\lambda}I + \hat{T}_0^{\mathbf{x}})^{\frac{1}{2}}(\lambda I + \hat{T}_0^{\mathbf{x}}T_0^{\mathbf{x}})^{-1}(\lambda I + \hat{T}_0^{\mathbf{x}}\hat{T}_0^{\mathbf{x}})(\sqrt{\lambda}I + \hat{T}_0^{\mathbf{x}})^{-\frac{1}{2}}\right\|.
\end{aligned}$$

The proof of Proposition 3.1 can be completed by the same method as employed in Proposition 5.1 in [14]. We omit it here.

Proposition 3.2. *(Indefinite kernel case). Under Assumptions 1 and 2, for any* $\delta \in (0, 1)$*, with confidence* $1 - \frac{13\delta}{4} - me^{-\theta}$*, it holds that*

$$\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}} \leq 4\kappa^3(M + \kappa)\lambda^{-\frac{3}{4}} U_2U_4U_5U_6, \tag{3.1}$$

where

$$U_5 = \lambda^{-1}\|T_0^{\mathbf{x}} - T_0^{\hat{\mathbf{x}}}\| + \lambda^{-1}\|\hat{T}_0^{\mathbf{x}} - \hat{T}_0^{\hat{\mathbf{x}}}\| + 1, \tag{3.2}$$

$$\begin{aligned}
U_6 = &\left(\|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0} + (9C + 4)(m^{-\frac{1}{2}} + 1)\log\frac{4}{\delta}\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\|\right)\\
&+ \|f_D\|_{K_0}\left(\|T_0^{\mathbf{x}} - T_0^{\hat{\mathbf{x}}}\| + \|\hat{T}_0^{\mathbf{x}} - \hat{T}_0^{\hat{\mathbf{x}}}\|\right).
\end{aligned} \tag{3.3}$$

Proof. Similar to Proposition 4.3 in [20], by the expressions (2.22) and (2.23) of f_D and $f_{\hat{D}}$, we have the following error decomposition:

$$\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}} \leq \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \mathcal{T}_4, \tag{3.4}$$

where

$$\mathcal{T}_1 = 4\kappa^4\lambda^{-\frac{3}{4}} U_2U_4U_5\|\hat{S}_0^*\mathbf{y} - S_0^*\mathbf{y}\|_{K_0}, \tag{3.5}$$

$$\mathcal{T}_2 = 4\kappa^2\lambda^{-\frac{7}{4}} U_2U_4 \left\|S_0^*\mathbf{y}\right\|_{K_0} \left\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\right\| \left(\left\|T_0^{\mathbf{x}} - T_0^{\hat{\mathbf{x}}}\right\| + \left\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\right\|\right), \tag{3.6}$$

$$\mathcal{T}_3 = 2\lambda^{-\frac{3}{4}} U_2U_4 \left\|S_0^*\mathbf{y}\right\|_{K_0} \left\|\hat{T}_0^{\mathbf{x}} - \hat{T}_0^{\hat{\mathbf{x}}}\right\|, \tag{3.7}$$

$$\begin{aligned}
\mathcal{T}_4 = 4\kappa^4 U_2U_4\|f_D\|_{K_0}\Big[&\lambda^{-\frac{7}{4}}\left(\left\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\right\| + \left\|T_0^{\mathbf{x}} - T_0^{\hat{\mathbf{x}}}\right\|\right)^2\\
&+ \lambda^{-\frac{1}{4}}\left\|T_0^{\mathbf{x}} - T_0^{\hat{\mathbf{x}}}\right\| + \lambda^{-\frac{3}{4}}\left\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\right\|\Big].
\end{aligned} \tag{3.8}$$

Then we only need to estimate the upper bound of $\left\|S_0^*\mathbf{y}\right\|_{K_0}$ under the moment assumption (1.8). Combining the upper bound of $\frac{1}{m}\sum_{i=1}^m |y_i|$ in Lemma 2.3 and the definition of S_0^* yields with confidence $1 - \delta/4$,

$$\left\|S_0^*\mathbf{y}\right\|_{K_0} \leq (9C + 4)(m^{-\frac{1}{2}} + 1)\kappa M\log\frac{4}{\delta}. \tag{3.9}$$

Then the desired error bound follows by substituting (3.9) into (3.6) and (3.7), respectively:

$$\begin{aligned}\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}} \leq 4\kappa^3(M+\kappa)\lambda^{-\frac{3}{4}} U_2 U_4 U_5 \Big[&\|\hat{S}_0^{*}\mathbf{y} - S_0^{*}\mathbf{y}\|_{K_0} + (9C+4)(m^{-\frac{1}{2}}+1)\log\frac{4}{\delta}\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\| \\ &+ \|f_D\|_{K_0}(\|T_0^{\mathbf{x}} - T_0^{\hat{\mathbf{x}}}\| + \|\hat{T}_0^{\mathbf{x}} - \hat{T}_0^{\hat{\mathbf{x}}}\|)\Big]. \end{aligned} \tag{3.10}$$

This proves Proposition 3.2. □

Following the error decomposition scheme in (2.21), we can get the general bound for $\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}}$ with the indefinite kernel immediately by combining the upper bounds obtained in Propositions 2.2, 3.1, and 3.2.

Proposition 3.3. *(Indefinite kernel case). Under the assumptions of Theorem 1, for any* $\delta \in (0,1)$*, with confidence* $1 - \frac{13\delta}{4} - me^{-\theta}$*, it holds that*

$$\begin{aligned}\left\|f_{\hat{D}} - f_\rho\right\|_{\rho_{X_\mu}} \leq\ & 4\kappa^3(M+\kappa)\lambda^{-\frac{3}{4}} U_2 U_4 U_5 U_6 \\ &+ U_4(2U_1U_2^2 + U_2U_3\|g_\rho\|_{\rho_{X_\mu}}\lambda) + c_r\lambda^{\min\{1,\frac{r}{2}\}}. \end{aligned} \tag{3.11}$$

When K is positive semi-definite, the different decompositions for the terms $\|f_D - f_\lambda\|_{\rho_{X_\mu}}$ and $\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}}$ can be performed as follows to refine the error estimate. Based on Proposition 3.1, it is uncomplicated to derive the following decomposition of $\|f_D - f_\lambda\|_{\rho_{X_\mu}}$.

Proposition 3.4. *(Positive semi-definite kernel case). Let K be a positive semi-definite kernel. Under Assumption 3 for* $g_\rho \in L^2_{\rho_{X_\mu}}$*, for any* $\delta \in (0,1)$*, with confidence* $1 - \frac{3\delta}{2}$ *for* $0 < r \leq \frac{3}{2}$ *and* $1 - 2\delta$ *for* $r > \frac{3}{2}$*, it holds that*

$$\|f_D - f_\lambda\|_{\rho_{X_\mu}} \leq 2U_1U_2^2 + U_2U_3\|g_\rho\|_{\rho_{X_\mu}}\lambda^{\min\{1,\frac{r}{2}+\frac{3}{4}\}}. \tag{3.12}$$

Proposition 3.5. *(Positive semi-definite kernel case). Under the assumptions of Theorem 2, for any* $\delta \in (0,1)$*, with confidence* $1 - \frac{3\delta}{4} - me^{-\theta}$*, it holds that*

$$\begin{aligned}\left\|f_{\hat{D}} - f_D\right\|_{\rho_{X_\mu}} \leq\ & 2\lambda^{-\frac{1}{4}}\hat{U}_2\left\|\hat{S}_0^{*}\mathbf{y} - S_0^{*}\mathbf{y}\right\|_{K_0} \\ &+ 6(9C+4)(m^{-\frac{1}{2}}+1)\log\frac{4}{\delta}\kappa M\lambda^{-\frac{3}{4}}\hat{U}_2\left\|\hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}}\right\|, \end{aligned} \tag{3.13}$$

where

$$\hat{U}_2 = \left\|\left(\sqrt{\lambda}I + L_{K_0}\right)^{\frac{1}{2}}\left(\sqrt{\lambda}I + \hat{T}_0^{\hat{\mathbf{x}}}\right)^{-\frac{1}{2}}\right\|.$$

Proof. Since K is positive semi-definite, $\hat{T}_0^{\mathbf{x}} = T_0^{\mathbf{x}}$ and $\hat{T}_0^{\hat{\mathbf{x}}} = T_0^{\hat{\mathbf{x}}}$. Thus we can rewrite the decomposition (3.1) for the indefinite kernel case as follows:

$$\|f_{\hat{D}} - f_D\|_{\rho_{X_\mu}} = \left\|L_{K_0}^{\frac{1}{2}}(f_{\hat{D}} - f_D)\right\|_{K_0} \leq \mathcal{T}_5 + \mathcal{T}_6, \tag{3.14}$$

where

$$\mathcal{T}_5 = \left\|L_{K_0}^{\frac{1}{2}}\left(\lambda I + \hat{T}_0^{\hat{\mathbf{x}}}\hat{T}_0^{\hat{\mathbf{x}}}\right)^{-1}\hat{T}_0^{\hat{\mathbf{x}}}\left(\hat{S}_0^{*}\mathbf{y} - S_0^{*}\mathbf{y}\right)\right\|_{K_0}, \tag{3.15}$$

$$\mathcal{T}_6 = \left\| L_{K_0}^{\frac{1}{2}} \left[\left(\lambda I + \hat{T}_0^{\hat{\mathbf{x}}} \hat{T}_0^{\hat{\mathbf{x}}} \right)^{-1} \hat{T}_0^{\hat{\mathbf{x}}} - \left(\lambda I + \hat{T}_0^{\mathbf{x}} \hat{T}_0^{\mathbf{x}} \right)^{-1} \hat{T}_0^{\mathbf{x}} \right] S_0^* \mathbf{y} \right\|_{K_0}. \tag{3.16}$$

Then we directly invoke the following upper bounds from Proposition 4.5 in [20].

$$\mathcal{T}_5 \leq 2\lambda^{-\frac{1}{4}} \hat{U}_2 \left\| \hat{S}_0^* \mathbf{y} - S_0^* \mathbf{y} \right\|_{K_0}, \tag{3.17}$$

$$\mathcal{T}_6 \leq 6\lambda^{-\frac{3}{4}} \hat{U}_2 \left\| S_0^* \mathbf{y} \right\|_{K_0} \left\| \hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}} \right\|. \tag{3.18}$$

By substituting (3.9) into (3.18), we have

$$\mathcal{T}_6 \leq 6(9C+4)(m^{-\frac{1}{2}}+1) \log \frac{4}{\delta} \kappa M \lambda^{-\frac{3}{4}} \hat{U}_2 \left\| \hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}} \right\|. \tag{3.19}$$

Finally we get our desired result by plugging (3.17) and (3.19) into (3.14). □

Then we provide the upper bound for $\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}}$ with the positive semi-definite kernel by combining Propositions 2.2, 3.4, and 3.5.

Proposition 3.6. *(Positive semi-definite kernel case). Under the assumptions of Theorem 2, for any* $\delta \in (0,1)$, *with confidence* $1 - \frac{7\delta}{4} - me^{-\theta}$ *for* $0 < r \leq \frac{3}{2}$ *and* $1 - \frac{9\delta}{4} - me^{-\theta}$ *for* $r > \frac{3}{2}$, *it holds that*

$$\begin{aligned} \left\| f_{\hat{D}} - f_\rho \right\|_{\rho_{X_\mu}} \leq\ & 2\lambda^{-\frac{1}{4}} \hat{U}_2 \left\| \hat{S}_0^* \mathbf{y} - S_0^* \mathbf{y} \right\|_{K_0} \\ & + 6(9C+4)(m^{-\frac{1}{2}}+1) \log \frac{4}{\delta} \kappa M \lambda^{-\frac{3}{4}} \hat{U}_2 \left\| \hat{T}_0^{\hat{\mathbf{x}}} - \hat{T}_0^{\mathbf{x}} \right\| \\ & + 2U_1 U_2^2 + U_2 U_3 \|g_\rho\|_{\rho_{X_\mu}} \lambda^{\min\left\{1, \frac{r}{2}+\frac{3}{4}\right\}} + c_r \lambda^{\min\left\{1, \frac{r}{2}\right\}}. \end{aligned} \tag{3.20}$$

Then the task is to estimate U_i $(i = 1,2,3,4,5,6)$ and $\hat{U}_2$ in (3.11) and (3.20).

Lemma 3.1. *Under Assumptions 1 and 3, for any* $\delta \in (0,1)$, *with confidence* $1-\delta$, *it holds that*

$$U_1 \leq c_\rho \kappa^{\max\{0, 2r-1\}} \lambda^{\min\left\{0, \frac{r}{2}-\frac{1}{4}\right\}} \mathcal{B}_{m,\lambda} \log \frac{4}{\delta} + c_r \lambda^{\min\left\{1, \frac{r}{2}\right\}}, \tag{3.21}$$

where $c_\rho = 2(C+1)M + \|g_\rho\|_{\rho_{X_\mu}}$.

Proof. Recall the upper error bound with U_1 proved in Proposition 5.2 of [14],

$$U_1 \leq c_\lambda \mathcal{B}_{m,\lambda} \log \frac{4}{\delta} + \|f_\rho - f_\lambda\|_{\rho_{X_\mu}}, \tag{3.22}$$

where

$$c_\lambda = 2(C+1)M\kappa^{-1} + \|f_\lambda\|_\infty \kappa^{-1}.$$

To obtain the upper bound for U_1, we only need to estimate $\|f_\lambda\|_\infty$. By (1.12), we have

$$\|f_\lambda\|_\infty \leq \sup_{\mu_x \in X_\mu} \sqrt{K_0(\mu_x, \mu_x)} \|f_\lambda\|_{K_0}.$$

Furthermore, it is known from $\|f\|_{\rho_{X_\mu}} = \|L_K^{\frac{1}{2}} f\|_K$ for all $f \in L^2_{\rho_{X_\mu}}$ and the expression (2.20) of f_λ that

$$\|f_\lambda\|_\infty^2 \leq \kappa^2 \|(\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{2+r} g_\rho\|_{K_0}^2$$

$$
\begin{aligned}
&= \kappa^2 \left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{\frac{3}{2}+r} L_{K_0}^{\frac{1}{2}} g_\rho \right\|_{K_0}^2 \\
&= \kappa^2 \left\| (\lambda I + L_{K_0}^2)^{-1} L_{K_0}^{\frac{3}{2}+r} \sum_{\ell=1}^{\infty} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}} \phi_\ell \right\|_{\rho_{X_\mu}}^2 .
\end{aligned}
$$

By the spectral decomposition lemma for operators (from [29], p.273), we proceed to expand the operator action:

$$
\begin{aligned}
\|f_\lambda\|_\infty^2 &\le \kappa^2 \left\| \sum_{\ell=1}^{\infty} \frac{\sigma_\ell^{\frac{3}{2}+r}}{\lambda + \sigma_\ell^2} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}} \phi_\ell \right\|_{\rho_{X_\mu}}^2 \\
&= \kappa^2 \sum_{\ell=1}^{\infty} \frac{\sigma_\ell^{3+2r}}{(\lambda + \sigma_\ell^2)^2} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}}^2 .
\end{aligned}
$$

When $0 < r < \frac{1}{2}$,

$$
\begin{aligned}
\sum_{\ell=1}^{\infty} \frac{\sigma_\ell^{3+2r}}{(\lambda + \sigma_\ell^2)^2} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}}^2 &= \sum_{\ell=1}^{\infty} \frac{(\sigma_\ell^2)^{\frac{3}{2}+r}}{(\lambda + \sigma_\ell^2)^{\frac{3}{2}+r}} \frac{1}{(\lambda + \sigma_\ell^2)^{2-\frac{3+2r}{2}}} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}}^2 \\
&\le \lambda^{r-\frac{1}{2}} \|g_\rho\|_{\rho_{X_\mu}}^2 ,
\end{aligned}
$$

and when $r \ge \frac{1}{2}$,

$$
\begin{aligned}
\sum_{\ell=1}^{\infty} \frac{\sigma_\ell^{3+2r}}{(\lambda + \sigma_\ell^2)^2} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}}^2 &= \sum_{\ell=1}^{\infty} \frac{(\sigma_\ell^2)^2}{(\lambda + \sigma_\ell^2)^2} \sigma_\ell^{2r-1} \langle g_\rho, \phi_\ell \rangle_{\rho_{X_\mu}}^2 \\
&\le \kappa^{4r-2} \|g_\rho\|_{\rho_{X_\mu}}^2 .
\end{aligned}
$$

This implies

$$
\|f_\lambda\|_\infty \le \kappa^{\max\{1,2r\}} \|g_\rho\|_{\rho_{X_\mu}} \lambda^{\min\left\{0, \frac{r}{2}-\frac{1}{4}\right\}}. \tag{3.23}
$$

By plugging (2.24) and (3.23) into (3.22), we complete the proof of Lemma 3.1. □

Then we need to invoke the following bounds of U_i $(i = 2, 3, 4)$ and $\hat{U}_2$ in [14].

Lemma 3.2. *For any* $\delta \in (0, 1)$*, with confidence* $1 - \frac{\delta}{2}$*, it holds that*

$$
U_2^2 \le \left\| \left(\sqrt{\lambda} I + L_{K_0} \right) \left(\sqrt{\lambda} I + \hat{T}_0^{\mathbf{x}} \right)^{-1} \right\| \le \left[1 + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda} \log \frac{4}{\delta} \right]^2 . \tag{3.24}
$$

Moreover, we have with confidence $1 - \frac{\delta}{2}$,

$$
U_2 = \left\| \left(\sqrt{\lambda} I + L_{K_0} \right)^{\frac{1}{2}} \left(\sqrt{\lambda} I + \hat{T}_0^{\mathbf{x}} \right)^{-\frac{1}{2}} \right\| \le (1 - t)^{-\frac{1}{2}}, \tag{3.25}
$$

where $c(t)\kappa^4 \left(\frac{\log \frac{4m}{\delta}}{m} \right)^2 \le \lambda \le \kappa^4$ with $0 < t \le \frac{1}{4}$.

Lemma 3.3. *Under Assumption 3, for any* $\delta \in (0,1)$*, with confidence* $1-\frac{\delta}{2}$ *for* $r \geq \frac{1}{2}$*, it holds that*

$$U_3 \leq \begin{cases} 2\lambda^{-\frac{3}{4}}, & \textit{if } 0 < r < \frac{1}{2}, \\ 2^{r+1}\left[1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\log\frac{4}{\delta}\right]^{2r-1}\lambda^{\frac{r}{2}-1}, & \textit{if } \frac{1}{2} \leq r \leq \frac{3}{2}, \\ (8r-4)\kappa^{2r-1}\left[m^{-\frac{1}{2}}\lambda^{-\frac{3}{4}}\log\frac{4}{\delta}+\lambda^{\min\left\{0,\frac{r}{2}-1\right\}}\right], & \textit{if } r > \frac{3}{2}. \end{cases} \tag{3.26}$$

Lemma 3.4. *For any* $\delta \in (0,1)$*, with confidence* $1-\frac{3\delta}{2}$*, it holds that*

$$U_4 \leq c'(t)\left(\log\frac{4}{\delta}\right)^2\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\right)^2, \tag{3.27}$$

where $c'(t) = 1+4(1-t)^{-\frac{1}{2}}+8\left[1-(2\sqrt{2}+1)t\right]^{-1}$.

Recall that $\hat{T}_0^{\mathbf{x}} = UT_1^{\mathbf{x}}U^*$ and $\hat{T}_0^{\hat{\mathbf{x}}} = UT_1^{\hat{\mathbf{x}}}U^*$, and combining the results in Lemma 2.2, we can get the following Lemma 3.5.

Lemma 3.5. *Under Assumptions 1, 2, and 5, with confidence* $1-me^{-\theta}$*, it holds that*

$$U_5 \leq \lambda^{-1}\kappa L(1+\sqrt{\theta})^{h_2}\frac{2^{\frac{2+h_2}{2}}B_k^{\frac{h_2}{2}}}{N^{\frac{h_2}{2}}}+\lambda^{-1}\kappa L(1+\sqrt{\theta})^{h_2}\frac{2^{\frac{2+h_1}{2}}B_k^{\frac{h_1}{2}}}{N^{\frac{h_1}{2}}}+1. \tag{3.28}$$

Lemma 3.6. *Under Assumptions 1, 2, and 5, for any* $\delta \in (0,1)$*, with confidence* $1-\frac{13\delta}{4}-me^{-\theta}$*, it holds that*

$$U_6 \leq \begin{cases} c_1'(t)\left(\log\frac{4}{\delta}\right)^{2r+2}\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\right)^{2r+2}N^{-\frac{h}{2}}(1+\sqrt{\theta})^{h'}\left(1+\|g_\rho\|_{\rho_{X_\mu}}+\lambda^{\frac{r}{2}-\frac{1}{4}}\right)\left(m^{-\frac{1}{2}}+1\right), & \textit{if } \frac{1}{2} \leq r \leq \frac{3}{2}, \\ c_2'(t)\left(\log\frac{4}{\delta}\right)^{3}\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\right)^{3}N^{-\frac{h}{2}}(1+\sqrt{\theta})^{h'}\left(2+\|g_\rho\|_{\rho_{X_\mu}}+\lambda^{\min\left\{\frac{3}{4},\frac{r}{2}-\frac{1}{4}\right\}}\right)\left(m^{-\frac{1}{2}}+1\right), & \textit{if } r > \frac{3}{2}, \end{cases} \tag{3.29}$$

where $c_i'(t)(i=1,2) = (9C+4)2^{\frac{h'+2}{2}}\kappa^{2r}L(\kappa+M)(c_i'+c_r')(B_k^{\frac{h_1}{2}}+B_k^{\frac{h_2}{2}})c'(t)(1-t)^{-\frac{1}{2}}$.

Proof. To estimate U_6, we need to derive the upper bound for $\|f_D\|_{K_0}$, which can be proved by the same method employed in Theorem 5.1 of [20]. When $\frac{1}{2} \leq r \leq \frac{3}{2}$, with confidence $1-3\delta$, it holds that

$$\begin{aligned} \|f_D\|_{K_0} \leq\ & c_1'c'(t)(1-t)^{-\frac{1}{2}}\left(\log\frac{4}{\delta}\right)^{2r+2}\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\right)^{2r+1} \\ & \times\left(\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}+\lambda^{\frac{r}{2}-\frac{1}{4}}\right)+c_r'\lambda^{\frac{2r-1}{4}}+\kappa^{2r-1}\|g_\rho\|_{\rho_{X_\mu}}, \end{aligned} \tag{3.30}$$

where $c_1' = 2c_\rho\kappa^{2r-1}+2^{r+1}\|g_\rho\|_{\rho_{X_\mu}}+2c_r$.

When $r > \frac{3}{2}$, with confidence $1-3\delta$, it holds that

$$\begin{aligned} \|f_D\|_{K_0} \leq\ & c_2'c'(t)(1-t)^{-\frac{1}{2}}\left(\log\frac{4}{\delta}\right)^{3}\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\right)^{2}\left(\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}+m^{-\frac{1}{2}}+\lambda^{\min\left\{\frac{3}{4},\frac{r}{2}-\frac{1}{4}\right\}}\right) \\ & +c_r'\lambda^{\min\left\{1,\frac{2r-1}{4}\right\}}+\kappa^{2r-1}\|g_\rho\|_{\rho_{X_\mu}}, \end{aligned} \tag{3.31}$$

where $c_2' = 2c_\rho\kappa^{2r-1}+(8r-4)\kappa^{2r-1}\|g_\rho\|_{\rho_{X_\mu}}+2c_r$. By substituting the bounds of (2.10), (2.11), (2.13), (3.30), and (3.31) into (3.3), we can obtain the upper bound of U_6. □

Note that under positive semi-definite kernels, Assumption 5 holds for $h_1 = h_2 = h$. The lemma below is given to provide the bound of $\hat{U}_2$.

Lemma 3.7. *Under Assumptions 1, 2, and 5, for any $\delta \in (0,1)$, with confidence $1 - \frac{\delta}{2} - me^{-\theta}$, it holds that*

$$\hat{U}_2 \leq 1 + \kappa L(1 + \sqrt{\theta})^h \frac{2^{\frac{h+2}{2}} B_k^{\frac{h}{2}}}{\lambda^{\frac{1}{2}} N^{\frac{h}{2}}} + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda} \log \frac{4}{\delta}. \tag{3.32}$$

4. Proof of main results

Now we are in a position to prove Theorem 1.

Proof. By substituting (3.21) and (3.25)–(3.29) into (3.11) in Proposition 3.3, we obtain the following bound of $\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}}$ by scaling $\frac{13\delta}{4}$ to δ and $me^{-\theta}$ to $e^{-\gamma}$. For $\frac{1}{2} \leq r \leq \frac{3}{2}$, we have with confidence $1 - \delta - e^{-\gamma}$,

$$\begin{aligned}\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \leq c_1 &\left(\log \frac{13}{\delta}\right)^{2r+4} (1 + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda})^{2r+4} \\ &\times \left[\mathcal{B}_{m,\lambda} + \lambda^{\frac{r}{2}} + \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} (1 + \sqrt{\log m + \gamma})^{h'} \left(m^{-\frac{1}{2}} + 1\right)\right] \\ &\times \left[1 + \lambda^{-1} N^{-\frac{h}{2}} (1 + \sqrt{\log m + \gamma})^{h'}\right] \left(1 + \|g_\rho\|_{\rho_{X_\mu}} + \lambda^{\frac{r}{2} - \frac{1}{4}}\right),\end{aligned}$$

where

$$c_1 = c'(t)(1-t)^{-1} \left[2c_\rho \kappa^{2r-1} + 2^{r+1} \|g_\rho\|_{\rho_{X_\mu}} + 3c_r + 2^{\frac{h'+6}{2}} \kappa^4 (\kappa + M) c_1'(t) \left(1 + LB_k^{\frac{h_1}{2}} + LB_k^{\frac{h_2}{2}}\right)\right].$$

Similarly, for $r > \frac{3}{2}$, we have with confidence $1 - \delta - e^{-\gamma}$,

$$\begin{aligned}\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \leq c_2 &\left(\log \frac{13}{\delta}\right)^{5} (1 + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda})^{5} \\ &\times \left[\mathcal{B}_{m,\lambda} + \lambda^{\frac{1}{4}} m^{-\frac{1}{2}} + \lambda^{\min\{1, \frac{r}{2}\}} + \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} (1 + \sqrt{\log m + \gamma})^{h'} \left(m^{-\frac{1}{2}} + 1\right)\right] \\ &\times \left[1 + \lambda^{-1} N^{-\frac{h}{2}} (1 + \sqrt{\log m + \gamma})^{h'}\right] \left(2 + \|g_\rho\|_{\rho_{X_\mu}} + \lambda^{\min\{\frac{3}{4}, \frac{r}{2} - \frac{1}{4}\}}\right),\end{aligned}$$

where

$$c_2 = c'(t)(1-t)^{-1} \left[2c_\rho \kappa^{2r-1} + (8r-4)\kappa^{2r-1} \|g_\rho\|_{\rho_{X_\mu}} + 3c_r + 2^{\frac{h'+6}{2}} \kappa^4 (\kappa + M) c_2'(t) \left(1 + LB_k^{\frac{h_1}{2}} + LB_k^{\frac{h_2}{2}}\right)\right].$$

This proves Theorem 1. □

Next we will prove Corollary 1 by choosing λ and N according to Theorem 1.

Proof. We choose the values of λ in (1.25) and N in (1.26). First, it is necessary to select a sufficiently large value of m such that

$$\kappa^4 c(t) \log^2(13m/\delta) m^{-2} \leq \lambda = \kappa^4 m^{-\beta} \leq \kappa^4.$$

It is sufficient to ensure that the following inequalities hold:

$$\begin{cases} \frac{1}{2}m^{2-\beta} \geq 2c(t)\log^2(13/\delta), \\ \frac{1}{2}m^{2-\beta} \geq 2c(t)\log^2 m. \end{cases}$$

The first inequality can be obtained when

$$m \geq \left[4c(t)(\log 13/\delta)^2\right]^{\frac{1}{2-\beta}}.$$

Next, we transform the second inequality into $\frac{1}{2}m^{\frac{2-\beta}{2}} \geq 2c(t)m^{-\frac{2-\beta}{2}}\log^2 m$. Let $f(m) = m^{-\frac{2-\beta}{2}}\log^2 m$. By the elementary derivative calculation, we find that the function $f(m)$ attains its maximum value when $m = e^{\frac{4}{2-\beta}}$, i.e., $f(e^{\frac{4}{2-\beta}}) = 16e^{-2}(2-\beta)^{-2}$. Therefore, from $\frac{1}{2}m^{\frac{2-\beta}{2}} \geq 2c(t)m^{-\frac{2-\beta}{2}}\log^2 m \geq 2c(t)16e^{-2}(2-\beta)^{-2}$, we can derive that

$$m \geq \left[2^{12}e^{-4}(2-\beta)^{-4}c^2(t)\right]^{\frac{1}{2-\beta}}.$$

Thus, it suffices to choose m such that

$$m \geq \max\left\{\left[4c(t)(\log 13/\delta)^2\right]^{\frac{1}{2-\beta}}, \left[2^{12}e^{-4}(2-\beta)^{-4}c^2(t)\right]^{\frac{1}{2-\beta}}\right\}.$$

Furthermore, recall that $\mathcal{N}(\lambda) \leq c'_\alpha \lambda^{-1/\alpha}$ with $c'_\alpha = \frac{\alpha c_\alpha}{\alpha-1}$. To get the fastest learning rates, we choose $\lambda = \kappa^4 m^{-\frac{2\alpha}{2\alpha r+1}}$ for $\frac{1}{2} \leq r \leq \frac{3}{2}$, and we have

$$\mathcal{N}(\lambda^{1/2}) \leq c'_\alpha \kappa^{-\frac{2}{\alpha}} m^{\frac{1}{2\alpha r+1}}.$$

Thus,

$$\mathcal{B}_{m,\lambda} = \frac{2\kappa}{\sqrt{m}}\left\{\frac{\kappa}{\sqrt{m}\lambda^{\frac{1}{4}}} + \sqrt{\mathcal{N}(\lambda^{\frac{1}{2}})}\right\} \leq 2\left(\kappa + \sqrt{c'_\alpha}\kappa^{1-\frac{1}{\alpha}}\right)m^{-\frac{\alpha r}{2\alpha r+1}}, \tag{4.1}$$

which implies

$$\mathcal{P}_{m,\lambda} = 1 + \lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda} \leq 1 + 2\left(1 + \sqrt{c'_\alpha}\kappa^{-\frac{1}{\alpha}}\right). \tag{4.2}$$

Let $N = m^{\frac{3\alpha+2\alpha r}{h(2\alpha r+1)}}\log m$. It holds that

$$\begin{aligned} \mathcal{D}_{m,N,\lambda} &\leq \kappa^{-3}m^{\frac{3\alpha}{4\alpha r+2}}m^{-\frac{3\alpha+2\alpha r}{4\alpha r+2}}(\log m)^{-\frac{h}{2}}(\log m)^{\frac{h'}{2}}\left(\frac{1}{\sqrt{\log m}} + \sqrt{\frac{\gamma+\log m}{\log m}}\right)^{h'} \\ &\leq (\log m)^{\frac{h'-h}{2}}\left(1 + \sqrt{1+\gamma}\right)^{h'} m^{-\frac{\alpha r}{2\alpha r+1}}, \end{aligned} \tag{4.3}$$

and

$$\lambda^{-\frac{1}{4}}\mathcal{D}_{m,N,\lambda} \leq (\log m)^{\frac{h'-h}{2}}\left(1 + \sqrt{1+\gamma}\right)^{h'}. \tag{4.4}$$

By substituting (4.1)–(4.4) into (1.21), we have

$$\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \le \tilde{c}_1 \left(\log \frac{13}{\delta}\right)^{2r+4} (1 + \sqrt{1+\gamma})^{2h'} (\log m)^{h'-h} m^{-\frac{\alpha r}{2\alpha r+1}},$$

where

$$\tilde{c}_1 = 2c_1 \left(3 + 2\sqrt{c'_\alpha}\kappa^{-\frac{1}{\alpha}}\right)^{2r+4} \left[2 + \kappa^{2r} + 2\left(\kappa + \sqrt{c'_\alpha}\kappa^{1-\frac{1}{\alpha}}\right)\right]\left(1 + \|g_\rho\|_{\rho_{X_\mu}} + \kappa^{2r-1}\right).$$

When $\frac{3}{2} < r \le 2$, the proof is analogous to that of the case where $\frac{1}{2} \le r \le \frac{3}{2}$ except that we can obtain

$$\lambda^{\frac{1}{4}} m^{-\frac{1}{2}} \le \kappa m^{-\frac{2\alpha r+\alpha+1}{4\alpha r+2}} \tag{4.5}$$

and

$$\lambda^{\frac{r}{2}} \le \kappa^{2r} m^{-\frac{\alpha r}{2\alpha r+1}}. \tag{4.6}$$

Then by substituting (4.1)–(4.6) into (1.21), we have

$$\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \le \tilde{c}_2 \left(\log \frac{13}{\delta}\right)^{5} (1 + \sqrt{1+\gamma})^{2h'} (\log m)^{h'-h} m^{-\frac{\alpha r}{2\alpha r+1}}, \tag{4.7}$$

where

$$\tilde{c}_2 = 2c_2 \left(3 + 2\sqrt{c'_\alpha}\kappa^{-\frac{1}{\alpha}}\right)^{5} \left[2 + \kappa + \kappa^{\min\{2r,4\}} + 2\left(\kappa + \sqrt{c'_\alpha}\kappa^{1-\frac{1}{\alpha}}\right)\right](2 + \|g_\rho\|_{\rho_{X_\mu}} + \kappa^{\min\{2r-1,3\}}).$$

When $r > 2$, the learning rate can be derived using the same method as employed in the aforementioned cases and thus its detailed derivation is omitted.

This completes the proof of Corollary 1. □

Then we proceed to prove Theorem 2 when the positive semi-definite K is used.

Proof. By substituting (2.10), (2.13), (3.21), (3.24), (3.26), and (3.32) into (3.20) in Proposition 3.6, we obtain the following bound of $\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}}$ by scaling $\frac{7\delta}{4}$ to δ and $me^{-\theta}$ to $e^{-\gamma}$. When $0 < r < \frac{1}{2}$, the following result holds with confidence $1 - \delta - e^{-\gamma}$:

$$\begin{aligned}
\|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \le{} & (9C+4)(12\kappa^2+2)(m^{-\frac{1}{2}}+1)LM\left(\log\frac{7}{\delta}\right)^2 \frac{(1+\sqrt{\theta})^h (2B_k)^{\frac{h}{2}}}{\lambda^{\frac{3}{4}} N^{\frac{h}{2}}} \mathcal{A}_{m,N,\lambda} && (4.8)\\
& + 2\left(c_\rho \lambda^{\frac{r}{2}-\frac{1}{4}} \mathcal{B}_{m,\lambda} \log\frac{7}{\delta} + c_r \lambda^{\frac{r}{2}}\right)\left(1 + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda} \log\frac{7}{\delta}\right)^2 && (4.9)\\
& + \left(1 + \lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda} \log\frac{7}{\delta}\right) 2\lambda^{-\frac{3}{4}} \|g_\rho\|_{\rho_{X_\mu}} \lambda^{\frac{r}{2}+\frac{3}{4}} + c_r \lambda^{\frac{r}{2}} \\
\le{} & d_0 \mathcal{A}^3_{m,N,\lambda} \left(\log\frac{7}{\delta}\right)^3 \left(\lambda^{\frac{r}{2}} + \lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} \left(1 + \sqrt{\gamma + \log m}\right)^h (m^{-\frac{1}{2}} + 1)\right), && (4.10)
\end{aligned}$$

where $d_0 = 2\left(c_\rho + \|g_\rho\|_{\rho_{X_\mu}}\right) + 3c_r + (9C+4)(12\kappa^2+2)LM(2B_k)^{\frac{h}{2}}$.

When $\frac{1}{2} \leq r \leq \frac{3}{2}$, with confidence $1 - \delta - e^{-\gamma}$, it holds that

$$\begin{aligned} \|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \leq{}& (9C+4)(12\kappa^2+2)(m^{-\frac{1}{2}}+1)LM\left(\log\frac{7}{\delta}\right)^2 \frac{(1+\sqrt{\theta})^h(2B_k)^{\frac{h}{2}}}{\lambda^{\frac{3}{4}}N^{\frac{h}{2}}}\mathcal{A}_{m,N,\lambda} \\ &+ 2\left(c_\rho\kappa^{2r-1}\mathcal{B}_{m,\lambda}\log\frac{7}{\delta} + c_r\lambda^{\frac{r}{2}}\right)\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\log\frac{7}{\delta}\right)^2 + \left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\log\frac{7}{\delta}\right) \\ &\times 2^{r+1}\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\log\frac{7}{\delta}\right)^{2r-1}\|g_\rho\|_{\rho_{X_\mu}}\lambda^{\frac{r}{2}} + c_r\lambda^{\frac{r}{2}} \\ \leq{}& d_1\mathcal{A}_{m,N,\lambda}^{2\max\{1,r\}}\left(\log\frac{7}{\delta}\right)^{\max\{3,2r+1\}}\left[\mathcal{B}_{m,\lambda} + \lambda^{\frac{r}{2}} + \lambda^{-\frac{3}{4}}N^{-\frac{h}{2}}\left(1+\sqrt{\gamma+\log m}\right)^h(m^{-\frac{1}{2}}+1)\right], \end{aligned} \tag{4.11}$$

where $d_1 = 2c_\rho\kappa^{2r-1} + 2^{r+1}\|g_\rho\|_{\rho_{X_\mu}} + 3c_r + (9C+4)(12\kappa^2+2)LM(2B_k)^{\frac{h}{2}}$.

When $r > \frac{3}{2}$, by scaling $\frac{9\delta}{4}$ to δ and $me^{-\theta}$ to $e^{-\gamma}$, the result holds with confidence $1 - \delta - e^{-\gamma}$:

$$\begin{aligned} \|f_{\hat{D}} - f_\rho\|_{\rho_{X_\mu}} \leq{}& (9C+4)(12\kappa^2+2)(m^{-\frac{1}{2}}+1)LM\left(\log\frac{9}{\delta}\right)^2 \frac{(1+\sqrt{\theta})^h(2B_k)^{\frac{h}{2}}}{\lambda^{\frac{3}{4}}N^{\frac{h}{2}}}\mathcal{A}_{m,N,\lambda} \\ &+ 2\left(c_\rho\kappa^{2r-1}\mathcal{B}_{m,\lambda}\log\frac{9}{\delta} + c_r\lambda^{\frac{r}{2}}\right)\left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\log\frac{9}{\delta}\right)^2 \\ &+ \left(1+\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda}\log\frac{9}{\delta}\right)(8r-4)\kappa^{2r-1}\left[m^{-\frac{1}{2}}\lambda^{-\frac{3}{4}}\log\frac{9}{\delta} + \lambda^{\min\left\{0,\frac{r}{2}-1\right\}}\right]\|g_\rho\|_{\rho_{X_\mu}}\lambda + c_r\lambda^{\min\left\{1,\frac{r}{2}\right\}} \\ \leq{}& d_2\mathcal{A}_{m,N,\lambda}^2\left(\log\frac{9}{\delta}\right)^3\left[\mathcal{B}_{m,\lambda} + \lambda^{1/4}m^{-1/2} + \lambda^{\min\left\{1,\frac{r}{2}\right\}} + \lambda^{-\frac{3}{4}}N^{-\frac{h}{2}}\left(1+\sqrt{\gamma+\log m}\right)^h(m^{-\frac{1}{2}}+1)\right], \end{aligned} \tag{4.12}$$

where $d_2 = 2c_\rho\kappa^{2r-1} + (8r-4)\kappa^{2r-1}\|g_\rho\|_{\rho_{X_\mu}} + 3c_r + (9C+4)(12\kappa^2+2)LM(2B_k)^{\frac{h}{2}}$.

We complete the proof of Theorem 2. □

Next we prove Corollary 2 by choosing λ and N according to Theorem 2.

Proof. For the positive semi-definite K, the constraint on m can be completely removed and the estimates can be further refined. To maximize the learning rate, we take the values of λ in (1.29) and N in (1.30).

When $0 < r < \frac{1}{2}$, we choose $\lambda = m^{-\frac{2\alpha}{\alpha+1}}$. Then

$$\mathcal{B}_{m,\lambda} = \frac{2\kappa}{\sqrt{m}}\left\{\frac{\kappa}{\sqrt{m}\lambda^{\frac{1}{4}}} + \sqrt{\mathcal{N}(\lambda^{\frac{1}{2}})}\right\} \leq 2(\kappa^2 + \kappa\sqrt{c'_\alpha})m^{-\frac{\alpha}{2\alpha+2}}, \tag{4.13}$$

and it follows that

$$\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda} \leq 2(\kappa^2 + \kappa\sqrt{c'_\alpha}). \tag{4.14}$$

Let $N = m^{\frac{3\alpha+2\alpha r}{h(\alpha+1)}}\log m$, and thus

$$\lambda^{-\frac{3}{4}}N^{-\frac{h}{2}}\left(1+\sqrt{\gamma+\log m}\right)^h \leq \left(1+\sqrt{1+\gamma}\right)^h m^{-\frac{\alpha r}{\alpha+1}}, \tag{4.15}$$

and

$$\lambda^{-\frac{1}{2}} N^{-\frac{h}{2}} \left(1 + \sqrt{\gamma + \log m}\right)^h \leq \left(1 + \sqrt{1+\gamma}\right)^h. \tag{4.16}$$

By substituting (4.13)–(4.16) into (1.28), we have

$$\begin{aligned} \left\|f_{\hat{D}} - f_\rho\right\|_{\rho_{X_\mu}} &\leq d_0 \left(\log \frac{7}{\delta}\right)^3 \left(1 + \sqrt{1+\gamma}\right)^{4h} \\ &\quad \times \left[1 + 2(\kappa^2 + \kappa\sqrt{c'_\alpha}) + 2\kappa L(2B_k)^{\frac{h}{2}}\right]^3 \left(m^{-\frac{\alpha r}{\alpha+1}} + m^{-\frac{2\alpha r+\alpha+1}{2\alpha+2}} + m^{-\frac{\alpha r}{\alpha+1}}\right) \\ &\leq \tilde{d}_0 \left(\log \frac{7}{\delta}\right)^3 \left(1 + \sqrt{1+\gamma}\right)^{4h} m^{-\frac{\alpha r}{\alpha+1}}, \end{aligned}$$

where $\tilde{d}_0 = 3d_0 \left[1 + 2(\kappa^2 + \kappa\sqrt{c'_\alpha}) + 2\kappa L(2B_k)^{\frac{h}{2}}\right]^3$.

When $\frac{1}{2} \leq r \leq 2$, we choose $\lambda = m^{-\frac{2\alpha}{2\alpha r+1}}$, and it holds that

$$\mathcal{B}_{m,\lambda} \leq 2(\kappa^2 + \kappa\sqrt{c'_\alpha}) m^{-\frac{\alpha r}{2\alpha r+1}}, \tag{4.17}$$

and

$$\lambda^{-\frac{1}{4}} \mathcal{B}_{m,\lambda} \leq 2(\kappa^2 + \kappa\sqrt{c'_\alpha}). \tag{4.18}$$

Moreover, let $N = m^{\frac{3\alpha+2\alpha r}{h(2\alpha r+1)}} \log m$, and we have

$$\lambda^{-\frac{3}{4}} N^{-\frac{h}{2}} \left(1 + \sqrt{\gamma + \log m}\right)^h \leq \left(1 + \sqrt{1+\gamma}\right)^h m^{-\frac{\alpha r}{2\alpha r+1}}, \tag{4.19}$$

and

$$\lambda^{-\frac{1}{2}} N^{-\frac{h}{2}} \left(1 + \sqrt{\gamma + \log m}\right)^h \leq \left(1 + \sqrt{1+\gamma}\right)^h. \tag{4.20}$$

By substituting (4.17)–(4.20) into (1.28), we can derive

$$\left\|f_{\hat{D}} - f_\rho\right\|_{\rho_{X_\mu}} \leq \begin{cases} \tilde{d}_1 \left(\log \frac{7}{\delta}\right)^{\max\{3,2r+1\}} \left(1 + \sqrt{1+\gamma}\right)^{h\max\{3,2r+1\}} m^{-\frac{\alpha r}{2\alpha r+1}}, & \text{if } \frac{1}{2} \leq r \leq \frac{3}{2}, \\ \tilde{d}_2 \left(\log \frac{9}{\delta}\right)^3 \left(1 + \sqrt{1+\gamma}\right)^{3h} m^{-\frac{\alpha r}{2\alpha r+1}}, & \text{if } \frac{3}{2} < r \leq 2, \end{cases}$$

where

$$\tilde{d}_1 = 8d_1(\kappa^2 + \kappa\sqrt{c'_\alpha}) \left[1 + 2(\kappa^2 + \kappa\sqrt{c'_\alpha}) + 2\kappa L(2B_k)^{\frac{h}{2}}\right]^{2\max\{1,r\}},$$

and

$$\tilde{d}_2 = 5d_2 \left[1 + 2(\kappa^2 + \kappa\sqrt{c'_\alpha}) + 2\kappa L(2B_k)^{\frac{h}{2}}\right]^3.$$

When $r > 2$, we choose $\lambda = m^{-\frac{2\alpha}{4\alpha+1}}$, and the following inequalities hold:

$$\mathcal{B}_{m,\lambda} \leq 2(\kappa^2 + \kappa\sqrt{c'_\alpha}) m^{-\frac{2\alpha}{4\alpha+1}}, \tag{4.21}$$

and

$$\lambda^{-\frac{1}{4}}\mathcal{B}_{m,\lambda} \leq 2(\kappa^2 + \kappa\sqrt{c'_\alpha}). \tag{4.22}$$

Let $N = m^{\frac{7\alpha}{h(4\alpha+1)}} \log m$, and we have

$$\lambda^{-\frac{3}{4}}N^{-\frac{h}{2}}\left(1 + \sqrt{\gamma + \log m}\right)^h \leq \left(1 + \sqrt{1+\gamma}\right)^h m^{-\frac{2\alpha}{4\alpha+1}}, \tag{4.23}$$

and

$$\lambda^{-\frac{1}{2}}N^{-\frac{h}{2}}\left(1 + \sqrt{\gamma + \log m}\right)^h \leq \left(1 + \sqrt{1+\gamma}\right)^h. \tag{4.24}$$

By substituting (4.21)–(4.24) into (1.28), we have

$$\left\|f_{\hat{D}} - f_\rho\right\|_{\rho_{X_\mu}} \leq \tilde{d}_2\left(\log\frac{9}{\delta}\right)^3\left(1 + \sqrt{1+\gamma}\right)^{3h} m^{-\frac{2\alpha}{4\alpha+1}}.$$

This completes the proof of Corollary 2. □

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is supported by the National Science Foundation for Young Scientists of China (Grant No. 12001328), the National Natural Science Foundation of China (Grant No. 11671213), the Development Plan of Youth Innovation Team of University in Shandong Province (No. 2021KJ067), the Shandong Provincial Natural Science Foundation of China (No. ZR2022MF223), and the Qilu University of Technology (Shandong Academy of Sciences) Talent Research Project (No. 2023RCKY133).

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. S. A. Dong, W. C. Sun, Learning rate of distribution regression with dependent samples, *J. Complex.*, **73** (2022), 101679. https://doi.org/10.1016/j.jco.2022.101679

2. N Mücke, Stochastic gradient descent meets distribution regression, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, PMLR, **130** (2021), 2143–2151.

3. Z. Yu, D. W. Ho, Z. Shi, D. X. Zhou, Robust kernel-based distribution regression, *Inverse Probl.*, **37** (2021), 105014. https://doi.org/10.1088/1361-6420/ac23c3

4. Z. Yu, D. W. Ho, Estimates on learning rates for multi-penalty distribution regression, *Appl. Comput. Harmon. Anal.*, **69** (2024), 101609. https://doi.org/10.1016/j.acha.2023.101609

5. A. J. Thorpe, M. M. Oishi, Stochastic optimal control via Hilbert space embeddings of distributions, in *Proceedings of the 60th IEEE Conference on Decision and Control (CDC)*, IEEE, (2021), 904–911. https://doi.org/10.1109/CDC45484.2021.9682801

6. X. Q. Zheng, H. W. Sun, Q. Wu, Regularized least square kernel regression for streaming data, *Commun. Math. Sci.*, **19** (2021), 1533–1548. https://doi.org/10.4310/CMS.2021.v19.n6.a4

7. F. Bauer, S. Pereverzev, L. Rosasco, On regularization algorithms in learning theory, *J. Complexity*, **23** (2007), 52–72. https://doi.org/10.1016/j.jco.2006.07.001

8. X. Guo, L. X. Li, Q. Wu, Modeling interactive components by coordinate kernel polynomial models, *Math. Found. Comput.*, **3** (2020), 263–277. https://doi.org/10.3934/mfc.2020010

9. Z. Szabó, B. K. Sriperumbudur, B. Póczos, A. Gretton, Learning theory for distribution regression, *J. Mach. Learn. Res.*, **17** (2016), 1–40.

10. Z. Y. Fang, Z. C. Guo, D. X. Zhou, Optimal learning rates for distribution regression, *J. Complexity*, **56** (2020), 101426. https://doi.org/10.1016/j.jco.2019.101426

11. B. Schölkopf, A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge, MA, 2002. https://doi.org/10.7551/mitpress/4175.001.0001

12. C. J. Liu, Gabor-based kernel PCA with fractional power polynomial models for face recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **26** (2004), 572–581. https://doi.org/10.1109/TPAMI.2004.1273927

13. Q. Guo, Distributed semi-supervised regression learning with coefficient regularization, *Results Math.*, **77** (2022), 63. https://doi.org/10.1007/s00025-021-01601-4

14. Z. C. Guo, L. Shi, Optimal rates for coefficient-based regularized regression, *Appl. Comput. Harmon. Anal.*, **47** (2019), 662–701. https://doi.org/10.1016/j.acha.2017.11.005

15. Q. Wu, Regularization networks with indefinite kernels, *J. Approx. Theory*, **166** (2013), 1–18. https://doi.org/10.1016/j.jat.2012.10.001

16. Q. Guo, P. X. Ye, Error analysis for ℓ_q-coefficient regularized moving least-square regression, *J. Inequal. Appl.*, **2018** (2018), 262. https://doi.org/10.1186/s13660-018-1856-y

17. T. Hu, R. J. Guo, Distributed robust regression with correntropy losses and regularization kernel networks, *Anal. Appl.*, **22** (2024), 689–725. https://doi.org/10.1142/S0219530523500355

18. B. Q. Su, H. W. Sun, Coefficient-based regularization network with variance loss for error, *Int. J. Wavelets Multiresolut. Inf. Process.*, **19** (2021), 2050055. https://doi.org/10.1142/S0219691320500551

19. S. A. Dong, W. C. Sun, Distributed learning and distribution regression of coefficient regularization, *J. Approx. Theory*, **263** (2021), 105523. https://doi.org/10.1016/j.jat.2020.105523

20. Y. Mao, L. Shi, Z. C. Guo, Coefficient-based regularized distribution regression, *J. Approximation Theory*, **297** (2024), 105995. https://doi.org/10.1016/j.jat.2023.105995

21. H. Z. Tong, Least squares regression under weak moment conditions, *J. Comput. Appl. Math.*, **458** (2025), 116336. https://doi.org/10.1016/j.cam.2024.116336

22. H. Z. Tong, M. Ng, Convergence rates of regularized Huber regression under weak moment conditions, *Anal. Appl.*, **23** (2025), 867–885. https://doi.org/10.1142/S0219530524410021

23. Z. C. Guo, D. X. Zhou, Concentration estimates for learning with unbounded sampling, *Adv. Comput. Math.*, **38** (2013), 207–223. https://doi.org/10.1007/s10444-011-9238-8

24. S. Y. Huang, Robust learning of Huber loss under weak conditional moment, *Neurocomputing*, **507** (2022), 191–198. https://doi.org/10.1016/j.neucom.2022.08.012

25. A. Caponnetto, E. D. Vito, Optimal rates for the regularized least-squares algorithm, *Found. Comput. Math.*, **7** (2007), 331–368. https://doi.org/10.1007/s10208-006-0196-8

26. Z. Szabó, A. Gretton, B. Póczos, B. Sriperumbudur, Two-stage sampled learning theory on distributions, in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*, PMLR, San Diego, USA, (2015), 948–957.

27. C. Wang, D. X. Zhou, Optimal learning rates for least squares regularized regression with unbounded sampling, *J. Complexity*, **27** (2011), 55–67. https://doi.org/10.1016/j.jco.2010.10.002

28. S. Smale, D. X. Zhou, Learning theory estimates via integral operators and their approximations, *Constr. Approximation*, **26** (2007), 153–172. https://doi.org/10.1007/s00365-006-0659-y

29. R. V. Kadison, J. R. Ringrose, *Fundamentals of the Theory of Operator Algebras. Volume II: Advanced Theory*, Academic Press, New York, 1986.

AIMS Press