Journal of Machine Learning Research 25 (2024) 1-56 Submitted 11/22; Revised 2/24; Published 4/24

Statistical Optimality of Divide and Conquer Kernel-based Functional Linear Regression

Jiading Liu 20210180088@FUDAN.EDU.CN
School of Mathematical Sciences
Fudan University, Shanghai 200433, China

Lei Shi* LEISHI@FUDAN.EDU.CN
School of Mathematical Sciences and Shanghai
Key Laboratory for Contemporary Applied Mathematics
Fudan University, Shanghai 200433, China
Shanghai Artificial Intelligence Laboratory
701 Yunjin Road, Shanghai 200232, China

Editor: Samuel Kaski

Abstract

Previous analysis of regularized functional linear regression in a reproducing kernel Hilbert space (RKHS) typically requires the target function to be contained in this kernel space. This paper studies the convergence performance of divide-and-conquer estimators in the scenario that the target function does not necessarily reside in the underlying RKHS. As a decomposition-based scalable approach, the divide-and-conquer estimators of functional linear regression can substantially reduce the algorithmic complexities in time and memory. We develop an integral operator approach to establish sharp finite sample upper bounds for prediction with divide-and-conquer estimators under various regularity conditions of explanatory variables and target function. We also prove the asymptotic optimality of the derived rates by building the mini-max lower bounds. Finally, we consider the convergence of noiseless estimators and show that the rates can be arbitrarily fast under mild conditions.

Keywords: functional linear regression, reproducing kernel Hilbert space, divide-and-conquer estimator, model misspecification, mini-max optimal rates

1. Introduction

Functional data analysis (FDA) has been an intense recent study, achieving remarkable success in a wide range of fields, including, among many others, chemometrics, linguistics, medicine and economics (see, e.g., Ramsay and Silverman, 2005; Wang et al., 2016). Under an FDA framework, the explanatory variable is usually a random function. We consider the following functional linear regression model to characterize the functional nature of explanatory variables. Let Y be a scalar response, and X be a random element taking values in $\mathcal{L}^2(\mathcal{T})$. Throughout the paper, we use $\mathcal{L}^2(\mathcal{T})$ to denote the Hilbert space of square integrable functions defined over a domain $\mathcal{T} \subseteq \mathbb{R}^D$ for some integer $D \geq 1$. In the

∗. Corresponding author

functional linear regression model, the dependence of Y and X is expressed as

$$Y = \int_{\mathcal{T}} \beta_0(t) X(t) dt + \epsilon, \tag{1.1}$$

where $\beta_0 \in \mathcal{L}^2(\mathcal{T})$ is the slope function and ϵ is a random noise independent of X with zero mean and bounded variance. The goal of functional linear regression is to construct an estimator $\hat{\beta}$ to approximate β_0 based on training samples of (X, Y). The performance of an estimator can be measured by the prediction risk, given by

$$\mathcal{R}(\hat{\beta}) := \mathbb{E}\left[\left(Y - \int_{\mathcal{T}} \hat{\beta}(t) X(t) dt\right)^2\right], \tag{1.2}$$

or equivalently, the excess prediction risk $\mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta_0)$.

The research on model (1.1) can be traced back to the 1990s (see, e.g., Hastie and Mallows, 1993; Marx, 1996; Cardot et al., 1999). Subsequently, a vast amount of literature has emerged to study the prediction and estimation problems under this model. A flourishing line of research is based on the functional principal component analysis (FPCA), leveraging spectral expansions of the covariance kernel of X and its empirical counterpart to estimate the slope function (see, e.g., Ramsay and Silverman, 2005; Yao et al., 2005; Cai and Hall, 2006; Hall and Horowitz, 2007). A necessary condition for the success of the FPCA-based approaches is that the slope function β_0 can be efficiently represented by the leading functional principal components, which, however, fails to hold in many applications. To address this issue, another influential line of research employs kernel-based estimators to approximate the target β_0 in a suitable reproducing kernel Hilbert space (RKHS) (see, e.g., Yuan and Cai, 2010; Cai and Yuan, 2012). More concretely, given a training sample set $S := \{(X_i, Y_i)\}_{i=1}^N$ consisting of N independent copies of (X, Y), one can employ an RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ induced by a reproducing kernel $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ to estimate β_0 through the regularized least squares (RLS) estimators defined by

$$\hat{\beta}_{S,\lambda} := \operatorname*{argmin}_{\beta \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^N \left(Y_i - \int_{\mathcal{T}} \beta(t) X_i(t) dt\right)^2 + \lambda \|\beta\|_K^2 \right\}. \tag{1.3}$$

Here we choose a tuning parameter $\lambda > 0$ to balance fidelity to the data and complexity of the estimators (measured by its squared $\mathcal{H}_K$ norm). According to the Representer Theorem proved in Yuan and Cai (2010), $\hat{\beta}_{S,\lambda}$ can be uniquely expressed as $\hat{\beta}_{S,\lambda}(\cdot) = \sum_{i=1}^N c_i \int_{\mathcal{T}} K(\cdot, t) X_i(t) dt$ with $(c_1, \cdots, c_N)^T = (\lambda N \mathbb{I}_N + \mathbb{K}_{\boldsymbol{X}})^{-1} \boldsymbol{Y}$, where $\mathbb{I}_N$ is the identity matrix on $\mathbb{R}^N$, $\mathbb{K}_{\boldsymbol{X}} \in \mathbb{R}^{N \times N}$ is the kernel matrix evaluated on $\boldsymbol{X} := \{X_1, \cdots, X_N\}$ with the $(i,j)-$entrance $[\mathbb{K}_{\boldsymbol{X}}]_{i,j} = \int_{\mathcal{T}} X_i(s) K(s,t) X_j(t) ds dt$, and $\boldsymbol{Y} := (Y_1, \cdots, Y_N)^T$. Under the assumption that the slope function β_0 belongs to the RKHS $\mathcal{H}_K$, it is shown in Cai and Yuan (2012) that the excess prediction risk of $\hat{\beta}_{S,\lambda}$ can achieve the mini-max optimal convergence rates.

In this paper, we aim to further advance the line of research on the kernel-based approach designed for functional linear regression model (1.1). Specially, we will study the convergence behavior of divided-and-conquer RLS estimators without requiring the unknown slope function β_0 to be contained in the RKHS $\mathcal{H}_K$. As a generalization of classical kernel ridge regression (see, for example, Murphy, 2012), algorithm (1.3) suffers from the same complexity issue that seriously limits its performance when dealing with massive data. To make the computational problem more tractable for large-scale sample sets, we implement algorithm (1.3) via the divide-and-conquer approach. We randomly partition the entire sample set S into m disjoint equal-sized subsets $S_1, \cdots, S_m$. On each S_j, a local estimator $\hat{\beta}_{S_j,\lambda}$ is obtained according to algorithm (1.3), i.e.,

$$\hat{\beta}_{S_j,\lambda}(\cdot) = \sum_{i:(X_i,Y_i)\in S_j} c_i \int_{\mathcal{T}} K(\cdot,t)X_i(t)dt \text{ where } (c_i)_{\{i:(X_i,Y_i)\in S_j\}} = (\lambda|S_j|\mathbb{I}_{|S_j|} + \mathbb{K}_{\boldsymbol{X}_j})^{-1}\boldsymbol{Y}_j.$$

Here $|S_j|$ denotes the cardinality of S_j, $\boldsymbol{X}_j$ is the set of X's sample in S_j, and $\boldsymbol{Y}_j \in \mathbb{R}^{|S_j|}$ is a vector composed of Y's sample in S_j. Divide-and-conquer RLS estimator is then computed by simply averaging $\{\hat{\beta}_{S_j,\lambda}\}_{j=1}^m$, which is given by

$$\overline{\beta}_{S,\lambda} := \frac{1}{m}\sum_{j=1}^m \hat{\beta}_{S_j,\lambda}. \tag{1.4}$$

This approach is appealing due to its easy exercisable partitions. By partitioning the sample set into m subsets of equal size and executing algorithm (1.3) on each subset concurrently, one can approximately diminish the computational complexities in terms of time and memory to $\frac{1}{m^2}$ of the initial requirements. In the context of regression analysis for massive data, divide-and-conquer kernel ridge regression and its variants have been extensively studied in statistics and machine learning communities (see, e.g., Zhang et al., 2015; Guo et al., 2017; Lin et al., 2017; Dumpert and Christmann, 2018; Mücke and Blanchard, 2018; Hamm and Steinwart, 2021; Sun and Wu, 2021; Hamm and Steinwart, 2022; Köhler and Christmann, 2022). In the present paper, we evaluate the prediction performance of averaged estimator $\overline{\beta}_{S,\lambda}$ in (1.4) via its excess prediction risk:

$$\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \tag{1.5}$$

in a more general setting which allows $\beta_0 \notin \mathcal{H}_K$. In supervised learning problem, if the target function resides outside the hypothesis space, for instance, the underlying RKHS in kernel-based regression, this scenario is often referred to as model misspecification (see, e.g., Rao, 1971; Bach, 2008). More recently, convergence behaviors of kernel ridge regression in model misspecification scenarios have been investigated in many works (see, e.g., Fischer and Steinwart, 2020; Lin et al., 2020; Sun and Wu, 2021), which show asymptotically minimax optimal rates in a variety of situations. In practice, canonical choices of $\mathcal{H}_K$ in (1.3) are the Sobolev spaces of smoothness s (see, for example, Yuan and Cai, 2010). Though such an RKHS is dense in $\mathcal{L}^2(\mathcal{T})$, the assumption that β_0 lies precisely in it is too restrictive in many real applications, as this assumption requires the derivatives of β_0 up to order $s-1$ are absolute continuous and its $s-$th derivative belongs to $\mathcal{L}^2(\mathcal{T})$. This raises the question of whether the global RLS estimator (1.3) and its averaged version (1.4) can still maintain excellent prediction performances in the model misspecification scenario $\beta_0 \notin \mathcal{H}_K$. We positively answer this question by establishing a tight convergence analysis with an integral operator technique. Furthermore, we also consider the noiseless circumstance when the model (1.1) has no additive noise. The noiseless condition means no ambiguity of the response Y given the explanatory variable X; in other words, the response Y is determined uniquely by the input X. The noiseless linear model has been widely adopted in many areas, including image classification and sound recognition (see, for example, Jun et al., 2019). The convergence of estimators in a noiseless model is very important but has not been considered till the very recent papers (see, e.g., Jun et al., 2019; Berthier et al., 2020; Sun and Wu, 2021).

The main contribution of this paper is to present new finite sample bounds on the prediction risk (1.5) concerning various regularity conditions. These conditions characterize the complexity of the prediction problem in functional linear regression model (1.1), which is measured through regularities of the explanatory variable X, optimum β_0, and their images under the kernel operators. See Section 2 and Section 3 for precise definitions and statements. Our analysis of convergence incorporates these regularity conditions into the integral operator techniques, substantially generalizing previously published bounds, which only consider the case $\beta_0 \in \mathcal{H}_K$, to the model misspecification scenario $\beta_0 \notin \mathcal{H}_K$ and the divide-and-conquer estimators. For prediction using the noisy model, the established convergence is tight as in most cases we prove upper and lower bounds on the performance of estimators that almost match. For prediction using the noiseless model, we prove that the estimators can converge with arbitrarily fast polynomial rates if the reproducing kernel or the covariance kernel is sufficiently smooth. Thus the estimators show some adaptivity to the complexity of the prediction problem. Besides, our analysis only requires the the kernel function to be square integrable, eliminating the uniformly bounded or even continuity assumptions required in previous literature, which is more in line with the practical application scenarios of functional data analysis.

The rest of this paper is organized as follows. We start in Section 2 with an introduction to notations, general assumptions, and some preliminary results. In Section 3, we describe the regularity conditions and present main theorems and their corollaries. In Section 4, we give further comments on these regularity conditions and main results and compare them with other related contributions. All proofs can be found in Section 5 and Appendix A.

2. Preliminaries

In this section we will provide basic notations and some preliminary results necessary for the further statement. We first recall some basic notations in operator theory (see, for example, Conway, 2000). Consider a linear operator $A : \mathcal{H} \to \mathcal{H}'$, where both $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ and $(\mathcal{H}', \langle\cdot,\cdot\rangle_{\mathcal{H}'})$ represent Hilbert spaces, equipped with their respective norms $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}'}$. The collection of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}'$ forms a Banach space when considered under operator norm $\|A\|_{\mathcal{H},\mathcal{H}'} = \sup_{\|f\|_{\mathcal{H}}=1} \|Af\|_{\mathcal{H}'}$, symbolized as $\mathscr{B}(\mathcal{H}, \mathcal{H}')$ or $\mathscr{B}(\mathcal{H})$ in cases where $\mathcal{H} = \mathcal{H}'$. In scenarios where $\mathcal{H}$ and $\mathcal{H}'$ are implicitly understood, the subscript is omitted, simplifying the operator norm notation to $\|\cdot\|$. The adjoint of A, denoted by A^*, satisfies the equality $\langle Af, f'\rangle_{\mathcal{H}} = \langle f, A^* f'\rangle_{\mathcal{H}'}$ for all $f \in \mathcal{H}$ and $f' \in \mathcal{H}'$. If A is an element of $\mathscr{B}(\mathcal{H}, \mathcal{H}')$, then its adjoint A^* belongs to $\mathscr{B}(\mathcal{H}', \mathcal{H})$, and the norms of A and A^* are equivalent. An operator $A \in \mathscr{B}(\mathcal{H})$ is identified as self-adjoint if $A^* = A$ and as nonnegative if it is self-adjoint and satisfies $\langle Af, f\rangle_{\mathcal{H}} \geq 0$ for every $f \in \mathcal{H}$. The operator norm of a nonnegative operator $A \in \mathscr{B}(\mathcal{H})$ has an equivalent expression:

$$\|A\| = \sup_{x\in\mathcal{H}, \|x\|_{\mathcal{H}}=1} \langle Ax, x\rangle_{\mathcal{H}}. \tag{2.1}$$

For $f \in \mathcal{H}$ and $f' \in \mathcal{H}'$, we introduce a rank-one operator $f \otimes f' : \mathcal{H} \to \mathcal{H}'$ defined by $f \otimes f'(h) = \langle f, h\rangle_{\mathcal{H}} f', \forall h \in \mathcal{H}$. Consider A to be a compact and nonnegative operator within $\mathscr{B}(\mathcal{H})$. According to the Spectral Theorem, it is guaranteed that an orthonormal basis $\{e_k\}_{k\geq 1}$, composed of A's eigenfunctions, exists within $\mathcal{H}$. This basis allows A to be expressed as $A = \sum_{k\geq 1} \lambda_k e_k \otimes e_k$, where λ_k denotes the non-negative eigenvalues of A in a descending sequence. These eigenvalues (with geometric multiplicities) may either form a finite set or approach zero as k increases indefinitely. Moreover, for any $r > 0$, we define the $r-$th power of A as $A^r = \sum_{k\geq 1} \lambda_k^r e_k \otimes e_k$, which is itself a nonnegative compact operator on $\mathcal{H}$. An operator A belonging to $\mathscr{B}(\mathcal{H}, \mathcal{H}')$ is identified as a Hilbert-Schmidt operator if, for a given orthonormal basis $\{e_k\}_{k\geq 1}$ of $\mathcal{H}$, the series $\sum_{k\geq 1} \|Ae_k\|^2_{\mathcal{H}'}$ converges. The collection of Hilbert-Schmidt operators constitutes a Hilbert space itself, equipped with the inner product defined by $\langle A, B\rangle_{HS} = \sum_{k\geq 1}\langle Ae_k, Be_k\rangle_{\mathcal{H}'}$, with $\|\cdot\|_{HS}$ representing the associated norm. In particular, a Hilbert-Schmidt operator A is compact and we have the following inequality to relate its two different norms:

$$\|A\| \leq \|A\|_{HS}. \tag{2.2}$$

An operator $A \in \mathscr{B}(\mathcal{H}, \mathcal{H}')$ is of trace class if $\sum_{k\geq 1}\langle (A^*A)^{1/2} e_k, e_k\rangle_{\mathcal{H}} < \infty$ for some (any) orthonormal basis $\{e_k\}_{k\geq 1}$ of $\mathcal{H}$. All trace class operators constitute a Banach space endowed with the norm $trace(A) := \sum_{k\geq 1}\langle (A^*A)^{1/2} e_k, e_k\rangle_{\mathcal{H}}$. It is obviously for any nonnegative operator $A \in \mathscr{B}(\mathcal{H})$,

$$trace(A) = \sum_{k\geq 1}\langle Ae_k, e_k\rangle_{\mathcal{H}}. \tag{2.3}$$

In the following, we fix a reproducing kernel Hilbert space (RKHS), denoted as $\mathcal{H}_K$, consisting of functions $f : \mathcal{T} \to \mathbb{R}$, where each evaluation functional on this space is bounded. Consequently, there exists a distinct symmetric nonnegative definite kernel function $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, known as the reproducing kernel, intrinsically linked to $\mathcal{H}_K$. Let $K_t : \mathcal{T} \to \mathbb{R}$ defined by $K_t(\cdot) = K(\cdot, t)$ for $t \in \mathcal{T}$ and denote by $\langle \cdot, \cdot\rangle_K$ the inner product of $\mathcal{H}_K$ which induces the norm $\|\cdot\|_K$. Then $K_t \in \mathcal{H}_K$ and the reproducing property

$$f(t) = \langle f, K_t\rangle_K$$

holds for all $t \in \mathcal{T}$ and $f \in \mathcal{H}_K$. Furthermore, it is well known in the literature (refer to Aronszajn (1950)) that any symmetric positive definite kernel K distinctly characterizes an RKHS, $\mathcal{H}_K$, for which K serves as the reproducing kernel. Throughout the paper, we assume that K is measurable on $\mathcal{T} \times \mathcal{T}$ such that

$$\int_{\mathcal{T}}\int_{\mathcal{T}} K^2(t, t')dtdt' < \infty.$$

Recall that $\mathcal{L}^2(\mathcal{T})$ is the Hilbert space of functions from $\mathcal{T}$ to $\mathbb{R}$ square-integrable with respect to Lebesgue measure. Denote by $\|\cdot\|_{\mathcal{L}^2}$ the corresponding norm of $\mathcal{L}^2(\mathcal{T})$ induced by the inner product $\langle f,g\rangle_{\mathcal{L}^2}=\int_{\mathcal{T}} f(t)g(t)dt$. The reproducing kernel K induces an integral operator $L_K:\mathcal{L}^2(\mathcal{T})\to\mathcal{L}^2(\mathcal{T})$, given by, for $f\in\mathcal{L}^2(\mathcal{T})$ and $t\in\mathcal{T}$,

$$L_K(f)(t)=\int_{\mathcal{T}} K(s,t)f(s)ds,$$

which is a nonnegative, compact operator on $\mathcal{L}^2(\mathcal{T})$. Then $L_K^{1/2}$ is well-defined and compact, and $L_K^{1/2}$ is an isomorphism from $\overline{\mathcal{H}_K}$, the closure of $\mathcal{H}_K$ in $\mathcal{L}^2(\mathcal{T})$, to $\mathcal{H}_K$, i.e., for each $f\in\overline{\mathcal{H}_K}$, $L_K^{1/2}f\in\mathcal{H}_K$ and

$$\|f\|_{\mathcal{L}^2}=\|L_K^{1/2}f\|_K. \tag{2.4}$$

Since we are mainly interested in the model misspecification scenario $\beta_o\notin\mathcal{H}_K$, we will always assume $\mathcal{H}_K$ is dense in $\mathcal{L}^2(\mathcal{T})$, i.e., $\mathcal{L}^2(\mathcal{T})=\overline{\mathcal{H}_K}$.

Besides the reproducing kernel K, another important kernel in our setting is the covariance kernel. Without loss of generality, we let the explanatory variable X satisfy $\mathbb{E}[X]=0$ and further assume $\mathbb{E}\left[\|X\|^2_{\mathcal{L}^2}\right]<\infty$. Then the covariance kernel $C:\mathcal{T}\times\mathcal{T}\to\mathbb{R}$, given by $C(s,t):=\mathbb{E}\left[X(s)X(t)\right],\forall s,t\in\mathcal{T}$, can define a compact nonnegative integral operator $L_C:\mathcal{L}^2(\mathcal{T})\to\mathcal{L}^2(\mathcal{T})$ via

$$L_C(f)(t)=\int_{\mathcal{T}} C(s,t)f(s)ds,\quad \forall f\in\mathcal{L}^2(\mathcal{T}) \text{ and } \forall t\in\mathcal{T}.$$

We next use L_K and L_C to give an expression of estimator $\hat\beta_{S,\lambda}$ in (1.3). Recall that $\mathcal{L}^2(\mathcal{T})=\overline{\mathcal{H}_K}$ and norms relation (2.4). We can express $\hat\beta_{S,\lambda}$ as $\hat\beta_{S,\lambda}=L_K^{1/2}\hat f_{S,\lambda}$ with

$$\hat f_{S,\lambda}=\operatorname*{argmin}_{f\in\mathcal{L}^2(\mathcal{T})}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\left\langle L_K^{1/2}f,X_i\right\rangle_{\mathcal{L}^2}\right)^2+\lambda\|f\|^2_{\mathcal{L}^2}\right\}.$$

Following the proof of Theorem 6.2.1 in Hsing and Eubank (2015), we can solve $\hat f_{S,\lambda}$ explicitly and obtain the following proposition.

Proposition 1 *The estimator $\hat\beta_{S,\lambda}$ in (1.3) can be expressed as $\hat\beta_{S,\lambda}=L_K^{1/2}\hat f_{S,\lambda}$ with*

$$\hat f_{S,\lambda}=(\lambda I+T_{\mathbf{X}})^{-1}\frac{1}{|S|}\sum_{(X_i,Y_i)\in S}L_K^{1/2}X_iY_i, \tag{2.5}$$

where I denotes the identity operator on $\mathcal{L}^2(\mathcal{T})$, $|S|=N$ is the cardinality of $S=\{(X_i,Y_i)\}_{i=1}^N$, and $T_{\mathbf{X}}:\mathcal{L}^2(\mathcal{T})\to\mathcal{L}^2(\mathcal{T})$ is an empirical operator with $\mathbf{X}=\{X_1,\cdots,X_N\}$ defined by

$$T_{\mathbf{X}}=\frac{1}{|S|}\sum_{X_i\in\mathbf{X}}L_K^{1/2}X_i\otimes L_K^{1/2}X_i. \tag{2.6}$$

Recall that $S=\cup_{j=1}^m S_j$ with $S_j\cap S_k=\emptyset$ for $j\neq k$ and $|S_j|=\frac{N}{m}$. One can define the empirical operators $T_{\mathbf{X}_j}$ with $\mathbf{X}_j=\{X_i:(X_i,Y_i)\in S_j\}$ according to (2.6) and compute the local estimator $\hat f_{S_j,\lambda}$ as (2.5), i.e.,

$$T_{\mathbf{X}_j}=\frac{1}{|S_j|}\sum_{X_i\in\mathbf{X}_j}L_K^{1/2}X_i\otimes L_K^{1/2}X_i$$

and

$$\hat{f}_{S_j,\lambda} = (\lambda I + T_{\mathbf{X}_j})^{-1} \frac{1}{|S_j|} \sum_{(X_i,Y_i)\in S_j} L_K^{1/2} X_i Y_i.$$

Then the averaged estimator $\overline{\beta}_{S,\lambda}$ in (1.4) is given by $\overline{\beta}_{S,\lambda} = L_K^{1/2}\overline{f}_{S,\lambda}$ with $\overline{f}_{S,\lambda} := \frac{1}{m}\sum_{j=1}^m \hat{f}_{S_j,\lambda}$.

To derive the upper bounds of excess prediction error, for any estimator $\hat{\beta} \in \mathcal{L}^2(\mathcal{T})$, we rewrite $\mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta_0)$ as

$$\mathcal{R}(\hat{\beta}) - \mathcal{R}(\beta_0) = \mathbb{E}\left[\left\langle X, \hat{\beta} - \beta_0 \right\rangle_{\mathcal{L}^2}^2\right] = \left\| L_C^{1/2}\left(\hat{\beta} - \beta_0\right)\right\|_{\mathcal{L}^2}^2. \tag{2.7}$$

Notice $T_{\mathbf{X}}$ and $\frac{1}{|S|}\sum_{(X_i,Y_i)\in S} L_K^{1/2} X_i Y_i$ are empirical versions of $L_K^{1/2} L_C L_K^{1/2}$ and $L_K^{1/2} L_C \beta_0$, respectively. We thus introduce intermediate function $f_\lambda := \left(\lambda I + L_K^{1/2} L_C L_K^{1/2}\right)^{-1} L_K^{1/2} L_C \beta_0$ which can be expected to approximate $\hat{f}_{S,\lambda}$ and its averaged version $\overline{f}_{S,\lambda}$. According to (2.7), we then split $\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)$ into two parts:

$$\begin{aligned}\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) &= \left\| L_C^{1/2}\left(L_K^{1/2}\overline{f}_{S,\lambda} - L_K^{1/2} f_\lambda + L_K^{1/2} f_\lambda - \beta_0\right)\right\|_{\mathcal{L}^2}^2 \\ &\le 2\mathscr{S}(S,\lambda) + 2\mathscr{A}(\lambda),\end{aligned} \tag{2.8}$$

where $\mathscr{S}(S,\lambda) := \left\| L_C^{1/2} L_K^{1/2}\overline{f}_{S,\lambda} - L_C^{1/2} L_K^{1/2} f_\lambda\right\|_{\mathcal{L}^2}^2$ and $\mathscr{A}(\lambda) := \left\| L_C^{1/2} L_K^{1/2} f_\lambda - L_C^{1/2}\beta_0\right\|_{\mathcal{L}^2}^2$.

In Section 3, we will describe the assumptions that are used to estimate $\mathscr{S}(S,\lambda)$ and $\mathscr{A}(\lambda)$, and then state the main results of this paper. Before that, we give a further characterization of the operators which are crucial in our estimation. For simplicity, let $T := L_K^{1/2} L_C L_K^{1/2}$ and $T_* := L_C^{1/2} L_K L_C^{1/2}$. Note that

$$T = L_K^{1/2} L_C^{1/2}\left(L_K^{1/2} L_C^{1/2}\right)^* \text{ and } T_* = \left(L_K^{1/2} L_C^{1/2}\right)^* L_K^{1/2} L_C^{1/2}.$$

Due to the compactness of $L_K^{1/2} L_C^{1/2}$, both T and T_* are compact and nonnegative operators on $\mathcal{L}^2(\mathcal{T})$. Singular value decomposition of $L_K^{1/2} L_C^{1/2}$ (see Hsing and Eubank, 2015, Theorem 4.3.1) leads to the following expansions:

$$\begin{aligned} L_K^{1/2} L_C^{1/2} &= \sum_{k\ge1} \sqrt{\mu_k}\varphi_k \otimes \phi_k, \\ L_C^{1/2} L_K^{1/2} &= \sum_{k\ge1} \sqrt{\mu_k}\phi_k \otimes \varphi_k, \\ T &= \sum_{k\ge1} \mu_k \phi_k \otimes \phi_k, \\ T_* &= \sum_{k\ge1} \mu_k \varphi_k \otimes \varphi_k, \end{aligned} \tag{2.9}$$

where $\{\mu_k\}_{k\ge1}$ is a non-negative, non-increasing and summable sequence, $\{\phi_k\}_{k\ge1}$ and $\{\varphi_k\}_{k\ge1}$ are two orthonormal bases of $L^2(\mathcal{T})$. Actually, for any $\mu_k > 0$, $\sqrt{\mu_k}$ is the singular values of $L_K^{1/2}L_C^{1/2}$ and the corresponding left and right singular vectors are given by φ_k and ϕ_k, which are the eigenvectors (with the same eigenvalue μ_k) of T and T_*. In particular, the system $\{\mu_k, \phi_k, \varphi_k\}_{k\geq 1}$ plays an important role in describing regularities of the explanatory variable X and the slope function β_0 which we will explain in details in Section 3.

3. Main Results

In this section, we will present our main theoretical results on the upper and lower bounds of excess prediction risk of divide-and-conquer estimator (1.4) for the functional linear regression model (1.1). These main results are based on several key assumptions, including the regularity conditions of the slope function and the functional explanatory variable. We begin by establishing a min-max lower bound for excess prediction risks.

3.1 Mini-max Convergence Lower Bound

In this subsection, we first introduce assumptions on the slope function β_0 and the random noise ϵ, then based on the two assumptions, we establish a min-max lower bound for excess prediction risks. We begin with the regularity assumption on the slope function β_0 which is expressed in terms of covariance operator L_C and operator T_* given in (2.9).

Assumption 1 (regularity condition of slope function) *The slope function β_0 in functional linear regression model* (1.1) *satisfies*

$$L_C^{1/2}\beta_0 = T_*^{\theta}(\gamma_0) \text{ with } 0 < \theta \leq 1/2 \text{ and } \gamma_0 \in \mathcal{L}^2(\mathcal{T}). \tag{3.1}$$

This assumption implies that $L_C^{1/2}\beta_0$ belongs to the range space of T_*^{θ} expressed as

$$\mathrm{ran}T_*^{\theta} := \left\{ f \in \mathcal{L}^2(\mathcal{T}) : \sum_{k\geq 1} \frac{\langle f, \varphi_k\rangle_{\mathcal{L}^2}^2}{\mu_k^{2\theta}} < \infty \right\},$$

where $\{\mu_k, \varphi_k\}_{k\geq 1}$ is the eigensystem of T_*. Then $\mathrm{ran}T_*^{\theta_1} \subseteq \mathrm{ran}T_*^{\theta_2}$ whenever $\theta_1 \geq \theta_2$. The regularity of functions within the range of T_*^{θ} is determined by the rate at which their expansion coefficients decrease, employing the set $\{\varphi_k\}_{k\geq 1}$. The stipulation in condition (3.1) signifies that the square of the inner product $\langle L_C^{1/2}\beta_0, \varphi_k\rangle$ in the $\mathcal{L}^2$ space diminishes more swiftly than the eigenvalues of T_* raised to the 2θ power. A larger value of θ thus correlates with more rapid attenuation rates, signifying higher regularities of $L_C^{1/2}\beta_0$. In particular, for $\theta = 0$ we have $\mathrm{ran}T_*^0 = \mathcal{L}^2(\mathcal{T})$ implying $\beta_0 \in \mathcal{L}^2(\mathcal{T})$ and $\beta_0 \in \mathcal{H}_K$ ensures regularity condition (3.1) is satisfied with $\theta = 1/2$ as $\mathrm{ran}T_*^{1/2} = \mathrm{ran}L_C^{1/2}L_K^{1/2}$. From this point of view, condition (3.1) allows $\beta_0 \notin \mathcal{H}_K$ which extends the previous regularity assumption in Yuan and Cai (2010) and Cai and Yuan (2012). This condition is also known as Hölder-type source condition involving the operator T_*, which is a classical smoothness assumption in the theory of inverse problems. Similar conditions defined by the operator L_K are widely used in the literature of learning theory (see, e.g., Bauer et al., 2007; Caponnetto and Vito, 2007; Smale and Zhou, 2007; Blanchard and Mücke, 2018). We will provide more discussions on Assumption 1 in Section 4.

Throughout of the paper, we assume the following noise condition.

Assumption 2 (noise condition) *The random noise* ϵ *in functional linear regression model* (1.1) *is independent of* X *satisfying* $\mathbb{E}[\epsilon] = 0$ *and* $\mathbb{E}[\epsilon^2] \leq \sigma^2$.

Now under Assumption 1 and 2, we propose the following theorem which establishes a mini-max lower bound for excess prediction risks. To this end, we also need to assume that $\{\mu_k\}_{k\geq 1}$, i.e., the eigenvalues of T_* (and T), satisfy a polynomially decaying condition. For two positive sequences $\{a_k\}_{k\geq 1}$ and $\{b_k\}_{k\geq 1}$, we say $a_k \lesssim b_k$ holds if there exits a constant $c > 0$ independent of k such that $a_k \leq c b_k, \forall k \geq 1$. In addition, $a_k \asymp b_k$ if and only if $a_k \lesssim b_k$ and $b_k \lesssim a_k$. For the sake of simplicity, we write $L_C^{1/2}\beta_0 \in \mathrm{ran} T_*^\theta$ if β_0 satisfies the regularity condition (3.1).

Theorem 2 (mini-max convergence lower bound) *Under Assumption 1 with* $0 < \theta \leq 1/2$ *and Assumption 2 with* $\sigma > 0$, *suppose that* $\{\mu_k\}_{k\geq 1}$ *satisfy* $\mu_k \asymp k^{-1/p}$ *for some* $0 < p \leq 1$. *Then excess prediction risks satisfy*

$$\lim_{\gamma\to 0} \inf_{N\to\infty} \lim_{\hat{\beta}_S} \inf \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\hat{\beta}_S) - \mathcal{R}(\beta_0) \geq \gamma N^{-\frac{2\theta}{2\theta+p}}\right\} = 1, \tag{3.2}$$

where the supremum is taken over all $\beta_0 \in \mathcal{L}^2(\mathcal{T})$ *satisfying* $L_C^{1/2}\beta_0 \in \mathrm{ran} T_*^\theta$ *and the infimum is taken over all possible predictors* $\hat{\beta}_S \in \mathcal{L}^2(\mathcal{T})$ *based on the training sample set* $S = \{(X_i, Y_i)\}_{i=1}^N$.

In Theorem 2 and subsequent statements, the case $p = 1$ corresponds to the case in which we only require $\{\mu_k\}_{k\geq 1}$ to be summable. The lower bound for $p = 1$ is also referred to as the capacity-independent optimum in some studies (see, for example, Yao et al., 2007). This means that the bound is optimal in the mini-max sense without the necessity for a capacity hypothesis, i.e., without the decaying condition of the eigenvalues $\{\mu_k\}_{k\geq 1}$.

3.2 Convergence Upper Bounds in Noised Case

In this subsection, we will establish three different upper bounds on the excess prediction risk of divide-and-conquer estimator (1.4) under Assumption 1 and 2. These upper bounds are based on three different regularity assumptions on the explanatory variable X, respectively. We first consider the upper bound on the convergence rate of the excess prediction risk (2.7) and show that the convergence rate of the lower bound established in Theorem 2 can be achieved by the divide-and-conquer estimator (1.4). The following assumption on the moment condition of the explanatory variable X plays a crucial role in establishing the upper bound of the convergence rate of (1.5).

Assumption 3 (regularity condition of explanatory variable I) *There exists a constant* $c_1 > 0$, *such that for any* $f \in \mathcal{L}^2(\mathcal{T})$,

$$\mathbb{E}\left[\langle X, f\rangle_{\mathcal{L}^2}^4\right] \leq c_1 \left[\mathbb{E}\langle X, f\rangle_{\mathcal{L}^2}^2\right]^2. \tag{3.3}$$

Assumption 3 has been introduced in Cai and Yuan (2012); Yuan and Cai (2010). Condition (3.3) articulates that the kurtosis of linear functionals applied to X remains constrained, a criterion particularly satisfied with $c_1 = 3$ in scenarios where X is modeled by a Gaussian process. For the convenience of further statements, define the effective dimension as

$$\mathcal{N}(\lambda) := \sum_{k\geq 1} \frac{\mu_k}{\lambda + \mu_k}, \tag{3.4}$$

where $\lambda > 0$ and $\{\mu_k\}_{k\geq 1}$ are non-negative eigenvalues of T (with geometric multiplicities) arranged in decreasing order. The effective dimension is widely used in the convergence analysis of kernel ridge regression (see, e.g., Caponnetto and Vito, 2007; Fischer and Steinwart, 2020; Lin et al., 2017; Zhang et al., 2015). Now under a polynomially decaying condition of eigenvalues $\{\mu_k\}_{k\geq 1}$, we can show in the following theorem that the convergence rate of the lower bound in Theorem 2 can be obtained by the divide-and-conquer RLS estimator $\overline{\beta}_{S,\lambda} = \frac{1}{m}\sum_{j=1}^m \hat{\beta}_{S_j,\lambda}$ with $S = \cup_{j=1}^m S_j = \{(X_i, Y_i)\}_{i=1}^N$ and $|S_j| = \frac{N}{m}$. We employ $o(\alpha_N)$ to denote a little-o sequence of $\{a_N\}_{N\geq 1}$ if $\lim_{N\to\infty} o(a_N)/a_N = 0$.

Theorem 3 (convergence upper bound I) *Under Assumption 1 with $0 < \theta \leq 1/2$, Assumption 2 with $\sigma > 0$ and Assumption 3 with $c_1 > 0$, suppose that $\{\mu_k\}_{k\geq 1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \leq 1$.*

1. *For $p/2 < \theta \leq 1/2$, there holds*

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim_{\beta_0} \sup \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma N^{-\frac{2\theta}{2\theta+p}}\right\} = 0 \tag{3.5}$$

provided that $m \leq o\left(N^{\frac{2\theta-p}{4\theta+2p}}\right)$ and $\lambda = N^{-\frac{1}{2\theta+p}}$.

2. *For $0 < \theta \leq p/2$, there holds*

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim_{\beta_0} \sup \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma N^{-\frac{\theta}{p}} (\log N)^{\frac{3\theta r}{p}}\right\} = 0 \tag{3.6}$$

provided that $m \leq (\log N)^r$ for some $r > 0$ and $\lambda = N^{-\frac{1}{2p}} (\log N)^{\frac{3r}{2p}}$, and

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim_{\beta_0} \sup \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma N^{-\frac{(1-r)\theta}{p}} \log N\right\} = 0 \tag{3.7}$$

provided that $m \leq N^r$ for some $0 \leq r < 1$ and $\lambda = N^{-\frac{1-r}{2p}} (\log N)^{\frac{1}{2\theta}}$.

Here the supremum is taken over all $\beta_0 \in \mathcal{L}^2(\mathcal{T})$ satisfying $L_C^{1/2}\beta_0 \in \mathrm{ran} T_*^\theta$ with $0 < \theta \leq 1/2$.

Actually, we will show that if the eigenvalue decay satisfies a polynomial upper bound of order $1/p$ with $0 < p < 1$ and if the regularity parameter θ satisfies $0 < \theta \leq p/2$, there holds

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim_{\beta_0} \sup \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma \lambda^{2\theta}\right\} = 0$$

provided that $m^2\lambda^{-2p} \leq o(N)$. From Theorem 3, the bound (3.5) implies when $\theta \in (p/2, 1/2]$, the excess prediction risk of $\overline{\beta}_{S,\lambda}$ attains the convergence rate of the lower bound given by Theorem 2 and is therefore rate-optimal. Additionally, if $\theta = p/2$, from (3.6), taking $m \leq (\log N)^r$ and $\lambda = N^{-\frac{1}{2p}}(\log N)^{\frac{3r}{2p}}$ with some $r > 0$ yields

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma N^{-\frac{1}{2}}(\log N)^{\frac{3r}{2}}\right\} = 0.$$

This convergence rate is also optimal up to a logarithmic factor. The bound (3.5) generalizes previous results of Cai and Yuan (2012), which only considered the case $\beta_0 \in \mathcal{H}_K$, to the model misspecification scenario $\beta_0 \notin \mathcal{H}_K$ and the divide-and-conquer estimators. Actually, when $\theta = 1/2$, taking $m = 1$ and $\lambda = N^{-\frac{1}{2\theta+p}}$, we recovery Theorem 2 of Cai and Yuan (2012) which establishes minimax upper bound for the estimator $\hat{\beta}_{S,\lambda}$ in (1.3) when $\beta_0 \in \mathcal{H}_K$.

We next introduce a higher-order moment condition on X such that one can establish the strong convergence in expectation. To this end, given a reproducing kernel K, we shall introduce various regularities of the explanatory variable X defined through its image under $L_K^{1/2}$. Recall that X is a random element taking values in $\mathcal{L}^2(\mathcal{T})$ with $\mathbb{E}[X] = 0$ and $\mathbb{E}\left[\|X\|_{\mathcal{L}^2}^2\right] < \infty$, and $\{\mu_k, \phi_k\}_{k\geq 1}$ is the eigensystem of T given by (2.9). Consider the principal component decomposition of $L_K^{1/2}X$ with respect to T (see for details, Ash and Gardner, 2014), which is expressed as

$$L_K^{1/2}X = \sum_{k\geq 1} \sqrt{\mu_k}\xi_k\phi_k \tag{3.8}$$

where the ξ_k's are zero-mean, uncorrelated real-valued random variables with $\mathbb{E}[\xi_k^2] = 1$. We assume the following moment condition to characterize the regularity of $L_K^{1/2}X$.

Assumption 4 (regularity condition of explanatory variable II) *For some integer $\ell \geq 2$, there exists a constant $\rho < \infty$ such that $\{\xi_k\}_{k\geq 1}$ in decomposition* (3.8) *satisfy $\sup_{k\geq 1} \mathbb{E}[\xi_k^{4\ell}] \leq \rho^{4\ell}$. Moreover, there exists a constant $c_2 > 0$ such that*

$$\mathbb{E}\left[\langle X, f\rangle_{\mathcal{L}^2}^8\right] \leq c_2^2\left[\mathbb{E}\langle X, f\rangle_{\mathcal{L}^2}^2\right]^4, \quad \forall f \in \mathcal{L}^2(\mathcal{T}). \tag{3.9}$$

Since $\mathbb{E}\left[\xi_k^2\right] = 1$, we always have $\rho \geq 1$. When X is a Gaussian random element taking value in $\mathcal{L}^2(\mathcal{T})$, Assumption 4 is satisfied for any integer $\ell \geq 2$. In fact, given an integer $\ell \geq 2$, the linear functionals of a Gaussian random element X satisfy

$$\mathbb{E}\left[\langle X, f\rangle_{\mathcal{L}^2}^{4\ell}\right] \leq (4\ell-1)!!\left[\mathbb{E}\langle X, f\rangle_{\mathcal{L}^2}^2\right]^{2\ell}, \quad \forall f \in \mathcal{L}^2(\mathcal{T}).$$

Then taking $f = L_K^{1/2}\phi_k$ implies Assumption 4 with $\rho = [(4\ell-1)!!]^{\frac{1}{4\ell}}$ and $c_2^2 = 105$ (by letting $\ell = 2$). We need condition (3.9) to bound $\mathbb{E}\left[\langle X, \beta_0 - L_K^{1/2}f_\lambda\rangle_{\mathcal{L}^2}^4\right]$ in the model misspecification scenario $\beta_0 \notin \mathcal{H}_K$, which is crucial in the estimation of

$$\mathscr{S}(S,\lambda) = \left\|L_C^{1/2}L_K^{1/2}\overline{f}_{S,\lambda} - L_C^{1/2}L_K^{1/2}f_\lambda\right\|_{\mathcal{L}^2}^2.$$

Now we can establish the following upper bounds of (2.7) in expectation under Assumption 1, 2 and 4.

Theorem 4 (convergence upper bound II) *Suppose that Assumption 1 is satisfied with $0 < \theta \leq 1/2$ and $\gamma_0 \in \mathcal{L}^2(\mathcal{T})$. Under Assumption 2 with $\sigma > 0$ and Assumption 4 with some integer $\ell \geq 2$ and $c_2 > 0$, take $\lambda \leq 1$, then if $2 \leq \ell < 8$, there holds*

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\leq 2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + 16\frac{\mathcal{N}(\lambda)}{N}(c_2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2) + 8c_2\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 \\
&\quad + b_1(\ell)\lambda^{\frac{\ell-8}{4}}\left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{4+2m}{N}\left(1 + \lambda^{2\theta}\mathcal{N}(\lambda)\right) \\
&\quad + b_2(\ell)\lambda^{\frac{\ell-8}{4}}\left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{4\sigma^2}{N}\mathcal{N}(\lambda). \qquad (3.10)
\end{aligned}
$$

If $\ell \geq 8$, there holds

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\leq 2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + 16\frac{\mathcal{N}(\lambda)}{N}(c_2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2) + 8c_2\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 \\
&\quad + b_1(\ell)\left[1 + \frac{m\mathcal{N}^2(\lambda)}{N} + \frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{4+2m}{N}\left(1 + \lambda^{2\theta}\mathcal{N}(\lambda)\right) \\
&\quad + b_2(\ell)\left[1 + \frac{m\mathcal{N}^2(\lambda)}{N} + \frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{4\sigma^2}{N}\mathcal{N}(\lambda). \qquad (3.11)
\end{aligned}
$$

Here $b_1(\ell)$ and $b_2(\ell)$ are constants only depending on ℓ and will be specified in the proof.

Recall that $a_N \lesssim b_N$ means that there exits a constant $c > 0$ independent of N such that $a_N \leq cb_N, \forall N \geq 1$. We obtain the following claims by using Theorem 4.

Corollary 5 *Under the assumptions of Theorem 4, suppose that $\{\mu_k\}_{k\geq 1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \leq 1$.*

1. *When $2 \leq \ell \leq 4$, there holds*

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\lesssim \max\left\{N^{\frac{2\theta(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{2\theta\ell(r-1)-8\theta}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{2\theta(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{4\theta\ell(r-1)-8\theta}{8+4p+8\theta+3p\ell}}\right\} \qquad (3.12)
\end{aligned}
$$

provided that

$$m \leq N^r \text{ for some } 0 \leq r \leq \tfrac{2\theta}{2\theta+p}$$

and

$$\lambda = \max\left\{N^{\frac{(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{\ell(r-1)-4}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{2\ell(r-1)-4}{8+4p+8\theta+3p\ell}}\right\}.$$

2. *When $5 \leq \ell \leq 7$, if $\frac{p\ell+8}{4\ell} \leq \theta \leq \frac{1}{2}$, then*

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim N^{-\frac{2\theta}{2\theta+p}} \qquad (3.13)$$

provided that

$$m \leq \min\left\{N^{\frac{8+p\ell-4p-4\theta\ell}{(4+2\ell)(2\theta+p)}}, N^{\frac{8+p\ell-8\theta-4\theta\ell}{(4+2\ell)(2\theta+p)}}\right\}$$

and

$$\lambda = N^{-\frac{1}{2\theta+p}};$$

if $0 < \theta < \frac{p\ell+8}{4\ell}$, *then*

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim \max\left\{N^{\frac{2\theta(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{2\theta\ell(r-1)-8\theta}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{2\theta(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{4\theta\ell(r-1)-8\theta}{8+4p+8\theta+3p\ell}}\right\} \quad (3.14)$$

provided that

$$m \leq N^r \text{ for some } 0 \leq r \leq \frac{2\theta}{2\theta+p}$$

and

$$\lambda = \max\left\{N^{\frac{(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{\ell(r-1)-4}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{2\ell(r-1)-4}{8+4p+8\theta+3p\ell}}\right\}.$$

3. *When* $\ell \geq 8$, *if* $\frac{p\ell+8}{2\ell+16} \leq \theta \leq \frac{1}{2}$, *then*

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim N^{-\frac{2\theta}{2\theta+p}} \quad (3.15)$$

provided that

$$m \leq \min\left\{N^{\frac{8+p\ell-4p-16\theta-2\theta\ell}{(12+\ell)(2\theta+p)}}, N^{\frac{8+p\ell-24\theta-2\theta\ell}{(12+\ell)(2\theta+p)}}\right\}$$

and

$$\lambda = N^{-\frac{1}{2\theta+p}};$$

if $0 < \theta < \frac{p\ell+8}{2\ell+16}$, *then*

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim \max\left\{N^{\frac{\theta(4+\ell)(r-1)}{4\theta+p\ell}}, N^{\frac{\theta\ell(r-1)-4\theta}{2p+4\theta+p\ell}}, N^{\frac{\theta(12+\ell)(r-1)}{4+4p+4\theta+p\ell}}, N^{\frac{\theta(8+\ell)(r-1)-4\theta}{4+6p+4\theta+p\ell}}\right\} \quad (3.16)$$

provided that

$$m \leq N^r \text{ for some } 0 \leq r \leq \frac{2\theta}{2\theta+p}$$

and

$$\lambda = \max\left\{N^{\frac{(4+\ell)(r-1)}{8\theta+2p\ell}}, N^{\frac{\ell(r-1)-4}{4p+8\theta+2p\ell}}, N^{\frac{(12+\ell)(r-1)}{8+8p+8\theta+2p\ell}}, N^{\frac{(8+\ell)(r-1)-4}{8+12p+8\theta+2p\ell}}\right\}.$$

According to Theorem 2, the expectation bounds (3.13) and (3.15) are minimax optimal. Due to the well-known Markov's inequality, convergence in expectation given by Theorem 4 and Corollary 5 is stronger, leading to bounds in a similar form as that of Theorem 3. However, the possible ranges of θ that achieve the optimal rates in (3.13) and (3.15), given respectively by $[\frac{p\ell+8}{4\ell}, 1/2]$ and $[\frac{p\ell+8}{2\ell+16}, 1/2]$ both of which are covered by $(p/2, 1/2]$, become smaller compared to the previous range of θ in the minimax bound (3.5). Moreover, we also observe from Corollary 5 that as the integer ℓ in Assumption 4 diverges to infinity, the possible ranges of θ that achieve the minimax expectation bounds will increase to $(p/2, 1/2]$ which is exactly the range of θ leading to the minimax bound (3.5). Motivated by this observation, we introduce another regularity condition on X to establish optimal expectation error bounds for any $\theta \in (0, 1/2]$.

Assumption 5 (regularity condition of explanatory variable III) *There exists a constant $\rho < \infty$ such that $\{\xi_k\}_{k\geq 1}$ in decomposition* (3.8) *satisfy* $\sup_{k\geq 1} |\xi_k| \leq \rho$ *and the fourth-order moment condition* (3.3) *is satisfied with $c_1 > 0$.*

One can verify that Assumption 5 holds true if the expansion of $L_K^{1/2} X$ in (3.8) is a summation of finite terms with each bounded ξ_k. Recall that the trace of operator T is given by

$$trace(T) := \sum_{k\geq 1} \mu_k. \tag{3.17}$$

Then we have the following theorem.

Theorem 6 (convergence upper bound III) *Suppose that Assumption 1 is satisfied with $0 < \theta \leq 1/2$ and $\gamma_0 \in \mathcal{L}^2(\mathcal{T})$. Under Assumption 2 with $\sigma > 0$ and Assumption 5 with $\rho, c_1 > 0$, take $\lambda \leq 1$, then there holds*

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\leq 2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + 16\frac{\mathcal{N}(\lambda)}{N}\left(c_1\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2\right) + 8c_1\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 \\
&\quad + c_3c_4\mu_1\frac{4+2m}{N\lambda^{2-2\theta}}\left(1 + \frac{m\mathcal{N}(\lambda)}{N}\right)\mathcal{N}^{\frac{1}{2}}(\lambda)\exp\left(-\frac{c_5 N}{2m\mathcal{N}(\lambda)}\right) \\
&\quad + c_4\mu_1\rho^2 trace(T)\frac{4\sigma^2}{N\lambda^2}\left(1 + \frac{m\mathcal{N}(\lambda)}{N}\right)\mathcal{N}^{\frac{1}{2}}(\lambda)\exp\left(-\frac{c_5 N}{2m\mathcal{N}(\lambda)}\right),
\end{aligned} \tag{3.18}$$

where c_3, c_4 and c_5 are universal constants which will be specified in the proof.

We further obtain a corollary of Theorem 6.

Corollary 7 *Under the assumptions of Theorem 6, suppose that $\{\mu_k\}_{k\geq 1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \leq 1$. Then there holds*

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim N^{-\frac{2\theta}{2\theta+p}} \tag{3.19}$$

provided that $m \leq o\left(\frac{N^{\frac{2\theta}{2\theta+p}}}{\log N}\right)$ and $\lambda = N^{-\frac{1}{2\theta+p}}$.

As far as we know, the expectation bound (3.19) establishes the first mini-max optimal rates for all possible $0 < \theta \leq 1/2$. One can refer to Section 4 for more discussions.

3.3 Convergence Upper Bounds in Noiseless Case

In this subsection, we establish fast convergence rates of the excess prediction risk (2.7) for noiseless functional linear model (i.e., $\epsilon = 0$ in (1.1)).

Theorem 8 (convergence upper bound IV) *Under Assumption 1 with $0 < \theta \le 1/2$, Assumption 2 with $\sigma = 0$ and Assumption 3 with $c_1 > 0$, suppose that $\{\mu_k\}_{k\ge1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \le 1$. Then for any $0 < \eta \le 1/2$, there holds*

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \ge \Gamma N^{-\frac{\theta(1-2\eta)}{p}}\right\} = 0 \tag{3.20}$$

provided that $m \le o\left(N^{\eta}\right)$ and $\lambda = N^{-\frac{1-2\eta}{2p}}$, where the supremum is taken over all $\beta_0 \in \mathcal{L}^2(\mathcal{T})$ satisfying $L_C^{1/2}\beta_0 \in \mathrm{ran}T_*^{\theta}$ with $0 < \theta \le 1/2$.

Follow from (3.20), given any $s > 0$ such that $0 < p < 2/s$ and $sp < \theta \le 1/2$, taking $0 < \eta \le \frac{1}{2} - \frac{sp}{2\theta}$, $m \le o(N^{\eta})$ and $\lambda = N^{-\frac{1-2\eta}{2p}}$ yields

$$\begin{aligned}
&\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \ge \Gamma N^{-s}\right\} \\
&\le \lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \ge \Gamma N^{-\frac{\theta(1-2\eta)}{p}}\right\} \\
&= 0,
\end{aligned}$$

where the inequality follows from $\frac{\theta(1-2\eta)}{p} \ge s$. The difference between noisy and noiseless models is significant: rates faster than N^{-1} for model (1.1) are impossible with non-zero additive noise, while we prove that the divided-and-conquer RLS estimators for the noiseless model can converge with arbitrarily fast polynomial rates when p is small enough. To our best knowledge, Theorem 8 and the related convergence rates (3.20) are new to the literature, constituting another contribution of this paper. We will prove all these results in Section 5.

4. Discussions and Comparisons

In this section, we first comment on Assumption 1 and then compare our convergence analysis with some related results. In the last, we review recent literature for noiseless linear model and point out some possible directions for future study.

4.1 Discussions on Assumption 1

Regularity condition (3.1) in Assumption 1 was first introduced by Fan et al. (2019) and then adopted in the subsequent works (see, for example, Chen et al., 2022). From the discussion in Section 3, we see that $\beta_0 \in \mathcal{H}_K$ implies condition (3.1) is satisfied with $\theta = 1/2$, while the former is equivalent to $\beta_0 = L_K^{1/2}\gamma_0$ for some $\gamma_0 \in \mathcal{L}^2(\mathcal{T})$. Actually, due to Theorem 3 in Chen et al. (2022), if $L_K \succeq \delta L_C^{\nu}$ for some $\delta > 0$ and $\nu > 0$, then for any $\beta_0 \in \mathcal{L}^2(\mathcal{T})$, there exists some $\gamma_0 \in \mathcal{L}^2(\mathcal{T})$ such that condition (3.1) is satisfied with

$\theta = 1/(2+2\nu)$. Here for any bounded self-adjoint operators A_1 and A_2 on $L^2(\mathcal{T})$, we write $A_1 \succeq A_2$ if $A_1 - A_2$ is nonnegative. As a special case when L_K and L_C can be simultaneously diagonalized, let $\{\rho_k\}_{k\geq 1}$ and $\{\lambda_k\}_{k\geq 1}$ be eigenvalues of L_K and L_C respectively (both are sorted in decreasing order with geometric multiplicities). When $\rho_k \asymp k^{-1/\omega}$ with $\omega > 1$ and $\lambda_k \asymp k^{-1/\tau}$ with $\tau > 1$, then $\beta_0 \in \mathrm{ran} L_K^s$ for some $s \in [0, 1/2]$ implies condition (3.1) is satisfied with $\theta = (\omega + 2s\tau)/(2\omega + 2\tau)$, where $\mathrm{ran} L_K^s$ denotes the range space of L_K^s. When K is an analytic kernel on $\mathcal{T}$, the eigenvalues of L_K decay exponentially, and then condition (3.1) can be satisfied for θ arbitrarily close to $1/2$ (but still strictly less than $1/2$). From the discussions above, Assumption 1 is mild and provides an intrinsic measurement for the complexity of the prediction problem through the regularity condition (3.1). Recently, under Assumption 1, Assumption 2 and some boundedness condition on K and C, the works Chen et al. (2022) and Guo et al. (2023) apply stochastic gradient descent to solve functional linear regression model (1.1) and establish convergence rates for prediction and estimation errors.

4.2 Comparisons with Relevant Results

Convergence performance of kernel ridge regression and its variants in model misspecification scenario has been intensively studied by many recent works (see, e.g., Pillaud et al., 2018; Shi, 2019; Fischer and Steinwart, 2020; Lin and Cevher, 2020; Lin et al., 2020; Sun and Wu, 2021). Among all available literature, the work Fischer and Steinwart (2020) obtained the best known convergence rates by applying the integral operator techniques combined with an embedding property (see condition (EMB) in Fischer and Steinwart, 2020). As far as we know, our paper is the first work to consider functional linear regression in a model misspecification scenario. To make a further comparison, we first introduce an embedding condition equivalent to the one in Fischer and Steinwart (2020) (i.e., condition (4.1) in this paper) under the functional linear regression setting. Then we apply this condition to derive convergence rates and compare them with related results in Section 3.

Assumption 6 (regularity condition of explanatory variable IV) *There exist constants $\kappa > 0$ and $0 < t \leq 1$ such that $\{\xi_k\}_{k\geq 1}$ in decomposition (3.8) satisfy*

$$\sum_{k\geq 1} \mu_k^t \xi_k^2 \leq \kappa^2. \tag{4.1}$$

Moreover, the fourth-order moment condition (3.3) is satisfied with some $c_1 > 0$.

Condition (4.1) actually describes the $\mathcal{L}^\infty$−embedding property of $T^{(t-1)/2} L_K^{1/2} X$ for $0 < t \leq 1$. Then we obtain the following result which also deserve attention in its own right.

Theorem 9 (convergence upper bound V) *Under Assumption 1 with $0 < \theta \leq 1/2$, Assumption 2 with $\sigma > 0$ and Assumption 6 with $0 < t \leq 1$, suppose that $\{\mu_k\}_{k\geq 1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \leq 1$.*

1. *When $\max\{0, t/2 - p/2\} < \theta \leq 1/2$, then*

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim N^{-\frac{2\theta}{2\theta+p}} \tag{4.2}$$

provided that

$$m \leq o\left(\frac{N^{\frac{2\theta+p-t}{2\theta+p}}}{\log N}\right) \text{ and } \lambda = N^{-\frac{1}{2\theta+p}}.$$

2. *When* $0 < \theta \leq \max\{0, t/2 - p/2\}$*, then*

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim N^{-\frac{2\theta}{t}}(\log N)^{-\frac{4\theta}{t}} \tag{4.3}$$

provided that

$$m \leq o(\log N) \text{ and } \lambda = N^{-\frac{1}{t}}(\log N)^{-\frac{2}{t}}.$$

The proof of Theorem 9 is also postponed to Section 5. Condition (4.1) characterizes the regularity of $L_K^{1/2}X$ through the parameter $t \in (0,1]$, of which the most general case is taking $t = 1$, i.e., $\sum_{k=1}^{\infty}\mu_k\xi_k^2 \leq \kappa^2$, or equivalently,

$$\left\|L_K^{1/2}X\right\|_{\mathcal{L}^2} \leq \kappa. \tag{4.4}$$

When $t = 1$, we obtain the mini-max rates in expectation for $\theta \in (1/2 - p/2, 1/2]$ from bound (4.2). However, as $p \downarrow 0$, which implies the eigenvalues of T_* decay even faster, the rate-optimal interval of θ is getting smaller. It seems unreasonable that higher regularity of T_* could instead reduce the possible range of θ that leads to the optimal convergence. This phenomenon is widely observed in the convergence analysis of regularized kernel regression for the model misspecification scenario (see, e.g., Shi, 2019; Fischer and Steinwart, 2020; Lin and Cevher, 2020; Lin et al., 2020). Note that verifying the embedding condition (4.1) for $t < 1$ is highly non-trivial. This condition is automatically satisfied for all $0 < t \leq 1$ if the expansion of $L_K^{1/2}X$ in (3.8) is a summation of finite terms with each bounded ξ_k. However, it is a wide-open question whether this condition holds for more general cases. It is also pointed out by Fischer and Steinwart (2020) that how to obtain the optimal rates for $t > p$ and $\theta \in (0, t/2 - p/2]$ is an outstanding problem that can not be addressed by introducing the embedding condition. Comparing Assumption 6 to Assumption 4 in Theorem 4 and Corollary 5, it is difficult to tell which one is more restrictive. But we think Assumption 4 is relatively more adaptive for functional linear regression model (1.1), since Assumption 6 excludes the most important case when X is the Gaussian random element in $\mathcal{L}^2(\mathcal{T})$. As we discussed in Section 3, Gaussian random element satisfies Assumption 4 for any integer $\ell \geq 2$. Moreover, due to Corollary 5, higher regularities indicated by larger ℓ in Assumption 4 or smaller p in the eigenvalue decaying of T_* will result in larger range of θ in which the estimators are rate-optimal. We believe that convergence analysis based on Assumption 4 is more insightful from this perspective. We then illustrate that in most cases, the index p can be close to zero arbitrarily if one of the kernels K and C is smooth enough. To this end, we need the following lemma.

Lemma 10 *Consider two nonnegative and compact operators* L_A *and* L_B *on a separable Hilbert space* $\mathcal{H}$*. Assume* $\overline{\operatorname{ran}(L_A^{1/2})} = \mathcal{H}$*, then we have*

$$\rho_k(L_A^{1/2}L_BL_A^{1/2}) \leq \rho_k(L_B)\|L_A\|,$$

where the $\rho_k(L_A^{1/2} L_B L_A^{1/2})$ *and* $\rho_k(L_B)$ *denote the k-th eigenvalue (sorted in decreasing order) of operators* $L_A^{1/2} L_B L_A^{1/2}$ *and* L_B*, respectively.*

We include its proof in Appendix A for the sake of completeness. Following from Lemma 10 with $\mathcal{H} = \mathcal{L}^2(\mathcal{T})$ and the fact that $\overline{\mathcal{H}_K} = \mathcal{L}^2(\mathcal{T})$, or equivalently $\overline{\text{ran}(L_K^{1/2})} = \mathcal{L}^2(\mathcal{T})$, we have $\mu_k = \rho_k(L_K^{1/2} L_C L_K^{1/2}) \leq \rho_k(L_C)\|L_K\| \lesssim \rho_k(L_C)$. Moreover, if $\overline{\text{ran}(L_C^{1/2})} = \mathcal{L}^2(\mathcal{T})$, one can deduce $\mu_k = \rho_k(L_C^{1/2} L_K L_C^{1/2}) \leq \rho_k(L_K)\|L_C\| \lesssim \rho_k(L_K)$ by the same argument. For example, when $\mathcal{T} = \mathbb{R}$ and K is the reproducing kernel of fractional Sobolev space $W^{\beta,2}(\mathbb{R})$ with $\beta > 1/2$, we have $\mu_k \lesssim \rho_k(L_K) \asymp k^{-2\beta}$ and then $p \leq \frac{1}{2\beta}$ can arbitrarily approach zero if K is smooth enough, i.e., β is sufficiently large. Another notable example is that when $\mathcal{T} = [0,1]^D$ for some integer $D \geq 1$ and X is a Gaussian random element in $\mathcal{L}^2(\mathcal{T})$ with zero mean and covariance kernel $C_\gamma : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ (which is called a square-exponential kernel) defined by $C_\gamma(x,x') := \exp(-\frac{\|x-x'\|^2}{\gamma^2}), \forall x, x' \in \mathcal{T}$, i.e., $X \sim \mathcal{N}(0, L_{C_\gamma})$. Here $\gamma > 0$ is a constant and L_{C_γ} denotes the covariance operator induced by C_γ. According to the existing literature about Gaussian process (see, for example, Kanagawa et al., 2018), we know that $\{\rho_k(L_{C_\gamma})\}_{k\geq 1}$ enjoys an exponential decay. For this case, we can prove that the divided-and-conquer RLS estimators are mini-max optimal for all possible $\theta \in (0, 1/2]$ according to Corollary 5.

We now compare Theorem 9 with Corollary 7 of Theorem 6. Under the uniformly boundedness condition on $\{\xi_k\}_{k\geq 1}$ in Assumption 5, we simplify the embedding condition (4.1) by only requiring the sequence $\{\mu_k^t\}_{k\geq 1}$ to be summable, i.e., $\sum_{k\geq 1} \mu_k^t < \infty$, which is satisfied for $t = p + \epsilon$ if $\mu_k \lesssim k^{-1/p}$. Here $\epsilon > 0$ can be arbitrarily small. Therefore, under the same assumptions of Corollary 7, the first claims in Theorem 6 ensures that for all sufficiently small $\epsilon > 0$ and $\epsilon/2 < \theta \leq 1/2$, there holds

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim N^{-\frac{2\theta}{2\theta+p}}$$

with $\lambda = N^{-\frac{1}{2\theta+p}}$ and $m \leq o\left(\frac{N^{\frac{2\theta-\epsilon}{2\theta+p}}}{\log N}\right)$. Since one can choose an arbitrarily small $\epsilon > 0$, the above result actually indicates the rate-optimal convergence for all $0 < \theta \leq 1/2$. We see from Corollary 7 in Section 3 that, under Assumption 5, one can obtain the same convergence result with a slightly better estimate on m which only requires $m \leq o\left(\frac{N^{\frac{2\theta}{2\theta+p}}}{\log N}\right)$.

When we finished this paper, we found that the work Tong (2021) also studies the divide and conquer functional linear regression but under a regularity condition of slope function different from (3.1) which actually requires $\beta_0 \in \mathcal{H}_K$, and a boundedness assumption equivalent to that (4.1) is satisfied with $t = 1$. And we also note that to achieve optimal convergence rate under condition $\beta_0 \in \mathcal{H}_K$, Theorem 2.1 in Tong (2021) requires the number of partitions $m = 1$, while with an additional assumption that the fourth-moment condition (3.3) is satisfied, Theorem 9 in this paper allows the number of partitions $m \leq o\left(\frac{N^{\frac{p}{1+p}}}{\log N}\right)$.

4.3 Relevant Works on Noiseless Linear Model

Recent works have intensively investigated the performance of various estimators within the context of a noiseless linear model. In typical learning tasks such as image classification, where human error is nearly impossible (e.g., misidentifying images of dogs as cats), it is reasonable to consider the output unambiguous for a given input. Consequently, many algorithms have incorporated noiseless models into these learning tasks. Theoretical analysis of noiseless models first appeared in studies on classification problems. For instance, Smale and Zhou (2007) demonstrates that, compared to models with noise, the convergence analysis of binary classification problems in noiseless models may exhibit a phenomenon known as "super convergence", where the convergence rate can be faster than N^{-1}. The studies by Jun et al. (2019); Sun and Wu (2021) explore the application of kernel-regularized least squares, revealing an improved rate of convergence for noiseless data relative to noisy scenarios. Furthermore, Berthier et al. (2020) delves into the utilization of stochastic gradient descent for addressing the noiseless linear model in a general Hilbert space, albeit concentrating solely on scenarios where the optimal predictor resides within this space. As far as we know, the convergence of estimator in RKHS as well as its divide-and-conquer counterpart has not been considered in the context of noiseless functional linear model. We establish the first convergence result in this setting when the optimal predictor is outside of the underlying RKHS. The framework and estimations developed in this paper can be extended to study more complex models of nonparametric supervised learning, such as models in Szabó et al. (2016); Guo et al. (2023); Mao (2024); Meunier et al. (2022), which we leave as future work.

5. Convergence Analysis

In this section, we first derive the upper bounds of convergence rates presented in Theorem 3 and Theorem 8. Then we establish the upper bounds in expectation presented in Theorem 4, Theorem 6, Theorem 9 and their corollaries. Last we prove the mini-max lower bound in Theorem 2.

5.1 Upper Rates and Upper Bounds

Recalling the decomposition (2.8), one can bound $\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)$ through estimating $\mathscr{S}(S,\lambda) = \left\| L_C^{1/2} L_K^{1/2} \overline{f}_{S,\lambda} - L_C^{1/2} L_K^{1/2} f_\lambda \right\|_{\mathcal{L}^2}^2$ and $\mathscr{A}(\lambda) = \left\| L_C^{1/2} L_K^{1/2} f_\lambda - L_C^{1/2} \beta_0 \right\|_{\mathcal{L}^2}^2$, respectively. We apply the following lemma to estimate $\mathscr{A}(\lambda)$.

Lemma 11 *Suppose Assumption 1 is satisfied with $0 < \theta \leq 1/2$ and $\gamma_0 \in \mathcal{L}^2(\mathcal{T})$. Then for any $\lambda > 0$, there holds*

$$\mathscr{A}(\lambda) \leq \lambda^{2\theta} \|\gamma_0\|_{\mathcal{L}^2}^2. \tag{5.1}$$

Proof Write $\gamma_0 = \sum_{k\geq 1} a_k \varphi_k$, according to singular value decomposition of T_* in (2.9), we have $L_C^{1/2}\beta_0 = T_*^\theta(\gamma_0) = \sum_{k\geq 1} \mu_k^\theta a_k \varphi_k$ and

$$L_C^{1/2} L_K^{1/2} f_\lambda = L_C^{1/2} L_K^{1/2} \left(\lambda I + T\right)^{-1} L_K^{1/2} L_C \beta_0 = \sum_{k=1}^{\infty} \frac{\mu_k^{1+\theta}}{\lambda + \mu_k} a_k \varphi_j.$$

Therefore,

$$\begin{aligned}\mathscr{A}(\lambda) &= \left\| L_C^{1/2}(L_K^{1/2} f_\lambda - \beta_0) \right\|_{\mathcal{L}^2}^2 \\ &= \sum_{k=1}^{\infty} \left(\frac{\mu_k^{1+\theta}}{\lambda + \mu_k} - \mu_j^\theta \right)^2 a_j^2 \\ &= \sum_{k=1}^{\infty} \frac{\lambda^2 \mu_k^{2\theta}}{(\lambda + \mu_k)^2} a_k^2.\end{aligned}$$

While we see that for $0 < \theta \le 1/2$,

$$\frac{t^\theta}{\lambda + t} \le \theta^\theta (1-\theta)^{1-\theta} \lambda^{\theta - 1} \le \lambda^{\theta - 1}, \quad \forall t > 0,$$

which implies that

$$\mathscr{A}(\lambda) = \sum_{k=1}^{\infty} \frac{\lambda^2 \mu_k^{2\theta}}{(\lambda + \mu_k)^2} a_k^2 \le \lambda^{2\theta} \sum_{k=1}^{\infty} a_k^2 = \lambda^{2\theta} \left\| \gamma_0 \right\|_{\mathcal{L}^2}^2.$$

The proof is then finished. ■

In the rest part of this subsection, we focus on estimating $\mathscr{S}(S, \lambda)$. Recall that $S = \cup_{j=1}^m S_j$ with $S_j \cap S_k = \emptyset$ for $j \neq k$, the empirical operator $T_{\mathbf{X}_j}$ is defined with $\mathbf{X}_j = \{X_i : (X_i, Y_i) \in S_j\}$ according to (2.6). For any $j = 1, 2, \ldots, m$, define the event

$$\mathcal{U}_j = \left\{ \mathbf{X}_j : \left\| (\lambda I + T)^{-1/2} (T_{\mathbf{X}_j} - T)(\lambda I + T)^{-1/2} \right\| \ge 1/2 \right\},$$

and denote its complement by $\mathcal{U}_j^c$. Let $\mathcal{U} = \cup_{j=1}^m \mathcal{U}_j$ be the union of above events. Then the complement of $\mathcal{U}$ is given by $\mathcal{U}^c = \cap_{j=1}^m \mathcal{U}_j^c$. Hereafter, let $\mathbb{I}_{\mathcal{E}}$ denote the indicator function of the event $\mathcal{E}$ and $\mathbb{P}(\mathcal{E}) = \mathbb{E}\left[\mathbb{I}_{\mathcal{E}}\right]$. We first give the following estimation

$$\begin{aligned}&\left\| (\lambda I + T)^{1/2} (\lambda I + T_{\mathbf{X}_j})^{-1} (\lambda I + T)^{1/2} \right\| \mathbb{I}_{\mathcal{U}_j^c} \\ &= \left\| (I - (\lambda I + T)^{-1/2} (T - T_{\mathbf{X}_j})(\lambda I + T)^{-1/2})^{-1} \right\| \mathbb{I}_{\mathcal{U}_j^c} \\ &\overset{(*)}{\le} 1 + \sum_{k=1}^{\infty} \left\| (\lambda I + T)^{-1/2} (T - T_{\mathbf{X}_j})(\lambda I + T)^{-1/2} \right\|^k \mathbb{I}_{\mathcal{U}_j^c} \\ &\le 1 + \sum_{k=1}^{\infty} \frac{1}{2^k} = 2,\end{aligned} \tag{5.2}$$

where inequality $(*)$ follows by expanding the inverse in Neumann series.

The following lemma plays a crucial role in bounding the convergence rate of $\mathscr{S}(S, \lambda)$.

Lemma 12 *For any $m \ge 1$, there holds*

$$\begin{aligned}&\mathbb{E}\left[\mathscr{S}(S, \lambda) \mathbb{I}_{\mathcal{U}^c}\right] \\ &\quad \le \frac{1}{m} \mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} (\hat{f}_{S_1, \lambda} - f_\lambda) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c} \right] + \left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[(\hat{f}_{S_1, \lambda} - f_\lambda) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2.\end{aligned} \tag{5.3}$$

Proof When $m \geq 2$, as

$$\mathscr{S}(S,\lambda) = \left\| L_C^{1/2} L_K^{1/2} \overline{f}_{S,\lambda} - L_C^{1/2} L_K^{1/2} f_\lambda \right\|_{\mathcal{L}^2}^2 = \left\| \frac{1}{m} \sum_{i=1}^m L_C^{1/2} L_K^{1/2} \hat{f}_{S_i,\lambda} - L_C^{1/2} L_K^{1/2} f_\lambda \right\|_{\mathcal{L}^2}^2,$$

then we have

$$\begin{aligned}
&\mathbb{E}[\mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}^c}] = \mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} \left(\frac{1}{m} \sum_{i=1}^m \hat{f}_{S_i,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}^c} \right] \\
&\overset{(i)}{=} \frac{1}{m^2} \sum_{i=1}^m \mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_i,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}^c} \right] \\
&\quad + \frac{1}{m^2} \sum_{i \neq j} \mathbb{E}\left[\left\langle L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_i,\lambda} - f_\lambda \right), L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_j,\lambda} - f_\lambda \right) \right\rangle_{\mathcal{L}^2} \mathbb{I}_{\mathcal{U}^c} \right] \\
&\overset{(ii)}{=} \frac{1}{m} \mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c} \right] \mathbb{P}\left(\cap_{j=2}^m \mathcal{U}_j^c \right) \\
&\quad + \frac{m(m-1)}{m^2} \mathbb{E}\left[\left\langle L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right), L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_2,\lambda} - f_\lambda \right) \right\rangle_{\mathcal{L}^2} \mathbb{I}_{\mathcal{U}_1^c} \mathbb{I}_{\mathcal{U}_2^c} \right] \mathbb{P}\left(\cap_{j=3}^m \mathcal{U}_j^c \right) \\
&\overset{(iii)}{\leq} \frac{1}{m} \mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c} \right] + \left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[\left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2.
\end{aligned}$$

Here equality (i) follows from the binomial expansion. Equality (ii) uses the fact that $\mathbb{I}_{\mathcal{U}^c} = \mathbb{I}_{\mathcal{U}_1^c} \mathbb{I}_{\mathcal{U}_2^c} \cdots \mathbb{I}_{\mathcal{U}_m^c}$ and for any $1 \leq i \neq j \leq m$, $\left(\hat{f}_{S_i,\lambda} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_i^c}$ and $\left(\hat{f}_{S_j,\lambda} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_j^c}$ are independent and identically distributed random elements. Inequality (iii) is from

$$\mathbb{E}\left[\langle L_C^{1/2} L_K^{1/2} (\hat{f}_{S_1,\lambda} - f_\lambda), L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_2,\lambda} - f_\lambda \right) \rangle_{\mathcal{L}^2} \mathbb{I}_{\mathcal{U}_1^c} \mathbb{I}_{\mathcal{U}_2^c} \right] = \left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[\left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2$$

This completes the proof. ■

For simplicity of notation, in the rest of this paper, we always denote

$$n := |S_1| = \frac{N}{m} \qquad \textit{and} \qquad \{(X_{1,i}, Y_{1,i})\}_{i=1}^n := S_1. \tag{5.4}$$

We establish the following bounds on the right hand side of (5.3) in Lemma 12.

Lemma 13 *Suppose that Assumption 1 is satisfied with* $0 < \theta \leq 1/2$ *and* $\gamma_0 \in \mathcal{L}^2(\mathcal{T})$, *Assumption 2 is satisfied with* $\sigma > 0$ *and Assumption 3 is satisfied with* $c_1 > 0$. *Then there hold*

$$\mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c} \right] \leq 8 \frac{m}{N} \mathcal{N}(\lambda) \left(c_1 \lambda^{2\theta} \|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2 \right) \tag{5.5}$$

and

$$\left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[\left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2 \leq 4 c_1 \frac{m}{N} \mathcal{N}(\lambda) \lambda^{2\theta} \|\gamma_0\|_{\mathcal{L}^2}^2, \tag{5.6}$$

where $\mathcal{N}(\lambda)$ *is the effective dimension given by* (3.4).

Proof We first prove the second inequality (5.6). Recalling (5.4), we can write

$$\hat{f}_{S_1,\lambda} = (\lambda I + T_{\mathbf{X}_1})^{-1} \frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} Y_{1,i}.$$

Then

$$\begin{aligned}
&\left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[\left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2 \\
&= \left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[\left((\lambda I + T_{\mathbf{X}_1})^{-1} \frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} Y_{1,i} - f_\lambda \right) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2 \\
&\overset{(i)}{=} \left\| L_C^{1/2} L_K^{1/2} \mathbb{E}\left[(\lambda I + T_{\mathbf{X}_1})^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda \right) \mathbb{I}_{\mathcal{U}_1^c} \right] \right\|_{\mathcal{L}^2}^2 \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c} \right] \\
&\leq \left\| L_C^{1/2} L_K^{1/2} (\lambda I + T)^{-1/2} \right\|^2 \left\| (\lambda I + T)^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} (\lambda I + T)^{1/2} \right\|^2 \mathbb{I}_{\mathcal{U}_1^c} \\
&\quad \times \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right] \\
&\overset{(iii)}{\leq} 4\mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right].
\end{aligned}$$

Here equality (i) follows from the fact that ϵ is a centered random variable independent of X, inequality (ii) uses Jensen's inequality, and inequality (iii) is due to (5.2) and the calculation that

$$\begin{aligned}
\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T)^{-1/2} \right\|^2 &= \left\| (\lambda I + T)^{-1/2} L_K^{1/2} L_C L_K^{1/2} (\lambda I + T)^{-1/2} \right\| \\
&= \left\| (\lambda I + T)^{-1/2} T (\lambda I + T)^{-1/2} \right\| \leq 1.
\end{aligned}$$

Note that for any $1 \leq i \leq n$, $L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda$ is a zero-mean random element. Then we have

$$\begin{aligned}
&\mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} \left(\frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} \left(L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right] \\
&\leq \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} \right\|_{\mathcal{L}^2}^2 \right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{\infty}\mathbb{E}\left[\left\langle(\lambda I+T)^{-1/2}L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^2\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^2\right]\\
&\overset{(i)}{\le} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{\infty}\left[\mathbb{E}\left\langle(\lambda I+T)^{-1/2}L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}\left[\mathbb{E}\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}\\
&\overset{(ii)}{\le} \frac{c_1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{\infty}\mathbb{E}\left[\left\langle(\lambda I+T)^{-1/2}L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^2\right]\mathbb{E}\left[\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^2\right]\\
&= \frac{c_1}{n}\sum_{j=1}^{\infty}\frac{1}{\lambda+\mu_j}\langle T\phi_j,\phi_j\rangle_{\mathcal{L}^2}\left\|L_C^{1/2}\left(\beta_0-L_K^{1/2}f_\lambda\right)\right\|_{\mathcal{L}^2}^2\\
&\overset{(iii)}{=} \frac{c_1}{n}\mathcal{N}(\lambda)\mathscr{A}(\lambda)\overset{(iv)}{\le} c_1\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2.
\end{aligned}
$$

Here $\{\phi_j\}_{j=1}^{\infty}$ is given by the singular value decomposition of T in (2.9). Inequality (i) uses Cauchy-Schwartz inequality. Inequality (ii) is from Assumption 3. Equality (iii) follows from $\mathscr{A}(\lambda)=\left\|L_C^{1/2}\left(\beta_0-L_K^{1/2}f_\lambda\right)\right\|_{\mathcal{L}^2}^2$ and the calculation that $\sum_{j=1}^{\infty}\frac{1}{\lambda+\mu_j}\langle T\phi_j,\phi_j\rangle_{\mathcal{L}^2}=\sum_{j=1}^{\infty}\frac{\mu_j}{\lambda+\mu_j}=\mathcal{N}(\lambda)$. Inequality (iv) is due to Lemma 11 and $n=N/m$. Combining the above two estimations, we have completed the proof of (5.6).

Next we prove the first inequality (5.5). According to the expression of $\hat{f}_{S_1,\lambda}$ and the triangular inequality, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\hat{f}_{S_1,\lambda}-f_\lambda\right)\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1^c}\right]\\
&\le 2\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}L_K^{1/2}X_{1,i}\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}-\lambda f_\lambda\right)\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1^c}\right] \qquad (5.5a)\\
&\quad+2\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}L_K^{1/2}X_{1,i}\epsilon_{1,i}\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1^c}\right], \qquad (5.5b)
\end{aligned}
$$

where $\epsilon_{1,i}:=Y_{1,i}-\langle\beta_0,X_{1,i}\rangle_{\mathcal{L}^2}$. We have bounded the term (5.5a) in the proof of (5.6), which is given by

$$
\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}L_K^{1/2}X_{1,i}\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}-\lambda f_\lambda\right)\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1^c}\right]\\
&\le 4c\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2.
\end{aligned} \qquad (5.7)
$$

Note that for any $1\le i\le n$, $L^{1/2}X_{1,i}\epsilon_{1,i}$ is also a zero-mean random element. Analogously, one can bound (5.5b) as

$$
\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^{n}L_K^{1/2}X_{1,i}\epsilon_{1,i}\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1^c}\right]
$$

$$\leq \left\| L_C^{1/2} L_K^{1/2} (\lambda I + T)^{-1/2} \right\|^2 \left\| (\lambda I + T)^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} (\lambda I + T)^{1/2} \right\|^2 \mathbb{I}_{\mathcal{U}_1^c}$$
$$\times \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} \frac{1}{n} \sum_{i=1}^n L_K^{1/2} X_{1,i} \epsilon_{1,i} \right\|_{\mathcal{L}^2}^2 \right]$$
$$\leq 4\mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} \frac{1}{n} \sum_{i=1}^n L_K^{1/2} X_{1,i} \epsilon_{1,i} \right\|_{\mathcal{L}^2}^2 \right] = \frac{4}{n^2} \sum_{i=1}^n \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \epsilon_{1,i}^2 \right]$$
$$\overset{(*)}{\leq} \frac{4\sigma^2}{n^2} \sum_{i=1}^n \sum_{j=1}^\infty \mathbb{E}\left[\left\langle (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i}, \phi_j \right\rangle_{\mathcal{L}^2}^2 \right] = \frac{4\sigma^2}{n} \sum_{j=1}^\infty \frac{1}{\lambda + \mu_j} \langle T\phi_j, \phi_j \rangle_{\mathcal{L}^2} = 4\frac{m}{N} \mathcal{N}(\lambda) \sigma^2.$$

Here inequality $(*)$ is from Assumption 2.

Then we obtain inequality (5.5) and the proof of Lemma 13 is finished. ■

We also need the following lemma to estimate the probability of event $\mathcal{U}_1$. Recall that $\mathcal{U}_1$ is defined as

$$\mathcal{U}_1 = \left\{ \mathbf{X}_1 : \left\| (\lambda I + T)^{-1/2} (T_{\mathbf{X}_1} - T)(\lambda I + T)^{-1/2} \right\| \geq 1/2 \right\}.$$

Lemma 14 *Suppose that Assumption 3 is satisfied with $c_1 > 0$, then*

$$\mathbb{P}(\mathcal{U}_1) \leq 4c_1 \frac{m}{N} \mathcal{N}^2(\lambda), \tag{5.8}$$

where $\mathcal{N}(\lambda)$ is the effective dimension given by (3.4).

Proof Recall (5.4). We first bound $\mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} (T_{\mathbf{X}_1} - T)(\lambda I + T)^{-1/2} \right\|^2 \right]$ as

$$\mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} (T_{\mathbf{X}_1} - T) (\lambda I + T)^{-1/2} \right\|^2 \right]$$
$$\overset{(i)}{\leq} \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} (T_{\mathbf{X}_1} - T) (\lambda I + T)^{-1/2} \right\|_{HS}^2 \right]$$
$$= \sum_{j=1}^\infty \sum_{k=1}^\infty \mathbb{E}\left[\left\langle (\lambda I + T)^{-1/2} \left(\frac{1}{n} \sum_{i=1}^n L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} - T \right) (\lambda I + T)^{-1/2} \phi_j, \phi_k \right\rangle_{\mathcal{L}^2}^2 \right]$$
$$\overset{(ii)}{\leq} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^\infty \sum_{k=1}^\infty \frac{1}{\lambda + \mu_j} \frac{1}{\lambda + \mu_k} \mathbb{E}\left[\left\langle L_K^{1/2} X_{1,i}, \phi_j \right\rangle_{\mathcal{L}^2}^2 \left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2}^2 \right]$$
$$\overset{(iii)}{\leq} \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^\infty \sum_{k=1}^\infty \frac{1}{\lambda + \mu_j} \frac{1}{\lambda + \mu_k} \left[\mathbb{E} \left\langle L_K^{1/2} X_{1,i}, \phi_j \right\rangle_{\mathcal{L}^2}^4 \right]^{\frac{1}{2}} \left[\mathbb{E} \left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2}^4 \right]^{\frac{1}{2}}$$
$$\overset{(iv)}{\leq} \frac{c_1}{n^2} \sum_{i=1}^n \sum_{j=1}^\infty \sum_{k=1}^\infty \frac{1}{\lambda + \mu_j} \frac{1}{\lambda + \mu_k} \mathbb{E}\left[\left\langle L_K^{1/2} X_{1,i}, \phi_j \right\rangle_{\mathcal{L}^2}^2 \right] \mathbb{E}\left[\left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2}^2 \right]$$
$$= \frac{c_1}{n} \sum_{j=1}^\infty \sum_{k=1}^\infty \frac{1}{\lambda + \mu_j} \frac{1}{\lambda + \mu_k} \langle T\phi_j, \phi_j \rangle_{\mathcal{L}^2} \langle T\phi_k, \phi_k \rangle_{\mathcal{L}^2}$$

$$= \frac{c_1}{n} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \frac{\mu_j}{\lambda + \mu_j} \frac{\mu_k}{\lambda + \mu_k} = c_1 \frac{m}{N} \mathcal{N}^2(\lambda).$$

Here $\{\phi_j\}_{j=1}^{\infty}$ is given by the singular value decomposition of T in (2.9). Inequality (i) follows from (2.2). Inequality (ii) is from the fact that for any $1 \le i \le n$, $L_K^{1/2} X_i \otimes L_K^{1/2} X_i - T$ is a zero-mean random element. Inequality (iii) uses Cauchy-Schwartz inequality. Inequality (iv) applies Assumption 3.

Combining the above estimation with Chebyshev inequality, we obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{U}_1) &= \mathbb{P}\left(\left\{\mathbf{X}_1 : \left\|(\lambda I + T)^{-1/2} (T_{\mathbf{X}_1} - T) (\lambda I + T)^{-1/2}\right\| \ge 1/2\right\}\right) \\
&\le 4\mathbb{E}\left[\left\|(\lambda I + T)^{-1/2} (T_{\mathbf{X}_1} - T) (\lambda I + T)^{-1/2}\right\|^2\right] \\
&\le 4c_1 \frac{m}{N} \mathcal{N}^2(\lambda).
\end{aligned}$$

Then we obtain the desired result and complete the proof. ■

The following lemma provides an estimation of $\mathcal{N}(\lambda)$ under the polynomial decaying condition of the eigenvalues.

Lemma 15 *Suppose that $\{\mu_k\}_{k\ge 1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \le 1$, then there holds*

$$\mathcal{N}(\lambda) \lesssim \lambda^{-p}, \quad \forall 0 < \lambda \le 1. \tag{5.9}$$

The estimation in Lemma 15 can be found in Guo et al. (2017); Lin et al. (2017); Guo and Shi (2019).

We have established preliminary estimations for Theorem 3 and Theorem 8. We are in the position to prove these two theorems. To this end, we also need to introduce the notations $o_{\mathbb{P}}(\cdot)$ and $\mathcal{O}_{\mathbb{P}}(\cdot)$. For a sequence of random variables $\{\xi_k\}_{k=1}^{\infty}$, we write $\xi_k \le o_{\mathbb{P}}(1)$ if

$$\lim_{k\to\infty} \mathbb{P}\left(|\xi_k| \ge d\right) = 0, \forall d > 0.$$

And we write $\xi_k \le \mathcal{O}_{\mathbb{P}}(1)$ if

$$\lim_{D\to\infty} \sup_{k\ge 1} \mathbb{P}\left(|\xi_k| \ge D\right) = 0.$$

In addition, suppose that there is a positive sequence $\{a_k\}_{k=1}^{\infty}$. Then we write $\xi_k \le o_{\mathbb{P}}(a_k)$ if $\xi_k/a_k \le o_{\mathbb{P}}(1)$, and $\xi_k \le \mathcal{O}_{\mathbb{P}}(a_k)$ if $\xi_k/a_k \le \mathcal{O}_{\mathbb{P}}(1)$.

Proof of Theorem 3. Combining the decomposition (2.8) and (5.1) in Lemma 11 yields

$$\begin{aligned}
\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) &\le 2\mathscr{S}(S,\lambda) + 2\mathscr{A}(\lambda) \\
&\le 2\mathscr{S}(S,\lambda) + 2\lambda^{2\theta} \|\gamma_0\|_{\mathcal{L}^2}^2.
\end{aligned} \tag{5.10}$$

We first decompose $\mathscr{S}(S,\lambda)$ as

$$\mathscr{S}(S,\lambda) = \mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}} + \mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}^c}. \tag{5.11}$$

For the term $\mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}}$, following from (5.8) in Lemma 14, we have

$$\mathbb{E}\left[\mathbb{I}_{\mathcal{U}}\right] = \mathbb{P}(\mathcal{U}) \leq \sum_{j=1}^{m} \mathbb{P}(\mathcal{U}_j) = m\mathbb{P}(\mathcal{U}_1) \leq 4c_1 \frac{m^2}{N}\mathcal{N}^2(\lambda).$$

Then using Markov's inequality, we can write

$$\mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}} \leq \mathcal{O}_{\mathbb{P}}\left(\frac{m^2}{N}\mathcal{N}^2(\lambda)\right)\mathscr{S}(S,\lambda). \tag{5.12}$$

For the term $\mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}^c}$, combining (5.3) in Lemma 12 with (5.5) and (5.6) in Lemma 13 yields

$$\mathbb{E}\left[\mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}^c}\right] \leq 8\frac{\mathcal{N}(\lambda)}{N}\left(c_1\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2\right) + 4c_1\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2.$$

Then using Markov's inequality, we can write

$$\mathscr{S}(S,\lambda)\mathbb{I}_{\mathcal{U}^c} \leq \mathcal{O}_{\mathbb{P}}\left(\frac{\mathcal{N}(\lambda)}{N} + \frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\right). \tag{5.13}$$

Therefore, combining (5.11), (5.12) and (5.13), we have

$$\left[1 - \mathcal{O}_{\mathbb{P}}\left(\frac{m^2}{N}\mathcal{N}^2(\lambda)\right)\right]\mathscr{S}(S,\lambda) \leq \mathcal{O}_{\mathbb{P}}\left(\frac{\mathcal{N}(\lambda)}{N} + \frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\right).$$

Then applying the estimation of $\mathcal{N}(\lambda)$ (5.9) in Lemma 15, taking $\lambda \leq 1$, we can write

$$\left[1 - \mathcal{O}_{\mathbb{P}}\left(\frac{m^2}{N}\lambda^{-2p}\right)\right]\mathscr{S}(S,\lambda) \leq \mathcal{O}_{\mathbb{P}}\left(\frac{\lambda^{-p}}{N} + \frac{m}{N}\lambda^{2\theta-p}\right). \tag{5.14}$$

Take m and λ satisfying $m^2\lambda^{-2p} \leq o(N)$ and $\lambda \leq 1$, then (5.14) implies that

$$\left[1 - o_{\mathbb{P}}(1)\right]\mathscr{S}(S,\lambda) \leq \mathcal{O}_{\mathbb{P}}\left(\frac{\lambda^{-p}}{N} + \frac{m}{N}\lambda^{2\theta-p}\right), \text{ as } \mathcal{O}_{\mathbb{P}}\left(\frac{m^2}{N}\lambda^{-2p}\right) \leq o_{\mathbb{P}}(1).$$

Thus, we can write

$$\mathscr{S}(S,\lambda) \leq \mathcal{O}_{\mathbb{P}}\left(\frac{\lambda^{-p}}{N} + \frac{m}{N}\lambda^{2\theta-p}\right).$$

Combining the above estimation with (5.10) yields

$$\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \leq \mathcal{O}_{\mathbb{P}}\left(\lambda^{2\theta} + \frac{\lambda^{-p}}{N} + \frac{m}{N}\lambda^{2\theta-p}\right) \tag{5.15}$$

provided that

$$m^2\lambda^{-2p} \leq o(N) \text{ and } \lambda \leq 1.$$

When $p/2 < \theta \leq 1/2$, take $m \leq o\left(N^{\frac{2\theta-p}{4\theta+2p}}\right)$ and $\lambda = N^{-\frac{1}{2\theta+p}}$, then there hold $m^2\lambda^{-2p} \leq o(N)$ and $\lambda \leq 1$. Therefore, following from (5.15), we can write

$$\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \leq \mathcal{O}_{\mathbb{P}}\left(N^{-\frac{2\theta}{2\theta+p}}\right),$$

or equivalently,

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma N^{-\frac{2\theta}{2\theta+p}}\right\} = 0.$$

This completes the proof of (3.5).

When $0 < \theta \leq p/2$, take m and λ satisfying $m^2\lambda^{-2p} \leq o(N)$ and $\lambda \leq 1$, then following from (5.15), one can calculate

$$\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \leq \mathcal{O}_{\mathbb{P}}\left(\lambda^{2\theta}\right),$$

or equivalently,

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma\lambda^{2\theta}\right\} = 0,$$

which further implies (3.6) and (3.7). The proof of Theorem 3 is then completed. ■

Now we turn to prove Theorem 8.

Proof of Theorem 8. Recalling (5.4), under the noiseless condition, we can write

$$\hat{f}_{S_1,\lambda} = (\lambda I + T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^{n} L_K^{1/2} X_{1,i}\langle X_{1,i}, \beta_0\rangle_{\mathcal{L}^2}.$$

Then we can give an improved estimation of the left hand side of (5.5) as

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\hat{f}_{S_1,\lambda} - f_\lambda\right)\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c}\right] \\
&= \mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I + T_{\mathbf{X}_1})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} L^{1/2}X_{1,i}\langle X_{1,i}, \beta_0 - L_K^{1/2}f_\lambda\rangle_{\mathcal{L}^2} - \lambda f_\lambda\right)\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c}\right] \\
&\overset{(*)}{\leq} 4c_1\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2, \qquad (5.16)
\end{aligned}$$

where inequality $(*)$ follows from (5.7).

Employing (5.16) and following the same arguments in the proof of Theorem 3, we have

$$.\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \leq \mathcal{O}_{\mathbb{P}}\left(\lambda^{2\theta} + \frac{m}{N}\lambda^{2\theta-p}\right), \qquad (5.17)$$

provided that

$$m^2\lambda^{-2p} \leq o(N) \text{ and } \lambda \leq 1.$$

For any $0 < \eta \leq 1/2$, take $m \leq o(N^{\eta})$ and $\lambda = N^{-\frac{1-2\eta}{2p}}$, then there hold $m^2\lambda^{-2p} \leq o(N)$ and $\lambda \leq 1$. Therefore, following from (5.17), one can calculate

$$\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \leq \mathcal{O}_{\mathbb{P}}\left(\lambda^{2\theta}\right) \leq \mathcal{O}_{\mathbb{P}}\left(N^{-\frac{\theta(1-2\eta)}{p}}\right),$$

or equivalently,

$$\lim_{\Gamma\to\infty} \sup_{N\to\infty} \lim \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0) \geq \Gamma N^{-\frac{\theta(1-2\eta)}{p}}\right\} = 0.$$

We have obtained (3.20). The proof of Theorem 8 is then finished. ■

We next aim to prove Theorem 4 and Corollary 5. We also need several lemmas before proving them.

When Assumption 3 is enhanced to Assumption 4, we can estimate the probability of event $\mathcal{U}_1$ better than Lemma 14.

Lemma 16 *Suppose that Assumption 4 is satisfied with some integer $\ell \geq 2$. Then there holds*

$$\mathbb{P}(\mathcal{U}_1) \leq c(\ell)2^{4\ell}\rho^{4\ell}\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\ell}, \tag{5.18}$$

where $c(\ell)$ is a constant only depends on ℓ and $\mathcal{N}(\lambda)$ is given by (3.4)*.*

Lemma 16 can be proved by employing Markov's inequality combined with the following lemma.

Lemma 17 *Suppose that Assumption 4 is satisfied with some integer $\ell \geq 2$. Then*

$$\mathbb{E}\left[\left\|(\lambda I + T)^{-1/2}\left(T_{\boldsymbol{X}_1} - T\right)(\lambda I + T)^{-1/2}\right\|_{HS}^{2\ell}\right] \leq c(\ell)2^{2\ell}\rho^{4\ell}\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\ell} \tag{5.19}$$

and

$$\mathbb{E}\left[\left\|(\lambda I + T)^{-1/2}\left(T_{\boldsymbol{X}_1} - T\right)\right\|_{HS}^{2\ell}\right] \leq c(\ell)2^{2\ell}\rho^{4\ell} trace^{\ell}(T)\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\ell}, \tag{5.20}$$

where $trace(T) = \sum_{j=1}^{\infty}\mu_j$ denotes the trace of operator T, $\mathcal{N}(\lambda)$ is the effective dimension given by (3.4)*, and $c(\ell)$ is a constant only depends on ℓ.*

Proof We first prove inequality (5.19). Recalling (5.4), for brevity of notations, we define

$$Q_i := (\lambda I + T)^{-1/2}\left(L_K^{1/2}X_{1,i} \otimes L_K^{1/2}X_{1,i} - T\right)(\lambda I + T)^{-1/2}, \quad i = 1, 2, \cdots, n.$$

Then we can write

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I + T)^{-1/2}\left(T_{\mathbf{X}_1} - T\right)(\lambda I + T)^{-1/2}\right\|_{HS}^{2\ell}\right] \\
&= \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{i=1}^{n} Q_i, \frac{1}{n}\sum_{j=1}^{n} Q_j \right\rangle_{HS}^{\ell}\right] \\
&= \frac{1}{n^{2\ell}}\sum_{i_1=1}^{n}\cdots\sum_{i_\ell=1}^{n}\sum_{j_1=1}^{n}\cdots\sum_{j_\ell=1}^{n}\mathbb{E}\left[\langle Q_{i_1}, Q_{j_1}\rangle_{HS}\cdots\langle Q_{i_\ell}, Q_{j_\ell}\rangle_{HS}\right].
\end{aligned}$$

When the indices in group $\{i_1, ..., i_\ell, j_1, ..., j_\ell\}$ are all distinct, then following from the independence, there holds $\mathbb{E}\left[\langle Q_{i_1}, Q_{j_1}\rangle_{HS} \cdots \langle Q_{i_\ell}, Q_{j_\ell}\rangle_{HS}\right] = 0$. We denote the set of all index-distinct groups by $\Omega(n,\ell)$. Let $\Theta(n,\ell) = \{1, ..., n\}^{2\ell} \backslash \Omega(n,\ell)$. Using these notations, we can write

$$
\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I + T)^{-1/2}(T_{\mathbf{X}_1} - T)(\lambda I + T)^{-1/2}\right\|_{HS}^{2\ell}\right] \\
=&\frac{1}{n^{2\ell}} \sum_{\{i_1,...,i_\ell,j_1,...,j_\ell\} \in \Theta(n,\ell)} \mathbb{E}\left[\langle Q_{i_1}, Q_{j_1}\rangle_{HS} ... \langle Q_{i_\ell}, Q_{j_\ell}\rangle_{HS}\right].
\end{aligned} \tag{5.21}
$$

We estimate the cardinality of $\Theta(n,k)$ as

$$
\begin{aligned}
|\Theta(n,\ell)| &= |\Theta_\ell(n,\ell)| + ... + |\Theta_1(n,\ell)| \\
&\le (2\ell)! \left[\binom{n}{\ell} + \binom{n}{\ell-1}(\ell-1)^2 + \cdots + \binom{n}{1}\right] \le (2\ell)! \ell^{2\ell+1} n^\ell := c(\ell) n^\ell,
\end{aligned} \tag{5.22}
$$

where $c(\ell) := (2\ell)!\ell^{2\ell+1}$. Let $\Theta_i(n,\ell)$ denote a subset of $\Theta(n,\ell)$ consisting of all groups with exactly i different indices. Then $\Theta(n,\ell) = \cup_{i=1}^{\ell} \Theta_i(n,\ell)$ and $|\Theta_i(n,\ell)| \le (2\ell)!\binom{n}{i} i^{2(\ell-i)} \le (2\ell)!\ell^{2\ell} n^\ell$.

For any $\{i_1, \cdots, i_\ell, j_1, \cdots, j_\ell\} \in \Theta(n,\ell)$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\langle Q_{i_1}, Q_{j_1}\rangle_{HS} \cdots \langle Q_{i_\ell}, Q_{j_\ell}\rangle_{HS}\right] \\
&\le \mathbb{E}\left[\|Q_{i_1}\|_{HS} \|Q_{j_1}\|_{HS} \cdots \|Q_{i_\ell}\|_{HS} \|Q_{j_\ell}\|_{HS}\right] \\
&\overset{(\dagger)}{\le} \left[\mathbb{E}\|Q_{i_1}\|_{HS}^{2\ell}\right]^{\frac{1}{2\ell}} \left[\mathbb{E}\|Q_{j_1}\|_{HS}^{2\ell}\right]^{\frac{1}{2\ell}} \cdots \left[\mathbb{E}\|Q_{i_\ell}\|_{HS}^{2\ell}\right]^{\frac{1}{2\ell}} \left[\mathbb{E}\|Q_{j_\ell}\|_{HS}^{2\ell}\right]^{\frac{1}{2\ell}},
\end{aligned} \tag{5.23}
$$

where inequality $(\dagger)$ uses Hölder inequality. Then we further bound $\mathbb{E}\left[\|Q_i\|_{HS}^{2\ell}\right]$ for any $1 \le i \le n$, which is given

$$
\mathbb{E}\left[\|Q_i\|_{HS}^{2\ell}\right] \tag{5.24}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left\|(\lambda I + T)^{-1/2}\left(L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} - T\right)(\lambda I + T)^{-1/2}\right\|_{HS}^{2\ell}\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \frac{1}{\lambda + \mu_j}\frac{1}{\lambda + \mu_k} \langle (X_{1,i} \otimes X_{1,i} - T)\phi_j, \phi_k\rangle_{\mathcal{L}^2}^2\right)^\ell\right] \\
&= \sum_{j_1=1}^{\infty} \cdots \sum_{j_\ell=1}^{\infty}\sum_{k_1=1}^{\infty} \cdots \sum_{k_\ell=1}^{\infty} \mathbb{E}\Bigg[\frac{1}{\lambda + \mu_{j_1}}\frac{1}{\lambda + \mu_{k_1}} \left\langle \left(L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} - T\right)\phi_{j_1}, \phi_{k_1}\right\rangle_{\mathcal{L}^2}^2 \\
&\quad \times \cdots \times \frac{1}{\lambda + \mu_{j_\ell}}\frac{1}{\lambda + \mu_{k_\ell}} \left\langle \left(L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} - T\right)\phi_{j_\ell}, \phi_{k_\ell}\right\rangle_{\mathcal{L}^2}^2\Bigg] \\
&\overset{(*)}{\le} \sum_{j_1=1}^{\infty}\sum_{k_1=1}^{\infty} \frac{1}{\lambda + \mu_{j_1}}\frac{1}{\lambda + \mu_{k_1}} \left[\mathbb{E}\left\langle \left(L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} - T\right)\phi_{j_1}, \phi_{k_1}\right\rangle^{2\ell}\right]^{\frac{1}{\ell}} \times \cdots \times \\
&\quad \sum_{j_\ell=1}^{\infty}\sum_{k_\ell=1}^{\infty} \frac{1}{\lambda + \mu_{j_\ell}}\frac{1}{\lambda + \mu_{k_\ell}} \left[\mathbb{E}\left\langle \left(L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} - T\right)\phi_{j_\ell}, \phi_{l_\ell}\right\rangle_{\mathcal{L}^2}^{2\ell}\right]^{\frac{1}{\ell}},
\end{aligned} \tag{5.25}
$$

where inequality $(*)$ also uses Hölder inequality. It remains to estimate

$$\mathbb{E}\left[\left\langle\left(L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}-T\right)\phi_j,\phi_k\right\rangle_{\mathcal{L}^2}^{2\ell}\right],\quad \forall 1\leq i\leq n \text{ and } \forall 1\leq j,k<\infty.$$

When $j\neq k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\left(L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}-T\right)\phi_j,\phi_k\right\rangle_{\mathcal{L}^2}^{2\ell}\right]\\
&=\mathbb{E}\left[\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^{2\ell}\langle L_K^{1/2}X_{1,i},\phi_k\rangle_{\mathcal{L}^2}^{2\ell}\right]\\
&\overset{(i)}{\leq}\left[\mathbb{E}\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^{4\ell}\right]^{\frac{1}{2}}\left[\mathbb{E}\left\langle L_K^{1/2}X_{1,i},\phi_k\right\rangle_{\mathcal{L}^2}^{4\ell}\right]^{\frac{1}{2}}\overset{(ii)}{\leq}\rho^{4\ell}\mu_j^\ell\mu_k^\ell,
\end{aligned}$$

where inequality (i) is from Cauchy-Schwarz inequality and inequality (ii) uses Assumption 4.

When $j=k$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle\left(L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}-T\right)\phi_j,\phi_j\right\rangle_{\mathcal{L}^2}^{2\ell}\right]\\
&=\mathbb{E}\left[\left(\left\langle L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}\phi_j,\phi_j\right\rangle_{\mathcal{L}^2}-\mu_j\right)^{2\ell}\right]\\
&=2^{2\ell}\mathbb{E}\left[\left(\frac{1}{2}\left\langle L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}\phi_j,\phi_j\right\rangle_{\mathcal{L}^2}-\frac{1}{2}\mu_j\right)^{2\ell}\right]\\
&\overset{(i)}{\leq}2^{2\ell-1}\left(\mathbb{E}\left[\left\langle L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}\phi_j,\phi_j\right\rangle_{\mathcal{L}^2}^{2\ell}\right]+\mu_j^{2\ell}\right)\\
&=2^{2\ell-1}\left(\mathbb{E}\left[\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^{4\ell}\right]+\mu_j^{2\ell}\right)\overset{(ii)}{\leq}2^{2\ell}\rho^{4\ell}\mu_j^{2\ell},
\end{aligned}$$

where inequality (i) is due to Jensen's inequality and inequality (ii) follows from Assumption 4 and the fact that $\rho\geq 1$.

Combining the above estimations, for any $1\leq i\leq n$ and $1\leq j,k<\infty$, there holds

$$\mathbb{E}\left[\left\langle\left(L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}-T\right)\phi_j,\phi_k\right\rangle^{2\ell}\right]\leq 2^{2\ell}\rho^{4\ell}\mu_j^\ell\mu_k^\ell. \tag{5.26}$$

Recall that $n=N/m$. Combining (5.21), (5.22), (5.23), (5.24) and (5.26) yields

$$\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\left(T_{\mathbf{X}_1}-T\right)(\lambda I+T)^{-1/2}\right\|_{HS}^{2\ell}\right]\leq c(\ell)2^{2\ell}\rho^{4\ell}\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^\ell.$$

This completes the proof of (5.19).

Analogously, we can demonstrate the second inequality (5.20) through

$$\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\left(L_K^{1/2}X_{1,i}\otimes L_K^{1/2}X_{1,i}-T\right)\right\|_{HS}^{2\ell}\right]\leq 2^{2\ell}\rho^{4\ell}trace^\ell(T)\mathcal{N}^\ell(\lambda)$$

and

$$\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\left(T_{\mathbf{X}_1}-T\right)\right\|_{HS}^{2\ell}\right] \leq c(\ell) 2^{2\ell}\rho^{4\ell} trace^{\ell}(T)\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\ell}.$$

The proof of Lemma 17 is then finished. ■

The following lemma plays a key role in estimating the upper bound of $\mathscr{S}(S,\lambda)$ under Assumption 4.

Lemma 18 *Suppose that Assumption 1 is satisfied with* $0<\theta\leq 1/2$ *and* $\gamma_0\in\mathcal{L}^2(\mathcal{T})$. *Under Assumption 2 and Assumption 4, taking* $\lambda\leq 1$ *yields*

$$\begin{aligned}&\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\frac{1}{|S_1|}\left(\sum_{X\in\boldsymbol{X}_1} L_K^{1/2}X\langle X,\beta_0-L_K^{1/2}f_\lambda\rangle_{\mathcal{L}^2}-\lambda f_\lambda\right)\right\|_{\mathcal{L}^2}^4\right]\\ &\leq c_6^2\frac{m^2}{N^2}(1+\lambda^{4\theta}\mathcal{N}^2(\lambda)),\end{aligned} \tag{5.27}$$

where c_6 *is a universal constant and* $\mathcal{N}(\lambda)$ *is given by* (3.4).

Proof Recalling (5.4), for simplicity of notations, we define

$$\alpha_i := L_K^{1/2}X_{1,i}\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}-\lambda f_\lambda,\quad i=1,2,\cdots,n.$$

We begin with the proof of the first inequality (5.27). Note that $\left\{(\lambda I+T)^{-1/2}\alpha_i\right\}_{i=1}^n$ are independent operator-valued zero-mean random elements. Then we can write

$$\begin{aligned}&\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^4\right]\\ &=\frac{1}{n^4}\sum_{i_1=1}^\infty\sum_{i_2=1}^\infty\sum_{j_1=1}^\infty\sum_{j_2=1}^\infty\mathbb{E}\Big[\left\langle(\lambda I+T)^{-1/2}\alpha_{i_1},(\lambda I+T)^{-1/2}\alpha_{i_2}\right\rangle_{\mathcal{L}^2}\\ &\quad\times\left\langle(\lambda I+T)^{-1/2}\alpha_{j_1},(\lambda I+T)^{-1/2}\alpha_{j_2}\right\rangle_{\mathcal{L}^2}\Big]\\ &=\frac{1}{n^4}\sum_{\{i_1,i_2,j_1,j_2\}\in\Theta(n,2)}\mathbb{E}\Big[\left\langle(\lambda I+T)^{-1/2}\alpha_{i_1},(\lambda I+T)^{-1/2}\alpha_{i_2}\right\rangle_{\mathcal{L}^2}\\ &\quad\times\left\langle(\lambda I+T)^{-1/2}\alpha_{j_1},(\lambda I+T)^{-1/2}\alpha_{j_2}\right\rangle_{\mathcal{L}^2}\Big],\end{aligned} \tag{5.28}$$

where $\Theta(n,2)=\{1,...,n\}^4\backslash\Omega(n,2)$ and $\Omega(n,2)$ denotes the set of all index-distinct group $\{i_1,i_2,j_1,j_2\}$. Then

$$|\Theta(n,2)|\leq 4!\left[\binom{n}{2}+\binom{n}{1}\right]\leq 24n^2,\quad\forall n\geq 1. \tag{5.29}$$

And for any $\{i_1, i_2, j_1, j_2\} \in \Theta(n, 2)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\langle(\lambda I+T)^{-1/2}\alpha_{i_1},(\lambda I+T)^{-1/2}\alpha_{i_2}\rangle_{\mathcal{L}^2}\langle(\lambda I+T)^{-1/2}\alpha_{j_1},(\lambda I+T)^{-1/2}\alpha_{j_2}\right\rangle_{\mathcal{L}^2}\right] \\
&\leq \mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\alpha_{i_1}\right\|_{\mathcal{L}^2}\left\|(\lambda I+T)^{-1/2}\alpha_{i_2}\right\|_{\mathcal{L}^2}\left\|(\lambda I+T)^{-1/2}\alpha_{j_1}\right\|_{\mathcal{L}^2}\left\|(\lambda I+T)^{-1/2}\alpha_{j_2}\right\|_{\mathcal{L}^2}\right] \\
&\overset{(*)}{\leq} \left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}\alpha_{i_1}\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{4}}\left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}\alpha_{i_2}\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{4}} \\
&\quad\times\left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}\alpha_{j_1}\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{4}}\left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}\alpha_{j_2}\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{4}},
\end{aligned} \tag{5.30}$$

where inequality $(*)$ uses Hölder inequality.

It remains to estimate $\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\alpha_i\right\|_{\mathcal{L}^2}^4\right], \forall 1 \leq i \leq n$. For brevity of notations, we define

$$\widetilde{\alpha}_i := L_K^{1/2}X_{1,i}\left\langle X_{1,i}, \beta_0 - L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}, \quad i = 1, 2, \cdots, n.$$

Then we see that $\alpha_i = \widetilde{\alpha}_i - \lambda f_\lambda$ and for any $1 \leq i \leq n$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I+T_0)^{-1/2}\widetilde{\alpha}_i\right\|_{\mathcal{L}^2}^4\right] \\
&= \mathbb{E}\left[\left(\sum_{j=1}^{\infty}\left\langle(\lambda I+T)^{-1/2}\widetilde{\alpha}_i, \phi_j\right\rangle_{\mathcal{L}^2}^2\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j_1=1}^{\infty}\sum_{j_2=1}^{\infty}\frac{1}{\lambda+\mu_{j_1}}\frac{1}{\lambda+\mu_{j_2}}\left\langle\widetilde{\alpha}_i,\phi_{j_1}\right\rangle_{\mathcal{L}^2}^2\left\langle\widetilde{\alpha}_i,\phi_{j_2}\right\rangle_{\mathcal{L}^2}^2\right] \\
&\overset{(\dagger)}{\leq} \sum_{j_1=1}^{\infty}\sum_{j_2=1}^{\infty}\frac{1}{\lambda+\mu_{j_1}}\frac{1}{\lambda+\mu_{j_2}}\left[\mathbb{E}\left\langle\widetilde{\alpha}_i,\phi_{j_1}\right\rangle_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}\left[\mathbb{E}\left\langle\widetilde{\alpha}_i,\phi_{j_2}\right\rangle_{\mathcal{L}^2}^4\right]^{\frac{1}{2}},
\end{aligned} \tag{5.31}$$

where inequality $(\dagger)$ uses Cauchy-Schwartz inequality. We further bound $\mathbb{E}\left[\langle\widetilde{\alpha}_i,\phi_j\rangle_{\mathcal{L}^2}^4\right]$ as

$$\begin{aligned}
&\mathbb{E}[\langle\widetilde{\alpha}_i,\phi_j\rangle_{\mathcal{L}^2}^4] \\
&= \mathbb{E}\left[\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^4\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^4\right] \\
&\overset{(i)}{\leq} \left[\mathbb{E}\left\langle L_K^{1/2}X_i,\phi_j\right\rangle_{\mathcal{L}^2}^8\right]^{\frac{1}{2}}\left[\mathbb{E}\left\langle X_i,\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^8\right]^{\frac{1}{2}} \\
&\overset{(ii)}{\leq} c_2\rho^4\mu_j^2\left[\mathbb{E}\left\langle X_i,\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^2\right]^2 = c_2\rho^4\mu_j^2\mathscr{A}^4(\lambda) \overset{(iii)}{\leq} c_2\rho^4\mu_j^2\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta},
\end{aligned} \tag{5.32}$$

where inequality (i) again uses Cauchy-Schwartz inequality, inequality (ii) is due to Assumption 4 and inequality (iii) is from Lemma 11.

Combining (5.31) and (5.32) yields

$$\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\widetilde{\alpha}_i\right\|_{\mathcal{L}^2}^4\right] \leq c_2\rho^4\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta}\mathcal{N}^2(\lambda), \quad \forall 1 \leq i \leq n.$$

Then for any $1 \le i \le n$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\alpha_i\right\|_{\mathcal{L}^2}^4\right] = \mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}(\tilde{\alpha}_i - \lambda f_\lambda)\right\|_{\mathcal{L}^2}^4\right] \\
&\overset{(i)}{\le} 8\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\widetilde{\alpha}_i\right\|_{\mathcal{L}^2}^4\right] + 8\left\|(\lambda I+T)^{-1/2}\lambda f_\lambda\right\|_{\mathcal{L}^2}^4 \\
&\overset{(ii)}{=} 8\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}\widetilde{\alpha}_i\right\|_{\mathcal{L}^2}^4\right] + 8\left\|(\lambda I+T)^{-1/2}\lambda(\lambda I+T)^{-1}L_K^{1/2}L_C^{1/2}T_*^\theta(\gamma_0)\right\|_{\mathcal{L}^2}^4 \qquad (5.33) \\
&\overset{(iii)}{\le} 8c_2\rho^4\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta}\mathcal{N}^2(\lambda) + 8\left\|(\lambda I+T)^{-1}\lambda\right\|^4\left\|(\lambda I+T)^{-1/2}L_K^{1/2}L_C^{1/2}\right\|^4\left\|T_*^\theta\right\|^4\|\gamma_0\|_{\mathcal{L}^2}^4 \\
&\le 8c_2\rho^4\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta}\mathcal{N}^2(\lambda) + 8\left\|T_*^\theta\right\|^4\|\gamma_0\|_{\mathcal{L}^2}^4 = 8c_2\rho^4\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta}\mathcal{N}^2(\lambda) + 8\mu_1^{4\theta}\|\gamma_0\|_{\mathcal{L}^2}^4,
\end{aligned}$$

where inequality (i) uses the triangular inequality, inequality (ii) follows from Assumption 1 and the expression of f_λ and inequality (iii) applies the above estimation.

Recall that $n = N/m$ and take $\lambda \le 1$. Combining with (5.28), (5.29), (5.30) and (5.33), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I+T_0)^{-1/2}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^4\right] \\
&\le \frac{192m^2}{N^2}\left(c_2\rho^4||\gamma_0||_{\mathcal{L}^2}^4\lambda^{4\theta}\mathcal{N}^2(\lambda) + \mu_1^{4\theta}||\gamma_0||_{\mathcal{L}^2}^4\right) \le c_6^2\frac{m^2}{N^2}(1+\lambda^{4\theta}\mathcal{N}^2(\lambda)),
\end{aligned}$$

where $c_6^2 := 192\left(c_2\rho^4||\gamma_0||_{\mathcal{L}^2}^4 + \max\{\mu_1^2, 1\}||\gamma_0||_{\mathcal{L}^2}^4\right)$. This completes the proof of Lemma 18. ■

We propose the following lemma to decompose $\mathbb{E}[\mathscr{S}(S,\lambda)]$.

Lemma 19 *For any $m \ge 1$, there holds*

$$\mathbb{E}\left[\mathscr{S}(S,\lambda)\right] \le \frac{1}{m}\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\hat{f}_{S_1,\lambda} - f_\lambda\right)\right\|_{\mathcal{L}^2}^2\right] + \left\|L_C^{1/2}L_K^{1/2}\mathbb{E}\left[(\hat{f}_{S_1,\lambda} - f_\lambda)\right]\right\|_{\mathcal{L}^2}^2. \qquad (5.34)$$

Proof When $m \ge 2$, as

$$\mathscr{S}(S,\lambda) = \left\|L_C^{1/2}L_K^{1/2}\overline{f}_{S,\lambda} - L_C^{1/2}L_K^{1/2}f_\lambda\right\|_{\mathcal{L}^2}^2 = \left\|\frac{1}{m}\sum_{i=1}^m L_C^{1/2}L_K^{1/2}\hat{f}_{S_i,\lambda} - L_C^{1/2}L_K^{1/2}f_\lambda\right\|_{\mathcal{L}^2}^2,$$

we can write

$$\begin{aligned}
\mathbb{E}[\mathscr{S}(S,\lambda)] &= \mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\frac{1}{m}\sum_{i=1}^m\hat{f}_{S_i,\lambda} - f_\lambda\right)\right\|_{\mathcal{L}^2}^2\right] \\
&\overset{(i)}{=} \frac{1}{m^2}\sum_{i=1}^m\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\hat{f}_{S_i,\lambda} - f_\lambda\right)\right\|_{\mathcal{L}^2}^2\right]
\end{aligned}$$

$$+ \frac{1}{m^2} \sum_{i \neq j} \mathbb{E} \left[\left\langle L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_i,\lambda} - f_\lambda \right), L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_j,\lambda} - f_\lambda \right) \right\rangle_{\mathcal{L}^2} \right]$$

$$\overset{(ii)}{\leq} \frac{1}{m} \mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right] + \left\| L_C^{1/2} L_K^{1/2} \mathbb{E} \left[\left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right] \right\|_{\mathcal{L}^2}^2 .$$

where equality (i) follows from the binomial expansion and inequality (ii) is from

$$\mathbb{E} \left[\langle L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right), L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_2,\lambda} - f_\lambda \right) \rangle_{\mathcal{L}^2} \right] = \left\| L_C^{1/2} L_K^{1/2} \mathbb{E} \left[\left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right] \right\|_{\mathcal{L}^2}^2 .$$

When $m = 1$, (5.34) is obvious.

Thus, we have completed the proof of Lemma 19. ■

Now we are in the position to prove Theorem 4.

Proof of Theorem 4. Combining (2.8), (5.1) and (5.34) yields

$$\begin{aligned} &\mathbb{E} \left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \right) \right] \leq 2\mathbb{E} \left[\mathscr{S}(S, \lambda) \right] + 2\mathscr{A}(\lambda) \qquad (5.35) \\ &\leq \frac{2}{m} \mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} (\hat{f}_{S_1,\lambda} - f_\lambda) \right\|_{\mathcal{L}^2}^2 \right] + 2 \left\| L_C^{1/2} L_K^{1/2} \mathbb{E} \left[(\hat{f}_{S_1,\lambda} - f_\lambda) \right] \right\|_{\mathcal{L}^2}^2 + 2\mathscr{A}(\lambda) \\ &\leq \frac{2}{m} \mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} (\hat{f}_{S_1,\lambda} - f_\lambda) \right\|_{\mathcal{L}^2}^2 \right] + 2 \left\| L_C^{1/2} L_K^{1/2} \mathbb{E} \left[(\hat{f}_{S_1,\lambda} - f_\lambda) \right] \right\|_{\mathcal{L}^2}^2 + 2\lambda^{2\theta} \|\gamma_0\|_{\mathcal{L}^2}^2 . \end{aligned}$$

In the following part of the proof, we aim to bound the terms $\mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} (\hat{f}_{S_1,\lambda} - f_\lambda) \right\|_{\mathcal{L}^2}^2 \right]$ and $\left\| L_C^{1/2} L_K^{1/2} \mathbb{E} \left[(\hat{f}_{S_1,\lambda} - f_\lambda) \right] \right\|_{\mathcal{L}^2}^2$, respectively. Recalling (5.4) and $Y_{1,i} = \langle X_{1,i}, \beta_0 \rangle_{\mathcal{L}^2} + \epsilon_{1,i}, \forall 1 \leq i \leq n$, for simplicity of notations, let

$$\alpha_i := L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda, \quad i = 1, 2, \cdots, n.$$

Then

$$\hat{f}_{S_1,\lambda} - f_\lambda = (\lambda I + T_{\mathbf{X}_1})^{-1} \frac{1}{n} \sum_{i=1}^{n} \left(\alpha_i + L_K^{1/2} X_{1,i} \epsilon_{1,i} \right).$$

Using this expression, we can bound $\mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right]$ as

$$\begin{aligned} &\mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \right] \\ &= \mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1^c} \right] + \mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} \left(\hat{f}_{S_1,\lambda} - f_\lambda \right) \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \\ &\overset{(i)}{\leq} 8 \frac{m}{N} \mathcal{N}(\lambda) \left(c_2 \lambda^{2\theta} \|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2 \right) + 2\mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} \frac{1}{n} \sum_{i=1}^{n} \alpha_i \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \\ &\quad + 2\mathbb{E} \left[\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} \frac{1}{n} \sum_{i=1}^{n} L_K^{1/2} X_{1,i} \epsilon_{1,i} \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \qquad (5.36) \end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} 8\frac{m}{N}\mathcal{N}(\lambda)\left(c_2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2+\sigma^2\right)+2\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\quad+\frac{2\sigma^2}{n^2}\sum_{i=1}^n\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}L_K^{1/2}X_{1,i}\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right].
\end{aligned}
$$

Here inequality (i) follows from (5.5) in Lemma 13 and the triangular inequality. Inequality (ii) is due to Assumption 2.

We next bound $\left\|L_C^{1/2}L_K^{1/2}\mathbb{E}\left[\hat{f}_{S_1,\lambda}-f_\lambda\right]\right\|_{\mathcal{L}^2}^2$ as

$$
\begin{aligned}
&\left\|L_C^{1/2}L_K^{1/2}\mathbb{E}\left[\hat{f}_{S_1,\lambda}-f_\lambda\right]\right\|_{\mathcal{L}^2}^2\\
&\overset{(i)}{=}\left\|L_C^{1/2}L_K^{1/2}\mathbb{E}\left[(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right]\right\|_{\mathcal{L}^2}^2\\
&\leq\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1^c}\right]\\
&\quad+\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\overset{(ii)}{\leq} 4c_2\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2+\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right].
\end{aligned}
\tag{5.37}
$$

Here equality (i) is from Assumption 2. Inequality (ii) follows from Jensen's inequality and (5.6) in Lemma 13.

The key point in the rest of the proof is to estimate $\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]$ and $\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}L_K^{1/2}X_{1,i}\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right], \forall 1\leq i\leq n$. For the first term, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\leq\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T)^{-1/2}\right\|^2\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\overset{(i)}{\leq}\mathbb{E}\left[\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\overset{(ii)}{\leq}\left[\mathbb{E}\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^4\mathbb{I}_{\mathcal{U}_1}\right]^{\frac{1}{2}}\left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}\frac{1}{n}\sum_{i=1}^n\alpha_i\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}
\end{aligned}
\tag{5.38}
$$

$$\overset{(iii)}{\leq} \left[\mathbb{E}\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^8\right]^{\frac{1}{4}} \mathbb{P}^{\frac{1}{4}}(\mathcal{U}_1)\left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}\frac{1}{n}\sum_{i=1}^n \alpha_i\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}.$$

Here inequality (i) follows from the fact that

$$\begin{aligned}\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T)^{-1/2}\right\|^2 &= \left\|(\lambda I+T)^{-1/2}L_K^{1/2}L_C L_K^{1/2}(\lambda I+T)^{-1/2}\right\| \\ &= \left\|(\lambda I+T)^{-1/2}T(\lambda I+T)^{-1/2}\right\| \leq 1.\end{aligned}$$

Inequalities (ii) and (iii) are from Cauchy-Schwartz inequality.

Analogously, for the second term, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}L_K^{1/2}X_{1,i}\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1}\right] \\ &\leq \left[\mathbb{E}\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^8\right]^{\frac{1}{4}} \mathbb{P}^{\frac{1}{4}}(\mathcal{U}_1)\left[\mathbb{E}\left\|(\lambda I+T)^{-1/2}L_K^{1/2}X_{1,i}\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}.\end{aligned} \tag{5.39}$$

While we can write

$$\begin{aligned}&\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}L_K^{1/2}X_{1,i}\right\|_{\mathcal{L}^2}^4\right] \\ &= \mathbb{E}\left[\left(\sum_{j=1}^\infty \frac{1}{\lambda+\mu_j}\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^2\right)^2\right] \\ &= \sum_{j=1}^\infty\sum_{k=1}^\infty \frac{1}{\lambda+\mu_j}\frac{1}{\lambda+\mu_k}\mathbb{E}\left[\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^2\left\langle L_K^{1/2}X_{1,i},\phi_k\right\rangle_{\mathcal{L}^2}^2\right] \\ &\overset{(i)}{\leq} \sum_{j=1}^\infty\sum_{k=1}^\infty \frac{1}{\lambda+\mu_j}\frac{1}{\lambda+\mu_k}\left[\mathbb{E}\left\langle L_K^{1/2}X_{1,i},\phi_j\right\rangle_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}\left[\mathbb{E}\left\langle L_K^{1/2}X_{1,i},\phi_k\right\rangle_{\mathcal{L}^2}^4\right]^{\frac{1}{2}} \\ &\overset{(ii)}{\leq} \rho^4\sum_{j=1}^\infty\sum_{k=1}^\infty \frac{\mu_j}{\lambda+\mu_j}\frac{\mu_k}{\lambda+\mu_k} = \rho^4\mathcal{N}^2(\lambda).\end{aligned} \tag{5.40}$$

Here $\{\phi_k\}_{k=1}^\infty$ is given by the singular value decomposition of T in (2.9). Inequality (i) is from Cauchy-Schwartz inequality. Inequality (ii) is due to the decomposition of $L_K^{1/2}X$ (3.8) and Assumption 4.

For the term $\mathbb{E}\left[\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^8\right]$, first applying the second-order decomposition, which was introduced in Guo et al. (2017); Lin et al. (2017); Guo and Shi (2019), to $(\lambda I+T_{\mathbf{X}_1})^{-1}$ yields that

$$\begin{aligned}(\lambda I+T_{\mathbf{X}_1})^{-1} &= (\lambda I+T)^{-1}+(\lambda I+T_{\mathbf{X}_1})^{-1}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1} \\ &= (\lambda I+T)^{-1}+(\lambda I+T)^{-1}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1} \\ &\quad +(\lambda I+T)^{-1}(T-T_{\mathbf{X}_1})(\lambda I+T_{\mathbf{X}_1})^{-1}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1}.\end{aligned} \tag{5.41}$$

If $2 \leq \ell < 8$, applying the above second-order decomposition of $(\lambda I + T_{\mathbf{X}_1})^{-1}$ and taking $\lambda \leq 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^8\right] \\
&\leq (1+\mu_1)^{8-\ell}\frac{1}{\lambda^{8-\ell}}\mathbb{E}\left[\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^\ell\right] \\
&\overset{(i)}{=} (1+\mu_1)^{8-\ell}\frac{3^\ell}{\lambda^{8-\ell}}\mathbb{E}\Bigg[\Big\|\frac{1}{3}I+\frac{1}{3}(\lambda I+T)^{-1/2}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1/2} \\
&\quad+\frac{1}{3}(\lambda I+T)^{-1/2}(T-T_{\mathbf{X}_1})(\lambda I+T_{\mathbf{X}_1})^{-1}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1/2}\Big\|^\ell\Bigg] \\
&\overset{(ii)}{\leq} (1+\mu_1)^{8-\ell}\frac{3^{\ell-1}}{\lambda^{8-\ell}}\Bigg\{\|I\|^\ell+\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1/2}\right\|^\ell\right] \\
&\quad+\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}(T-T_{\mathbf{X}_1})\right\|^{2\ell}\right]\frac{1}{\lambda^\ell}\Bigg\} \qquad (5.42)\\
&\overset{(iii)}{\leq} (1+\mu_1)^{8-\ell}\frac{3^{\ell-1}}{\lambda^{8-\ell}}\Bigg\{\|I\|^\ell+\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}(T-T_{\mathbf{X}_1})(\lambda I+T)^{-1/2}\right\|_{HS}^{2\ell}\right]^{\frac{1}{2}} \\
&\quad+\mathbb{E}\left[\left\|(\lambda I+T)^{-1/2}(T-T_{\mathbf{X}_1})\right\|_{HS}^{2\ell}\right]\frac{1}{\lambda^\ell}\Bigg\} \\
&\overset{(iv)}{\leq} (1+\mu_1)^{8-\ell}\frac{3^{\ell-1}}{\lambda^{8-\ell}}\left[1+c^{\frac{1}{2}}(\ell)2^\ell\rho^{2\ell}\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{2}}+c(\ell)2^{2\ell}\rho^{4\ell}trace^\ell(T)\frac{1}{\lambda^\ell}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^\ell\right] \\
&\leq (1+\mu_1)^8 3^7 c(8)2^{16}\rho^{32}\max\left\{1, trace^8(T)\right\}\frac{1}{\lambda^{8-\ell}}\left[1+\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{2}}+\frac{1}{\lambda^\ell}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^\ell\right] \\
&= c_7^4\lambda^{\ell-8}\left[1+\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{2}}+\lambda^{-\ell}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^\ell\right],
\end{aligned}$$

where $c_7^4 := (1+\mu_1)^8 3^7 c(8)2^{16}\rho^{32}\max\left\{1, trace^8(T)\right\}$. Here equality (i) is from the second-order decomposition of $(\lambda I + T_{\mathbf{X}_1})^{-1}$ (5.41). Inequality (ii) uses Jensen's inequality. Inequality (iii) is due to Cauchy-Schwartz inequality and (2.2). Inequality (iv) follows from estimations (5.19) and (5.20) in Lemma 17.

Analogously, if $\ell \geq 8$, applying the second-order decomposition of $(\lambda I + T_{\mathbf{X}_1})^{-1}$ (5.41) and taking $\lambda \leq 1$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\lambda I+T)^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}(\lambda I+T)^{1/2}\right\|^8\right] \\
&\leq 3^7\left[1+c^{\frac{1}{2}}(8)2^8\rho^{16}\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^4+c(8)2^{16}\rho^{32}trace^8(T)\frac{1}{\lambda^8}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^8\right] \\
&\leq c_7^4\left[1+\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^4+\frac{1}{\lambda^8}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^8\right], \qquad (5.43)
\end{aligned}$$

where $c_7^4 = (1+\mu_1)^8 3^7 c(8) 2^{16} \rho^{32} \max\left\{1, trace^8(T)\right\}$.

We can now prove (3.10).

If $2 \le \ell < 8$, taking $\lambda \le 1$, for the term $\mathbb{E}\left[\left\|L_C^{1/2} L_K^{1/2}(\lambda I + T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n \alpha_i\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1}\right]$, combining (5.38) and (5.42) with (5.18) in Lemma 16 and (5.27) in Lemma 18, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2} L_K^{1/2}(\lambda I + T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^n \alpha_i\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1}\right] \\
&\le c_6 c_7 c^{\frac{1}{4}}(\ell) 2^\ell \rho^\ell \lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{2}} + \lambda^{-\ell}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\ell}\right]^{\frac{1}{4}} \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \frac{m}{N}\left(1 + \lambda^{4\theta}\mathcal{N}^2(\lambda)\right)^{\frac{1}{2}} \\
&\le c_6 c_7 c^{\frac{1}{4}}(\ell) 2^\ell \rho^\ell \lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right] \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \frac{m}{N}\left(1 + \lambda^{2\theta}\mathcal{N}(\lambda)\right) \\
&= b_1(\ell) \lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right] \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \frac{m}{N}\left(1 + \lambda^{2\theta}\mathcal{N}(\lambda)\right),
\end{aligned}$$

where $b_1(\ell) := c_6 c_7 c^{\frac{1}{4}}(\ell) 2^\ell \rho^\ell$.

For the term $\mathbb{E}\left[\left\|L_C^{1/2} L_K^{1/2}(\lambda I + T_{\mathbf{X}_1})^{-1} L_K^{1/2} X_{1,i}\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1}\right], \forall 1 \le i \le n$, combining (5.39), (5.40) and (5.42) with (5.18) in Lemma 16, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2} L_K^{1/2}(\lambda I + T_{\mathbf{X}_1})^{-1} L_K^{1/2} X_{1,i}\right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1}\right] \\
&\le c_7 c^{\frac{1}{4}}(\ell) 2^\ell \rho^{\ell+2} \lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right] \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \mathcal{N}(\lambda) \\
&= b_2(\ell) \lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right] \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \mathcal{N}(\lambda),
\end{aligned}$$

where $b_2(\ell) := c_7 c^{\frac{1}{4}}(\ell) 2^\ell \rho^{\ell+2}$.

Then recall that $n = N/m$, combining the above two estimations with (5.35), (5.36) and (5.37) yields

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\le 2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + 16\frac{\mathcal{N}(\lambda)}{N}\left(c_2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 + \sigma^2\right) + 8c_2\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2 \\
&\quad + b_1(\ell)\lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right] \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \frac{4+2m}{N}\left(1 + \lambda^{2\theta}\mathcal{N}(\lambda)\right) \\
&\quad + b_2(\ell)\lambda^{\frac{\ell-8}{4}} \left[1 + \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^{\frac{\ell}{4}}\right] \left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}} \frac{4\sigma^2}{N}\mathcal{N}(\lambda).
\end{aligned}$$

This completes the proof of (3.10).

We next give the proof of (3.11).

If $\ell \geq 8$, taking $\lambda \leq 1$, employing (5.43) and following the same arguments in the proof of (3.10), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}\left(\lambda I+T_{\mathbf{X}_1}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\leq c_6c_7c^{\frac{1}{4}}(\ell)2^{\ell}\rho^{\ell}\left[1+\frac{m\mathcal{N}^2(\lambda)}{N}+\frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{m}{N}\left(1+\lambda^{2\theta}\mathcal{N}(\lambda)\right)\\
&=b_1(\ell)\left[1+\frac{m\mathcal{N}^2(\lambda)}{N}+\frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{m}{N}\left(1+\lambda^{2\theta}\mathcal{N}(\lambda)\right),
\end{aligned}$$

where $b_1(\ell)=c_6c_7c^{\frac{1}{4}}(\ell)2^{\ell}\rho^{\ell}$.

And

$$\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}L_K^{1/2}X_{1,i}\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\leq c_3c^{\frac{1}{4}}(\ell)2^{\ell}\rho^{\ell+2}\left[1+\frac{m\mathcal{N}^2(\lambda)}{N}+\frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\mathcal{N}(\lambda)\\
&=b_2(\ell)\left[1+\frac{m\mathcal{N}^2(\lambda)}{N}+\frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\mathcal{N}(\lambda),
\end{aligned}$$

where $b_2(\ell)=c_7c^{\frac{1}{4}}(\ell)2^{\ell}\rho^{\ell+2}$.

And then

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda})-\mathcal{R}(\beta_0)\right)\right]\\
&\leq 2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2+16\frac{\mathcal{N}(\lambda)}{N}(c_2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2+\sigma^2)+8c_2\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2\\
&\quad+b_1(\ell)\left[1+\frac{m\mathcal{N}^2(\lambda)}{N}+\frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{4+2m}{N}\left(1+\lambda^{2\theta}\mathcal{N}(\lambda)\right)\\
&\quad+b_2(\ell)\left[1+\frac{m\mathcal{N}^2(\lambda)}{N}+\frac{1}{\lambda^2}\left(\frac{m\mathcal{N}(\lambda)}{N}\right)^2\right]\left(\frac{m\mathcal{N}^2(\lambda)}{N}\right)^{\frac{\ell}{4}}\frac{4\sigma^2}{N}\mathcal{N}(\lambda).
\end{aligned}$$

We have completed the proof of inequality (3.11). The proof of Theorem 4 is then finished. ■

We next prove Corollary 5.

Proof of Corollary 5. We prove the desired bounds in three cases, respectively.

When $2 \le \ell \le 4$, taking $\lambda \le 1$, (3.10) and (5.9) imply

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\lesssim \lambda^{2\theta} + \frac{\lambda^{-p}}{N} + m\frac{\lambda^{2\theta-p}}{N} \\
&\quad + \lambda^{\frac{\ell-8}{4}}\left[1 + \left(\frac{m\lambda^{-2p}}{N}\right)^{\frac{\ell}{8}} + \lambda^{-\frac{\ell}{4}}\left(\frac{m\lambda^{-p}}{N}\right)^{\frac{\ell}{4}}\right]\left(\frac{m\lambda^{-2p}}{N}\right)^{\frac{\ell}{4}}\left(\frac{m}{N} + \frac{m\lambda^{2\theta-p}}{N} + \frac{\lambda^{-p}}{N}\right).
\end{aligned} \tag{5.44}$$

Noting that $\theta \le \frac{1}{2} \le \frac{p\ell+8}{4\ell}$, taking $m \le N^r$ for some $0 \le r \le \frac{2\theta}{2\theta+p}$, we have

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim \max\left\{N^{\frac{2\theta(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{2\theta\ell(r-1)-8\theta}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{2\theta(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{4\theta\ell(r-1)-8\theta}{8+4p+8\theta+3p\ell}}\right\}$$

provided that

$$\lambda = \max\left\{N^{\frac{(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{\ell(r-1)-4}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{2\ell(r-1)-4}{8+4p+8\theta+3p\ell}}\right\}.$$

This completes the proof of case $2 \le \ell \le 4$.

When $5 \le \ell \le 7$, taking $\lambda \le 1$, inequality (5.44) still holds. Then taking $\lambda = N^{-\frac{1}{2\theta+p}}$ yields

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim N^{\frac{2\theta}{2\theta+p}}$$

provided that $\frac{p\ell+8}{4\ell} \le \theta \le \frac{1}{2}$ and $m \le \min\left\{N^{\frac{8+p\ell-4p-4\theta\ell}{(4+2\ell)(2\theta+p)}}, N^{\frac{8+p\ell-8\theta-4\theta\ell}{(4+2\ell)(2\theta+p)}}\right\}$.

If $\theta\frac{p\ell+8}{4\ell}$, take $m \le N^r$ for some $0 \le r \le \frac{2\theta}{2\theta+p}$. Then we have

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim \max\left\{N^{\frac{2\theta(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{2\theta\ell(r-1)-8\theta}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{2\theta(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{4\theta\ell(r-1)-8\theta}{8+4p+8\theta+3p\ell}}\right\}$$

provided that

$$\lambda = \max\left\{N^{\frac{(4+\ell)(r-1)}{8+8\theta+2p\ell-\ell}}, N^{\frac{\ell(r-1)-4}{8+4p+8\theta+2p\ell-\ell}}, N^{\frac{(4+2\ell)(r-1)}{8+8\theta+3p\ell}}, N^{\frac{2\ell(r-1)-4}{8+4p+8\theta+3p\ell}}\right\}.$$

This completes the proof of case $5 \le \ell \le 7$.

When $\ell \ge 8$, taking $\lambda \le 1$, (3.11) and (5.9) imply

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \\
&\lesssim \lambda^{2\theta} + \frac{\lambda^{-p}}{N} + m\frac{\lambda^{2\theta-p}}{N} \\
&\quad + \left(1 + \frac{m\lambda^{-2p}}{N} + \frac{m^2\lambda^{-2p-2}}{N^2}\right)\left(\frac{m\lambda^{-2p}}{N}\right)^{\frac{\ell}{4}}\left(\frac{m}{N} + \frac{m\lambda^{2\theta-p}}{N} + \frac{\lambda^{-p}}{N}\right).
\end{aligned}$$

Taking $\lambda = N^{-\frac{1}{2\theta+p}}$ yields

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim N^{\frac{2\theta}{2\theta+p}}$$

provided that $\frac{p\ell+8}{2\ell+16} \leq \theta \leq 1/2$ and $m \leq \min\left\{N^{\frac{8+p\ell-4p-16\theta-2\theta\ell}{(12+\ell)(2\theta+p)}}, N^{\frac{8+p\ell-24\theta-2\theta\ell}{(12+\ell)(2\theta+p)}}\right\}$.

If $\theta < \frac{p\ell+8}{2\ell+16}$, take $m \leq N^r$ for some $0 \leq r \leq \frac{2\theta}{2\theta+p}$. Then we have

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim \max\left\{N^{\frac{\theta(4+\ell)(r-1)}{4\theta+p\ell}}, N^{\frac{\theta\ell(r-1)-4\theta}{2p+4\theta+p\ell}}, N^{\frac{\theta(12+\ell)(r-1)}{4+4p+4\theta+p\ell}}, N^{\frac{\theta(8+\ell)(r-1)-4\theta}{4+6p+4\theta+p\ell}}\right\}$$

provided that

$$\lambda = \max\left\{N^{\frac{(4+\ell)(r-1)}{8\theta+2p\ell}}, N^{\frac{\ell(r-1)-4}{4p+8\theta+2p\ell}}, N^{\frac{(12+\ell)(r-1)}{8+8p+8\theta+2p\ell}}, N^{\frac{(8+\ell)(r-1)-4}{8+12p+8\theta+2p\ell}}\right\}.$$

We have completed the proof of case $\ell \geq 8$. Then proof of Corollary 5 is then finished. ■

We next turn to prove Theorem 6 and Corollary 7. If Assumption 5 is satisfied, we can estimate the probability of event $\mathcal{U}_1$ better than Lemma 14 and Lemma 16.

Lemma 20 *Suppose that Assumption 5 is satisfied, then there holds*

$$\mathbb{P}(\mathcal{U}_1) \leq c_4^4\left(1 + \frac{m^2\mathcal{N}^2(\lambda)}{N^2}\right)\mathcal{N}(\lambda)\exp\left(-c_5\frac{N}{m\mathcal{N}(\lambda)}\right). \tag{5.45}$$

Where c_4 and c_5 are universal constants and $\mathcal{N}(\lambda)$ is the effective dimension given by (3.4).

Proof Our proof relies on the Bernstein's inequality for the sum of self-adjoint random operators (see, Lemma 25). Recalling (5.4), define

$$\zeta_i := (\lambda I + T)^{-1/2}L_K^{1/2}X_{1,i} \otimes L_K^{1/2}X_{1,i}(\lambda I + T)^{-1/2} \quad \textit{and} \quad \eta_i := \frac{1}{n}\left(\zeta_i - \mathbb{E}[\zeta_i]\right), \quad 1 \leq i \leq n.$$

Then we can write

$$(\lambda I + T)^{-1/2}(T_{\mathbf{X}_1} - T)(\lambda I + T)^{-1/2} = \sum_{i=1}^{n}\eta_i.$$

Using expression (3.8), we have

$$\begin{aligned}
&\left\|(\lambda I + T)^{-1/2}L_K^{1/2}X \otimes L_K^{1/2}X(\lambda I + T)^{-1/2}\right\| \\
&= \sup_{\|f\|_{\mathcal{L}^2}=1, \|g\|_{\mathcal{L}^2}=1}\left\langle(\lambda I + T)^{-1/2}L_K^{1/2}X \otimes L_K^{1/2}X(\lambda I + T)^{-1/2}f, g\right\rangle_{\mathcal{L}^2} \\
&\leq \left\|(\lambda I + T)^{-1/2}L_K^{1/2}X\right\|_{\mathcal{L}^2}^2 = \sum_{k=1}^{\infty}\frac{\mu_k}{\lambda + \mu_k}\xi_k^2 \\
&\overset{(*)}{\leq} \rho^2\sum_{k=1}^{\infty}\frac{\mu_k}{\lambda + \mu_k} = \rho^2\mathcal{N}(\lambda).
\end{aligned} \tag{5.46}$$

Here inequality $(*)$ is from Assumption 5.

Then for any $1 \leq i \leq n$, one can calculate

$$\|\eta_i\| = \left\| \frac{1}{n}\left(\zeta_i - \mathbb{E}[\zeta_i]\right) \right\| \overset{(i)}{\leq} \frac{1}{n}\|\zeta_i\| + \frac{1}{n}\mathbb{E}\left[\|\zeta_i\|\right] \overset{(ii)}{\leq} 2\rho^2 \frac{\mathcal{N}(\lambda)}{n}. \tag{5.47}$$

Here inequality (i) uses the triangle inequality and Jensen's inequality. Inequality (ii) follows from (5.46). And then we have

$$\begin{aligned}
&\left\| \mathbb{E}\left[(\overline{\eta})^2\right] \right\| = \left\| \mathbb{E}\left[\left(\sum_{i=1}^n \eta_i \right)^2 \right] \right\| \\
&\overset{(i)}{=} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2}=1} \sum_{i=1}^n \left\langle f, \mathbb{E}[\eta_i^2] f \right\rangle_{\mathcal{L}^2} \\
&= \frac{1}{n^2} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2}=1} \sum_{i=1}^n \left(\left\langle f, \mathbb{E}[\zeta_i^2] f \right\rangle_{\mathcal{L}^2} - \left\langle f, [\mathbb{E}\zeta_i]^2 f \right\rangle_{\mathcal{L}^2} \right) \\
&\leq \frac{1}{n^2} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2}=1} \sum_{i=1}^n \left\langle f, \mathbb{E}[\zeta_i^2] f \right\rangle_{\mathcal{L}^2} \qquad (5.48) \\
&= \frac{1}{n^2} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2}=1} \sum_{i=1}^n \mathbb{E}\left[\left\langle (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i}, f \right\rangle_{\mathcal{L}^2}^2 \left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \right] \\
&\overset{(ii)}{\leq} \rho^2 \frac{\mathcal{N}(\lambda)}{n^2} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2}=1} \sum_{i=1}^n \mathbb{E}\left[\left(\sum_{k=1}^\infty \sqrt{\frac{1}{\lambda + \mu_k}} \langle f, \phi_k \rangle_{\mathcal{L}^2} \left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2} \right)^2 \right] \\
&\overset{(iii)}{=} \rho^2 \frac{\mathcal{N}(\lambda)}{n^2} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2}=1} \sum_{i=1}^n \sum_{k=1}^\infty \frac{\mu_k}{\lambda + \mu_k} \langle f, \phi_k \rangle_{\mathcal{L}^2}^2 \leq \rho^2 \frac{\mathcal{N}(\lambda)}{n}.
\end{aligned}$$

Here equality (i) is due to the equivalent expression of the operator norm of a nonnegative operator (2.1) and the fact that $\mathbb{E}[\eta_i] = 0$. Inequality (ii) follows from the fact that

$$\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \leq \rho^2 \mathcal{N}(\lambda)$$

which is given by (5.46). Equality (iii) is from the fact that

$$\mathbb{E}\left[\left\langle L_K^{1/2} X, \phi_j \right\rangle_{\mathcal{L}^2} \left\langle L_K^{1/2} X, \phi_k \right\rangle_{\mathcal{L}^2} \right] = \langle T\phi_j, \phi_k \rangle_{\mathcal{L}^2} = \mu_k \delta_j^k,$$

We also need the following estimate given by

$$\begin{aligned}
trace\left(\mathbb{E}\left[(\overline{\eta})^2\right] \right) &\overset{(i)}{=} \sum_{k=1}^\infty \left\langle \mathbb{E}\left[\left(\sum_{i=1}^n \eta_i \right)^2 \right] \phi_k, \phi_k \right\rangle_{\mathcal{L}^2} \\
&= \sum_{i=1}^n \sum_{k=1}^\infty \left\langle \mathbb{E}\left[\eta_i^2\right] \phi_k, \phi_k \right\rangle_{\mathcal{L}^2} \leq \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^\infty \left\langle \mathbb{E}\left[\zeta_i^2\right] \phi_k, \phi_k \right\rangle_{\mathcal{L}^2} \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^\infty \frac{1}{\lambda + \mu_k} \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2}^2 \right]
\end{aligned}$$

$$\overset{(ii)}{\leq} \rho^2 \frac{\mathcal{N}(\lambda)}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{\infty} \frac{1}{\lambda + \mu_k} \mathbb{E}\left[\left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2}^2 \right] \overset{(iii)}{\leq} \rho^2 \frac{\mathcal{N}^2(\lambda)}{n}. \tag{5.49}$$

Here equality (i) is from the formulation of the trace norm of an operator (2.3). Inequality (ii) is due to the fact that $\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \leq \rho^2 \mathcal{N}(\lambda)$. Inequality (iii) follows from the calculation that $\sum_{k=1}^{\infty} \frac{1}{\lambda+\mu_k} \mathbb{E}\left[\left\langle L_K^{1/2} X_{1,i}, \phi_k \right\rangle_{\mathcal{L}^2}^2 \right] = \frac{\mu_k}{\lambda+\mu_k} = \mathcal{N}(\lambda)$.

Recall that $n = N/m$. Based on (5.47), (5.48) and (5.49), one can apply Lemma 25 with $L = 2\rho^2 \frac{m\mathcal{N}(\lambda)}{N}$, $v = \rho^2 \frac{m\mathcal{N}(\lambda)}{N}$, $d = \mathcal{N}(\lambda)$ and $s = 1/2$ to obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{U}_1) &= \mathbb{P}\left(\left\| \sum_{i=1}^{n} \eta_i \right\| \geq 1/2 \right) \\
&\leq \left[1 + 6 \left(\rho^2 \frac{4m\mathcal{N}(\lambda)}{N} + \rho^2 \frac{4m\mathcal{N}(\lambda)}{3N} \right)^2 \right] \mathcal{N}(\lambda) \exp\left(- \frac{3N}{32\rho^2 m \mathcal{N}(\lambda)} \right) \\
&\leq c_4^4 \left(1 + \frac{m^2 \mathcal{N}^2(\lambda)}{N^2} \right) \mathcal{N}(\lambda) \exp\left(-c_5 \frac{N}{m\mathcal{N}(\lambda)} \right),
\end{aligned}$$

where $c_4^4 := \left[1 + 6 \left(4\rho^2 + \frac{4\rho^2}{3} \right)^2 \right]$ and $c_5 := \frac{3}{32\rho^2}$. The proof is then completed. ■

Now we can prove Theorem 6.

Proof of Theorem 6. Under Assumption 5, there holds

$$\left\| L_K^{1/2} X \right\|_{\mathcal{L}^2} = \left(\sum_{k=1}^{\infty} \mu_k \xi_k^2 \right)^{\frac{1}{2}} \leq \rho \left(\sum_{k=1}^{\infty} \mu_k \right)^{\frac{1}{2}} = \rho \cdot trace^{\frac{1}{2}}(T). \tag{5.50}$$

Therefore, recalling (5.4), for any $1 \leq i \leq n$, we can write

$$\begin{aligned}
&\mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \\
&\leq \frac{\mu_1}{\lambda^2} \mathbb{E}\left[\left\| L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \overset{(i)}{\leq} \frac{\mu_1}{\lambda^2} \left[\mathbb{E} \left\| L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^4 \right]^{\frac{1}{2}} \mathbb{P}^{\frac{1}{2}}(\mathcal{U}_1) \\
&\overset{(ii)}{\leq} c_4 \mu_1 \rho^2 trace(T) \frac{1}{\lambda^2} \left(1 + \frac{m\mathcal{N}(\lambda)}{N} \right) \mathcal{N}^{\frac{1}{2}}(\lambda) \exp\left(- \frac{c_5 N}{2m\mathcal{N}(\lambda)} \right).
\end{aligned} \tag{5.51}$$

Here inequality (i) is from Cauchy-Schwartz inequality. Inequality (ii) follows from (5.45) in Lemma 20 and (5.51).

Then under Assumption 1 and 5, one can calculate

$$\begin{aligned}
&\mathbb{E}\left[\left\| L_K^{1/2} X \left\langle X, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} \right\|_{\mathcal{L}^2}^4 \right] \\
&\overset{(i)}{\leq} \rho^4 trace^2(T) \mathbb{E}\left[\left\langle X, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2}^4 \right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} c_1\rho^4 trace^2(T)\left[\mathbb{E}\left\langle X,\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}^2\right]^2\\
&= c_1\rho^4 trace^2(T)\left\|L_C^{1/2}\left(\beta_0-L_K^{1/2}f_\lambda\right)\right\|_{\mathcal{L}^2}^4\\
&\overset{(iii)}{\leq} c_1\rho^4 trace^2(T)\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta},
\end{aligned}
$$

where inequality (i) follows from (5.51), inequality (ii) uses the fourth-moment condition (3.3) and inequality (iii) is due to Lemma 11. Then employing the above bound and following the same estimates in the proof of (5.27), taking $\lambda \leq 1$, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left(L_K^{1/2}X_{1,i}\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}-\lambda f_\lambda\right)\right\|_{\mathcal{L}^2}^4\right]\\
&\leq \frac{192m^2}{N^2}\left(c_1\rho^4 trace^2(T)\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^{4\theta}+\mu_1^{4\theta}\|\gamma_0\|_{\mathcal{L}^2}^4\lambda^2\right)\leq c_3^2\frac{m^2}{N^2}\lambda^{4\theta},
\end{aligned}
\tag{5.52}
$$

where $c_3 := 192\left(c_1\rho^4 trace^2(T)||\gamma_0||_{\mathcal{L}^2}^4+\max\{\mu_1^2,1\}||\gamma_0||_{\mathcal{L}^2}^4\right)$.

For simplicity of notations, define

$$
\alpha_i := L_K^{1/2}X_{1,i}\left\langle X_{1,i},\beta_0-L_K^{1/2}f_\lambda\right\rangle_{\mathcal{L}^2}-\lambda f_\lambda,\quad i=1,2,\cdots,n.
$$

Then we can write

$$
\begin{aligned}
&\mathbb{E}\left[\left\|L_C^{1/2}L_K^{1/2}(\lambda I+T_{\mathbf{X}_1})^{-1}\frac{1}{n}\sum_{i=1}^{n}\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\\
&\leq \frac{\mu_1}{\lambda^2}\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha_i\right\|_{\mathcal{L}^2}^2\mathbb{I}_{\mathcal{U}_1}\right]\overset{(i)}{\leq}\frac{\mu_1}{\lambda^2}\left[\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\alpha_i\right\|_{\mathcal{L}^2}^4\right]^{\frac{1}{2}}\mathbb{P}^{\frac{1}{2}}(\mathcal{U}_1)\\
&\overset{(ii)}{\leq} c_3c_4\mu_1\frac{m}{N\lambda^{2-2\theta}}\left(1+\frac{m\mathcal{N}(\lambda)}{N}\right)\mathcal{N}^{\frac{1}{2}}(\lambda)\exp\left(-\frac{c_5N}{2m\mathcal{N}(\lambda)}\right).
\end{aligned}
\tag{5.53}
$$

Here inequality (i) is due to Cauchy-Schwartz inequality. Inequality (ii) is from (5.45) in Lemma 20 and (5.52).

Finally, employing estimates (5.51) and (5.53), we follow the same arguments in the proof of (3.10) and then obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda})-\mathcal{R}(\beta_0)\right)\right]\\
&\leq 2\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2+16\frac{\mathcal{N}(\lambda)}{N}\left(c_1\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2+\sigma^2\right)+8c_1\frac{m}{N}\mathcal{N}(\lambda)\lambda^{2\theta}\|\gamma_0\|_{\mathcal{L}^2}^2\\
&\quad+c_3c_4\mu_1\frac{4+2m}{N\lambda^{2-2\theta}}\left(1+\frac{m\mathcal{N}(\lambda)}{N}\right)\mathcal{N}^{\frac{1}{2}}(\lambda)\exp\left(-\frac{c_5N}{2m\mathcal{N}(\lambda)}\right)\\
&\quad+c_4\mu_1\rho^2 trace(T)\frac{4\sigma^2}{N\lambda^2}\left(1+\frac{m\mathcal{N}(\lambda)}{N}\right)\mathcal{N}^{\frac{1}{2}}(\lambda)\exp\left(-\frac{c_5N}{2m\mathcal{N}(\lambda)}\right)
\end{aligned}
$$

The proof of Theorem 6 is then finished. ■

We next turn to prove Corollary 7.

Proof of Corollary 7. Taking $m \leq o\left(\frac{N^{\frac{2\theta}{2\theta+p}}}{\log N}\right)$ and $\lambda = N^{-\frac{1}{2\theta+p}}$, (5.9) implies

$$m\frac{\mathcal{N}(\lambda)}{N} \lesssim m\frac{\lambda^{-p}}{N} \leq o\left(\frac{1}{\log N}\right).$$

Therefore, for any $r > 0$, there holds

$$\liminf_{N\to\infty} N^r \exp\left(-\frac{c_5 N}{2m\mathcal{N}(\lambda)}\right) = 0.$$

Then using Theorem 6, we obtain

$$\mathbb{E}\left[\mathcal{R}(\overline{\beta}_{S,\lambda}) - \mathcal{R}(\beta_0)\right] \lesssim \lambda^{2\theta} + \frac{\mathcal{N}(\lambda)}{N} + m\frac{\mathcal{N}(\lambda)}{N}\lambda^{2\theta} \lesssim N^{-\frac{2\theta}{2\theta+p}}.$$

The proof of Corollary 7 is then finished. ■

Finally, we will provide the proof of Theorem 9. Before that, we establish the following lemma to estimate $\mathbb{P}(\mathcal{U}_1)$ based on Assumption 6.

Lemma 21 *Suppose that Assumption 6 is satisfied, then there holds*

$$\mathbb{P}(\mathcal{U}_1) \leq \left[1 + 6\left(\frac{4m\kappa^2}{N\lambda^t} + \frac{4m\kappa^2}{3N\lambda^t}\right)^2\right]\mathcal{N}(\lambda)\exp\left(-\frac{3N\lambda^t}{32m\kappa^2}\right). \tag{5.54}$$

Proof Our proof relies on Lemma 25. Recalling (5.4) and the definition of $\mathcal{U}_1$ given by

$$\mathcal{U}_1 = \left\{\mathbf{X}_1 : \left\|(\lambda I + T)^{-1/2}(T_{\mathbf{X}_1} - T)(\lambda I + T)^{-1/2}\right\| \geq 1/2\right\},$$

let

$$\zeta_i := (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} (\lambda I + T)^{-1/2} \quad \textit{and} \quad \eta_i := \frac{1}{n}\left(\zeta_i - \mathbb{E}[\zeta_i]\right), \quad 1 \leq i \leq n.$$

Then

$$(\lambda I + T)^{-1/2}(T_{\mathbf{X}_1} - T)(\lambda I + T)^{-1/2} = \sum_{i=1}^{n} \eta_i.$$

Due to the decomposition of $L_K^{1/2} X$ in (3.8), there holds

$$\begin{aligned}
&\left\|(\lambda I + T)^{-1/2} L_K^{1/2} X \otimes L_K^{1/2} X (\lambda I + T)^{-1/2}\right\| \\
&= \sum_{\|f\|_{\mathcal{L}^2}=1, \|g\|_{\mathcal{L}^2}=1} \left\langle (\lambda I + T)^{-1/2} L_K^{1/2} X \otimes L_K^{1/2} X (\lambda I + T)^{-1/2} f, g \right\rangle_{\mathcal{L}^2} \\
&\leq \left\|(\lambda I + T)^{-1/2} L_K^{1/2} X\right\|_{\mathcal{L}^2}^2 = \sum_{k=1}^{\infty} \frac{\mu_k^{1-t}\mu_k^t \xi_k^2}{\lambda + \mu_k} \leq \frac{1}{\lambda^t}\sum_{k=1}^{\infty}\mu_k^t\xi_k^2 \overset{(*)}{\leq} \frac{\kappa^2}{\lambda^t}.
\end{aligned} \tag{5.55}$$

Here inequality $(*)$ follows from Assumption 6. Then for any $1 \le i \le n$,

$$\|\eta_i\| = \left\| \frac{1}{n} \left(\zeta_i - \mathbb{E}[\zeta_i] \right) \right\| \le \frac{1}{n} \|\zeta_i\| + \frac{1}{n} \mathbb{E}\left[\|\zeta_i\| \right] \overset{(\dagger)}{\le} \frac{2\kappa^2}{n}. \tag{5.56}$$

Here inequality $(\dagger)$ follows from (5.55) and the definition that $\zeta_i = (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \otimes L_K^{1/2} X_{1,i} (\lambda I + T)^{-1/2}$.

Following the same arguments of (5.48), we have

$$\begin{aligned}
&\left\| \mathbb{E}\left[(\overline{\eta})^2 \right] \right\| \\
&\le \frac{1}{n^2} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2} = 1} \sum_{i=1}^{n} \mathbb{E}\left[\left\langle (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i}, f \right\rangle_{\mathcal{L}^2}^2 \left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \right] \\
&= \frac{1}{n} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2} = 1} \mathbb{E}\left[\left\langle (\lambda I + T)^{-1/2} L_K^{1/2} X, f \right\rangle_{\mathcal{L}^2}^2 \left\| (\lambda I + T)^{-1/2} L_K^{1/2} X \right\|_{\mathcal{L}^2}^2 \right] \qquad (5.57) \\
&\overset{(i)}{\le} \frac{\kappa^2}{n\lambda^t} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2} = 1} \mathbb{E}\left[\left(\sum_{k=1}^{\infty} \sqrt{\frac{\mu_k}{\lambda + \mu_k}} \langle f, \phi_k \rangle_{\mathcal{L}^2} \xi_k \right)^2 \right] \\
&\overset{(ii)}{\le} \frac{\kappa^2}{n\lambda^t} \sup_{f \in \mathcal{L}^2(\mathcal{T}), \|f\|_{\mathcal{L}^2} = 1} \sum_{k=1}^{\infty} \frac{\mu_k}{\lambda + \mu_k} \langle f, \phi_k \rangle_{\mathcal{L}^2}^2 \le \frac{\kappa^2}{n\lambda^t}.
\end{aligned}$$

Here inequality (i) is from (5.55) and inequality (ii) is due to the decomposition of $L_K^{1/2} X$ in (3.8) and the fact that $\mathbb{E}[\xi_j \xi_k] = \delta_j^k$. And following the same arguments of (5.49), we have

$$\begin{aligned}
&trace(\mathbb{E}[(\overline{\eta})^2]) \\
&\le \frac{1}{n^2} \sum_{i=1}^{n} \sum_{k=1}^{\infty} \frac{1}{\lambda + \mu_k} \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \langle L_K^{1/2} X_{1,i}, \phi_k \rangle_{\mathcal{L}^2}^2 \right] \\
&= \frac{1}{n} \sum_{k=1}^{\infty} \frac{1}{\lambda + \mu_k} \mathbb{E}\left[\left\| (\lambda I + T)^{-1/2} L_K^{1/2} X \right\|_{\mathcal{L}^2}^2 \langle L_K^{1/2} X, \phi_k \rangle_{\mathcal{L}^2}^2 \right] \\
&\overset{(\dagger)}{\le} \frac{\kappa^2}{n\lambda^t} \sum_{k=1}^{\infty} \frac{\mu_k}{\lambda + \mu_k} \mathbb{E}\left[\xi_k^2 \right] \le \frac{\kappa^2 \mathcal{N}(\lambda)}{n\lambda^t}.
\end{aligned} \tag{5.58}$$

Here inequality $(\dagger)$ is from (5.55) and decomposition of $L_K^{1/2} X$ (3.8).

Recall that $n = N/m$. (5.56), (5.57) and (5.58) imply that we can employ Lemma 25 with $L = \frac{2m\kappa^2}{N\lambda^t}$, $v = \frac{m\kappa^2}{N\lambda^t}$, $d = \mathcal{N}(\lambda)$ and $s = 1/2$ to obtain

$$\begin{aligned}
\mathbb{P}(\mathcal{U}_1) &= \mathbb{P}\left(\left\| \sum_{i=1}^{n} \eta_i \right\| \ge 1/2 \right) \\
&\le \left[1 + 6 \left(\frac{4m\kappa^2}{N\lambda^t} + \frac{4m\kappa^2}{3N\lambda^t} \right)^2 \right] \mathcal{N}(\lambda) \exp\left(-\frac{3N\lambda^t}{32m\kappa^2} \right).
\end{aligned}$$

This completes the proof. ■

Now we are ready to prove Theorem 9.

Proof of Theorem 9. Recalling (5.4) and $Y_{1,i} = \langle X_{1,i}, \beta_0 \rangle_{\mathcal{L}^2} + \epsilon_{1,i}, \forall 1 \le i \le n$, we define

$$\alpha_i = L_K^{1/2} X_{1,i} \left\langle X_{1,i}, \beta_0 - L_K^{1/2} f_\lambda \right\rangle_{\mathcal{L}^2} - \lambda f_\lambda, \quad i = 1, 2, \cdots, n.$$

Under Assumption 6, we have

$$\left\| L_K^{1/2} X \right\|_{\mathcal{L}^2} = \left(\sum_{k=1}^{\infty} \mu_k \xi_k^2 \right)^{\frac{1}{2}} \le \mu_1^{\frac{1-t}{2}} \left(\sum_{k=1}^{\infty} \mu_k^t \xi_k^2 \right)^{\frac{1}{2}} \le \mu_1^{\frac{1-t}{2}} \kappa.$$

Employing the above estimation and (5.54), following the same arguments in the proof of Theorem 6, we have

$$\begin{aligned} &\mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} L_K^{1/2} X_{1,i} \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \\ &\lesssim \frac{1}{\lambda^2} \left(1 + \frac{m}{N\lambda^t} \right) \mathcal{N}^{\frac{1}{2}}(\lambda) \exp\left(-\frac{3N\lambda^t}{64m\kappa^2} \right) \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}\left[\left\| L_C^{1/2} L_K^{1/2} (\lambda I + T_{\mathbf{X}_1})^{-1} \frac{1}{n} \sum_{i=1}^{n} \alpha_i \right\|_{\mathcal{L}^2}^2 \mathbb{I}_{\mathcal{U}_1} \right] \\ &\lesssim \frac{m}{N\lambda^{2-2\theta}} \left(1 + \frac{m}{N\lambda^t} \right) \mathcal{N}^{\frac{1}{2}}(\lambda) \exp\left(-\frac{3N\lambda^t}{64m\kappa^2} \right) \end{aligned}$$

and then

$$\begin{aligned} \mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \right) \right] \lesssim{}& \lambda^{2\theta} + \frac{\mathcal{N}(\lambda)}{N} + \frac{m}{N} \mathcal{N}(\lambda) \lambda^{2\theta} \\ &+ \frac{m}{N\lambda^{2-2\theta}} \left(1 + \frac{m}{N\lambda^t} \right) \mathcal{N}^{\frac{1}{2}}(\lambda) \exp\left(-\frac{3N\lambda^t}{64m\kappa^2} \right) \\ &+ \frac{1}{N\lambda^2} \left(1 + \frac{m}{N\lambda^t} \right) \mathcal{N}^{\frac{1}{2}}(\lambda) \exp\left(-\frac{3N\lambda^t}{64m\kappa^2} \right). \end{aligned} \tag{5.59}$$

Recall that $\{\mu_k\}_{k \ge 1}$ satisfy $\mu_k \lesssim k^{-1/p}$ for some $0 < p \le 1$.

When $\max\{0, t/2 - p/2\} \le \theta \le 1/2$, taking $m \le o\left(\frac{N^{\frac{2\theta+p-t}{2\theta+p}}}{\log N} \right)$ and $\lambda = N^{-\frac{1}{2\theta+p}}$ yields that for any $r > 0$, there holds

$$\limsup_{N \to \infty} N^r \exp\left(-\frac{3N\lambda^t}{64m\kappa^2} \right) = 0, \text{ as } \frac{m}{N\lambda^t} \le o\left(\frac{1}{\log N} \right).$$

Then combining with (5.9) and (5.59), we have

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0) \right) \right] \lesssim \lambda^{2\theta} + \frac{\lambda^{-p}}{N} + \frac{m\lambda^{2\theta-p}}{N} \lesssim N^{-\frac{2\theta}{2\theta+p}}$$

When $\theta < \max\{0, t/2 - p/2\}$ which implies $t > p > 0$, taking $m \leq o(\log N)$ and $\lambda = N^{-\frac{1}{t}}(\log N)^{-\frac{2}{t}}$ yields that for any $r > 0$, there holds

$$\limsup_{N\to\infty} N^r \exp\left(-\frac{3N\lambda^t}{64m\kappa^2}\right) = 0, \text{ as } \frac{m}{N\lambda^t} \leq o\left(\frac{1}{\log N}\right).$$

Then combining with (5.9) and (5.59), we have

$$\mathbb{E}\left[\left(\mathcal{R}(\overline{\beta}_{\mathcal{S},\lambda}) - \mathcal{R}(\beta_0)\right)\right] \lesssim \lambda^{2\theta} + \frac{\lambda^{-p}}{N} + \frac{m\lambda^{2\theta-p}}{N} \lesssim N^{-\frac{2\theta}{t}}(\log N)^{-\frac{4\theta}{t}}.$$

We have completed the proof of Theorem 9. ■

5.2 Mini-max Lower Rates

In this subsection, we establish the lower bound in Theorem 2. Before that, we present some crucial results used in our proof. Our analysis of lower bound bases on Fano's method, which provides lower bound in nonparametric estimation problem and was proposed by Khas (1979). Fano's method has been a crucial method in minimax lower bound estimation problem since it was proposed, and has inspired many following studies (see, e.g., Yang and Barron, 1999; Guntuboyina, 2011; Candes and Davenport, 2013). The following lemma is a direct application of Fano's method (see, for example, Yang and Barron, 1999). To this end, recall that the Kullback-Leibler divergence (KL-divergence) of two probability measures P, Q on a general space $(\Omega, \mathscr{F})$ is defined as

$$D_{kl}(P\|Q) := \int_{\Omega} \log\left(\frac{dP}{dQ}\right) dP,$$

if P is absolutely continuous with respect to Q, and otherwise $D_{kl}(P\|Q) := \infty$. Recall that for $\beta \in \mathcal{L}^2(\mathcal{T})$, $L_C^{1/2}\beta \in \mathrm{ran}T_*^\theta$ if β satisfied the regularity condition (3.1), i.e.,

$$L_C^{1/2}\beta = T_*^\theta(\gamma) \text{ with } 0 < \theta \leq 1/2 \text{ and some } \gamma \in \mathcal{L}^2(\mathcal{T}).$$

Lemma 22 *Suppose that there exist constants $r, R > 0$ and $\beta_1, \beta_2, \cdots, \beta_L \in \mathcal{L}^2(\mathcal{T})$ for some integer $L \geq 2$, such that*

$$L_C^{1/2}\beta_i \in \mathrm{ran}T_*^\theta,\ \left\|L_C^{1/2}(\beta_i - \beta_j)\right\|_{\mathcal{L}^2} \geq 2r \text{ and } D_{kl}(P_i\|P_j) \leq R,\ \forall 1 \leq i \neq j \leq L, \tag{5.60}$$

where P_i denotes the joint probability distribution of (X, Y) with

$$Y = \int_{\mathcal{T}} \beta_i(t)X(t)dt + \epsilon.$$

Here ϵ is independent of X satisfying $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] \leq \sigma^2$. Then we have

$$\inf_{\hat{\beta}_S} \sup_{\beta_0} \mathbb{P}\left\{\mathcal{R}(\hat{\beta}_S) - \mathcal{R}(\beta_0) \geq r^2\right\} \geq 1 - \frac{NR + \log 2}{\log L}, \tag{5.61}$$

where the supremum is taken over all $\beta_0 \in \mathcal{L}^2(\mathcal{T})$ satisfying $L_C^{1/2}\beta_0 \in \mathrm{ran} T_*^\theta$ and the infimum is taken over all possible predictors $\hat{\beta}_S \in \mathcal{L}^2(\mathcal{T})$ based on the training sample set $S = \{(X_i, Y_i)\}_{i=1}^N$ consisting of independent copies of (X, Y) with

$$Y = \int_{\mathcal{T}} \beta_0(t) X(t) dt + \epsilon.$$

In the followings, we first construct a family of $\{\beta_i\}_{i=1}^L$ satisfying (5.60) with suitable r, R and L, and then apply Lemma 22 to establish the lower bound. Note that establishing a lower bound for a particular instance directly provides a lower bound for the general scenario. Consequently, it is adequate to examine the situation where ϵ represents a Gaussian random variable with zero mean, characterized by $\mathbb{E}[\epsilon^2] = \sigma^2$. The following lemma is from the formulation of KL-divergence of two Gaussian distribution (see Duchi, 2016, Example 2.7), which can further facilitate the calculation.

Lemma 23 *Suppose that ϵ is a zero-mean Gaussian random variable independent of X satisfying $\mathbb{E}[\epsilon^2] = \sigma^2$ ¿0. For $\beta_i \in \mathcal{L}^2(\mathcal{T}), i = 1, 2$, let P_i denote the joint probability distribution of (X, Y) with*

$$Y = \int_{\mathcal{T}} \beta_i(t) X(t) dt + \epsilon.$$

Then

$$D_{kl}(P_1 \| P_2) = \frac{1}{2\sigma^2} \left\| L_C^{1/2}(\beta_1 - \beta_2) \right\|_{\mathcal{L}^2}^2. \tag{5.62}$$

Our construction of $\{\beta_i\}_{i=1}^L$ relies on the following lemma which is known as Gilbert-Varshamov bound (see Duchi, 2016, Lemma 7.5).

Lemma 24 *Let $M \geq 8$. There exists a subset $\Lambda \subset \mathcal{H}_M = \{-1, 1\}^M$ of size $|\Lambda| \geq \exp(M/8)$ such that*

$$\left\| \iota - \iota' \right\|_1 = 2 \sum_{i=1}^M \mathbb{I}_{\{\iota_j \neq \iota'_j\}} \geq M/2$$

for any $\iota \neq \iota'$ with $\iota, \iota' \in \Lambda$.

Now we are in the position to prove Theorem 2.

Proof of Theorem 2. Recall that the eigenvalues of T_* denoted by $\{\mu_k\}_{k \geq 1}$ are sorted in decreasing order with geometric multiplicities and satisfy $\mu_k \asymp k^{-1/p}$ for some $0 < p \leq 1$, which implies there exists $c > 0$ independent of j such that

$$\mu_{k+1} \leq \mu_k \text{ and } c k^{-1/p} \leq \mu_k \leq \frac{1}{c} k^{-1/p}, \quad \forall k \geq 1. \tag{5.63}$$

We only consider the case that ϵ is from the Gaussian distribution $N(0, \sigma^2)$ and independent of X, then the Assumption 2 is satisfied with $\sigma > 0$.

For $L \geq 2$, we construct $\{\beta_i\}_{i=1}^L$ according to Lemma 24. Take $M = \lceil a N^{\frac{p}{p+2\theta}} \rceil$, which denotes the smallest integer larger than $a N^{\frac{p}{p+2\theta}}$ for some constant $a > 8$ to be specified later. Let $\iota^{(1)}, \cdots, \iota^{(L)} \in \{-1,+1\}^M$ be given by Lemma 24 with $L \geq \exp(M/8)$. Given $0 < \theta \leq 1/2$, define

$$L_C^{1/2}\beta_i = \sum_{k=M+1}^{2M} \frac{1}{\sqrt{M}} \mu_k^\theta \iota_{k-M}^{(i)} \varphi_k = T_*^\theta(\gamma_i), \quad i = 1, \cdots, L, \tag{5.64}$$

where $\{\varphi_k\}_{k\geq 1}$ are the eigenvectors (corresponding to eigenvalue μ_k) of T_* which constitutes the orthonormal bases of $L^2(\mathcal{T})$, and $\gamma_i = \sum_{k=M+1}^{2M} \frac{1}{\sqrt{M}} \iota_{k-M}^{(i)} \varphi_k$ satisfies $\|\gamma_i\|_{\mathcal{L}^2}^2 = 1$. Then $L_C^{1/2}\beta_i \in \mathrm{ran} T_*^\theta$ with $0 < \theta \leq 1/2$ for $i = 1, \cdots, L$.

We next determine the positive constants r and R in (5.60) for $\{\beta_i\}_{i=1}^L$ defined above. For $1 \leq i, j \leq L$, we apply Lemma 5.60 and (5.63) to obtain

$$\begin{aligned} \left\| L_C^{1/2}(\beta_i - \beta_j) \right\|_{\mathcal{L}^2}^2 &= \sum_{k=M+1}^{2M} \frac{1}{M} \mu_k^{2\theta} \left(\iota_{k-M}^{(i)} - \iota_{k-M}^{(j)} \right)^2 \\ &\geq \mu_{2M}^{2\theta} \frac{4}{M} \sum_{k=M+1}^{2M} \mathbb{I}_{\{\iota_{k-M}^{(i)} \neq \iota_{k-M}^{(j)}\}} \\ &\geq \mu_{2M}^{2\theta} \frac{4}{M} \frac{M}{4} \geq c^{2\theta} 2^{-\frac{2\theta}{p}} M^{-\frac{2\theta}{p}}, \end{aligned}$$

where the last two inequalities are from (5.63). Therefore, we can take $r = \frac{1}{2}\sqrt{c^{2\theta} 2^{-\frac{2\theta}{p}} M^{-\frac{2\theta}{p}}}$. To determine R, we turn to bound $\mathcal{D}_{kl}(P_i \| P_j)$ where $\{P_i\}_{i=1}^L$ are the joint probability distributions of (X, Y) with $Y = \langle X, \beta_i \rangle_{\mathcal{L}^2} + \epsilon$ and $\epsilon \sim N(0, \sigma^2)$. Then, using lemma 23 and (5.64) yields

$$\begin{aligned} \mathcal{D}_{kl}(P_i \| P_j) &= \frac{1}{2\sigma^2} \left\| L_C^{1/2} (\beta_i - \beta_j) \right\|_{\mathcal{L}^2}^2 \\ &= \frac{1}{2\sigma^2} \sum_{k=M+1}^{2M} \frac{1}{M} \mu_k^{2\theta} (\iota_{k-M}^{(i)} - \iota_{k-M}^{(j)})^2 \\ &\leq \frac{2}{\sigma^2} \mu_M^{2\theta} \leq \frac{2}{\sigma^2 c^{2\theta}} M^{-\frac{2\theta}{p}}, \end{aligned}$$

where the last two inequalities are also due to (5.63). Thus, we can take $R = \frac{2}{\sigma^2 c^{2\theta}} M^{-\frac{2\theta}{p}}$.

Finally, let $r = \frac{1}{2}\sqrt{c^{2\theta}2^{-\frac{2\theta}{p}}M^{-\frac{2\theta}{p}}}$, $R = \frac{2}{\sigma^2 c^{2\theta}}M^{-\frac{2\theta}{p}}$ in Lemma 22 with $L \geq \exp(M/8)$ and $M = \lceil aN^{\frac{p}{p+2\theta}} \rceil$. Then there holds

$$\begin{aligned}
&\inf_{\hat{\beta}_S}\sup_{\beta_0}\mathbb{P}\left\{\mathcal{R}(\hat{\beta}_S) - \mathcal{R}(\beta_0) \geq \frac{c^{2\theta}}{4}2^{-\frac{2\theta}{p}}a^{-\frac{2\theta}{p}}N^{-\frac{2\theta}{p+2\theta}}\right\}\\
&\geq 1 - \frac{\frac{2N}{\sigma^2 c^{2\theta}}M^{-\frac{2\theta}{p}} + \log 2}{M/8} \geq 1 - \frac{\frac{2N}{\sigma^2 c^{2\theta}}M^{-\frac{2\theta}{p}} + \log 2}{M/8}\\
&= 1 - \frac{16}{\sigma^2 c^{2\theta}}NM^{-\frac{2\theta+p}{p}} - \frac{8\log 2}{M}\\
&\geq 1 - a^{-\frac{2\theta+p}{p}}\frac{16}{\sigma^2 c^{2\theta}}N^{1-\frac{p}{2\theta+p}\cdot\frac{2\theta+p}{p}} - \frac{8\log 2}{aN^{\frac{p}{p+2\theta}}}\\
&= 1 - a^{-\frac{2\theta+p}{p}}\frac{16}{\sigma^2 c^{2\theta}} - \frac{8\log 2}{a}N^{-\frac{p}{p+2\theta}}.
\end{aligned}$$

Therefore, we have

$$\inf_{N\to\infty}\lim\inf_{\hat{\beta}_S}\sup_{\beta_0}\mathbb{P}\left\{\mathcal{R}(\hat{\beta}_S) - \mathcal{R}(\beta_0) \geq \frac{c^{2\theta}}{4}2^{-\frac{2\theta}{p}}a^{-\frac{2\theta}{p}}N^{-\frac{2\theta}{p+2\theta}}\right\} = 1 - a^{-\frac{2\theta+p}{p}}\frac{16}{\sigma^2 c^{2\theta}}$$

and then

$$\lim_{a\to\infty}\inf_{N\to\infty}\lim\inf_{\hat{\beta}_S}\sup_{\beta_0}\mathbb{P}\left\{\mathcal{R}(\hat{\beta}_S) - \mathcal{R}(\beta_0) \geq \frac{c^{2\theta}}{4}2^{-\frac{2\theta}{p}}a^{-\frac{2\theta}{p}}N^{-\frac{2\theta}{p+2\theta}}\right\} = 1.$$

Taking $\gamma = \frac{c^{2\theta}}{4}2^{-\frac{2\theta}{p}}a^{-\frac{2\theta}{p}}$, we have

$$\lim_{\gamma\to 0}\inf_{N\to\infty}\lim\inf_{\hat{\beta}_S}\sup_{\beta_0}\mathbb{P}\left\{\mathcal{R}(\hat{\beta}_S) - \mathcal{R}(\beta_0) \geq \gamma N^{-\frac{2\theta}{2\theta+p}}\right\} = 1.$$

This completes the proof of Theorem 2. ∎

Acknowledgments

The work described in this paper is supported by the National Natural Science Foundation of China [Grants Nos.12171039 and 12061160462] and Shanghai Science and Technology Program [Project No. 21JC1400600].

Appendix A.

The lemma below provides Bernstein's inequality for the sum of self-adjoint random operators on a Hilbert space. The proof of this lemma can be found in Minsker (2017).

Lemma 25 *Consider a finite sequence $\{\eta_i\}_{i\geq 1}$ of independent random self-adjoint operators on a separable Hilbert space H. Assume that*

$$E[\eta_i] = 0 \quad and \quad ||\eta_i|| \leq L \quad for \quad each \quad i$$

Define the random operator $\overline{\eta} := \sum_{i\geq 1} \eta_i$. Suppose there are constant $v, d > 0$ such that $\left\|E[\overline{\eta}^2]\right\| \leq v$ and trace $\left(E[\overline{\eta}^2]\right) \leq vd$. Then for all $s \geq 0$,

$$\mathbb{P}(||\overline{\eta}|| \geq s) \leq \left[1 + 6\left(\frac{v}{s^2} + \frac{L}{3s}\right)^2\right] d \exp\left(-\frac{s^2}{2(v + Ls/3)}\right)$$

We next give the proof of Lemma 10.

Proof of Lemma 10. Using the Courant-Fischer mini-max principle Theorem (see Hsing and Eubank, 2015, Theorem 4.2.7), there holds

$$\begin{aligned}
\rho_k(L_A^{1/2} L_B L_A^{1/2}) &= \max_{v_1, \cdots, v_k \in \mathcal{H}} \min_{v \in span\{v_1, \cdots, v_k\}} \frac{\langle L_A^{1/2} L_B L_A^{1/2} v, v\rangle_{\mathcal{H}}}{\|v\|_{\mathcal{H}}^2} \\
&= \max_{v_1, \cdots, v_k \in \mathcal{H}} \min_{v \in span\{v_1, \cdots, v_k\}} \frac{\langle L_B L_A^{1/2} v, L_A^{1/2} v\rangle_{\mathcal{H}}}{\|L_A^{1/2} v\|_{\mathcal{H}}^2} \frac{\|L_A^{1/2} v\|_{\mathcal{H}}^2}{\|v\|_{\mathcal{H}}^2} \\
&\leq \max_{L_A^{1/2} v_1, \cdots, L_A^{1/2} v_k \in \mathcal{H}} \min_{L_A^{1/2} v \in span\{L_A^{1/2} v_1, \cdots, L_A^{1/2} v_k\}} \frac{\langle L_B L_A^{1/2} v, L_A^{1/2} v\rangle_{\mathcal{H}}}{\|L_A^{1/2} v\|_{\mathcal{H}}^2} \|L_A\| \\
&\overset{(\dagger)}{\leq} \max_{e_1, \cdots, e_k \in \mathcal{H}} \min_{e \in span\{e_1, \cdots, e_k\}} \frac{\langle L_B e, e\rangle_{\mathcal{H}}}{\|e\|_{\mathcal{H}}^2} \|L_A\| = \rho_k(L_B)\|L_A\|,
\end{aligned}$$

where $\{v_1, \cdots, v_k\}$ and $\{e_1, \cdots, e_k\}$ are two groups of k linearly independent elements in $\mathcal{H}$. Inequality ($\dagger$) uses the fact that $L_A^{1/2} v_1, \cdots, L_A^{1/2} v_k$ are linearly independent which is deduced from the assumption $\overline{\text{ran}(L_A^{1/2})} = H$. The proof of Lemma 10 is then finished. ■

References

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

R. B. Ash and M. F. Gardner. *Topics in Stochastic Processes: Probability and Mathematical Statistics: A Series of Monographs and Textbooks.* Academic press, Cambridge, Massachusetts, 2014.

F. Bach. Consistency of the group lasso and multiple kernel learning. *The Journal of Machine Learning Research*, (40):1179–1225, 2008.

F. Bauer, S. Pereverzev and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

R. Berthier, F. Bach and P. Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

T. T. Cai and P. Hall. Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179, 2006.

T. T. Cai and M. Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

E. J. Candes and M. A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

H. Cardot, F. Ferraty and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, 1999.

X. Chen, B. Tang, J. Fan and X. Guo. Online gradient descent algorithms for functional data learning. *Journal of Complexity*, 70:101635, 2022.

J. B. Conway. *A Course in Operator Theory.* American Mathematical Society, Providence, Rhode Island, 2000.

J. Duchi. Lecture notes for statistics 311/electrical engineering 377. *https://stanford.edu/class/stats311/Lectures/full_notes.pdf,* 2016.

F. Dumpert and A. Christmann. Universal consistency and robustness of localized support vector machines. *Neurocomputing*, 315:96–106, 2018.

J. Fan, F. Lv and L. Shi. An RKHS approach to estimate individualized treatment rules based on functional predictors. *Mathematical Foundations of Computing*, 2(2):169, 2019.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

A. Guntuboyina. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

K. Guo and J. Fan and L. Zhu. Optimal prediction for kernel-based semi-functional linear regression. *Analysis and Applications*, 22(3):467–505, 2024.

X. Guo, Z. C. Guo and L. Shi. Capacity dependent analysis for functional online learning algorithms. *Applied and Computational Harmonic Analysis*, 67:101567, 2023.

Z. C. Guo, S. B. Lin and D. X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

Z. C. Guo and L. Shi. Optimal rates for coefficient-based regularized regression. *Applied and Computational Harmonic Analysis*, 47(3):662–701, 2019.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.

T. Hamm and I. Steinwart. Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *The Annals of Statistics*, 49(6):3153–3180, 2021.

T. Hamm and I. Steinwart. Intrinsic dimension adaptive partitioning for kernel methods. *SIAM Journal on Mathematics of Data Science*, 4(2):721–749, 2022.

T. Hastie and C. Mallows. Statistical view of some chemometrics regression tools: Discussion. *Technometrics*, 35(2):140–143, 1993.

T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with An Introduction to Linear Operators*. John Wiley & Sons, Hoboken, New Jersey, 2015.

K. S. Jun, A. Cutkosky and F. Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. *Advances in Neural Information Processing Systems*, 32:1–10, 2019.

M. Kanagawa, P. Hennig, D. Sejdinovic and B. K. Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

R. Z. Khas'minskii. A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications*, 23(4):794–798, 1979.

H. Köhler and A. Christmann. Total stability of SVMs and localized SVMs. *The Journal of Machine Learning Research*, 23(1):4305–4345, 2022.

J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *The Journal of Machine Learning Research*, 21(1):5852–5914, 2020.

J. Lin, A. Rudi, L. Rosasco and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.

S. B. Lin, X. Guo and D. X. Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

Y. Mao, L. Shi and Z. C. Guo. Coefficient-based regularized distribution regression. *Journal of Approximation Theory*, 297:105995, 2024.

B. D. Marx. Generalized linear regression on sampled signals with penalized likelihood. *Proceedings of The 11th International Workshop on Statistical Modeling, Orvietto*, 1996.

D. Meunier, M. Pontil and C. Ciliberto. Distribution regression with sliced wasserstein kernels. *International Conference on Machine Learning*, 15501–15523, 2022.

S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

N. Mücke and G. Blanchard. Parallelizing spectrally regularized kernel algorithms. *The Journal of Machine Learning Research*, 19(1):1069–1097, 2018.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective.* MIT press, Cambridge, Massachusetts, 2012.

L. Pillaud-Vivien, A. Rudi and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31:1–11, 2018.

J. O. Ramsay and B. W. Silverman. *Fuctional Data Analysis, 2nd ed.* Springer, New York, 2005.

P. Rao. Some notes on misspecification in multiple regressions. *The American Statistician*, 25(5):37–39, 1971.

L. Shi. Distributed learning with indefinite kernels. *Analysis and Applications*, 17(06):947–975, 2019.

S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

H. Sun and Q. Wu. Optimal rates of distributed regression with imperfect kernels. *The Journal of Machine Learning Research*, 22(171):1–34, 2021.

Z. Szabó, B. K. Sriperumbudur, B. Póczos and A. Gretton. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

H. Tong. Distributed least squares prediction for functional linear regression. *Inverse Problems*, 38(2):025002, 2021.

J. L. Wang, J. M. Chiou and H. G. Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

F. Yao, H. G. Müller and J. L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6):2873–2903, 2005.

Y. Yao, L. Rosasco and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

M. Yuan and T. T. Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.

Y. Zhang, J. Duchi and M. Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.