Journal of Machine Learning Research 23 (2022) 1-94 Submitted 9/21; Revised 7/22; Published 7/22

Functional Linear Regression with Mixed Predictors

Daren Wang DWANG24@ND.EDU
Department of ACMS
University of Notre Dame
Indiana, USA

Zifeng Zhao ZIFENG.ZHAO@ND.EDU
Mendoza College of Business
University of Notre Dame
Indiana, USA

Yi Yu YI.YU.2@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, UK

Rebecca Willett WILLETT@UCHICAGO.EDU
Department of Statistics
University of Chicago
Illinois, USA

Editor: Pradeep Ravikumar

Abstract

We study a functional linear regression model that deals with functional responses and allows for both functional covariates and high-dimensional vector covariates. The proposed model is flexible and nests several functional regression models in the literature as special cases. Based on the theory of reproducing kernel Hilbert spaces (RKHS), we propose a penalized least squares estimator that can accommodate functional variables observed on discrete sample points. Besides a conventional smoothness penalty, a group Lasso-type penalty is further imposed to induce sparsity in the high-dimensional vector predictors. We derive finite sample theoretical guarantees and show that the excess prediction risk of our estimator is minimax optimal. Furthermore, our analysis reveals an interesting phase transition phenomenon that the optimal excess risk is determined jointly by the smoothness and the sparsity of the functional regression coefficients. A novel efficient optimization algorithm based on iterative coordinate descent is devised to handle the smoothness and group penalties simultaneously. Simulation studies and real data applications illustrate the promising performance of the proposed approach compared to the state-of-the-art methods in the literature.

Keywords: Reproducing kernel Hilbert space, functional data analysis, high-dimensional regression, minimax optimality.

1. Introduction

Functional data analysis is the collection of statistical tools and results revolving around the analysis of data in the form of (possibly discrete samples of) functions, images and more general objects. Recent technological advancement in various application areas, including neuroscience (e.g. Petersen et al., 2019; Dai et al., 2019), medicine (e.g. Chen et al., 2017; Ratcliffe et al., 2002), linguistic (e.g. Hadjipantelis et al., 2015; Tavakoli et al., 2019), finance (e.g. Fan et al., 2014; Benko, 2007), economics (e.g. Yan, 2007; Ramsay and Ramsey, 2002), transportation (e.g. Chiou et al., 2014; Wagner-Muns et al., 2017), climatology (e.g. Fraiman et al., 2014; Bonner et al., 2014), and others, has spurred an increase in the popularity of functional data analysis.

The statistical research in functional data analysis has covered a wide range of topics and areas. We refer the readers to a recent comprehensive review (Wang et al., 2016). In this paper, we are concerned with a general functional linear regression model that deals with functional responses and accommodates both functional and vector covariates.

By rescaling if necessary, without loss of generality, we assume the domain of the functional variables is $[0,1]$. Let $A^*(\cdot,\cdot) : [0,1]\times[0,1] \to \mathbb{R}$ be a bivariate coefficient function and $\{\beta_j^*(\cdot) : [0,1] \to \mathbb{R}\}_{j=1}^p$ be a collection of p univariate coefficient functions. The functional linear regression model concerned in this paper is as follows:

$$Y_t(r) = \int_{[0,1]} A^*(r,s)X_t(s)\,\mathrm{d}s + \sum_{j=1}^p \beta_j^*(r)Z_{tj} + \epsilon_t(r), \quad r \in [0,1], \tag{1}$$

where $Y_t(r) : [0,1] \to \mathbb{R}$ is the functional response, $X_t(s) : [0,1] \to \mathbb{R}$ is the functional covariate, $Z_t = (Z_{tj})_{j=1}^p \in \mathbb{R}^p$ is the vector covariate and $\epsilon_t(r) : [0,1] \to \mathbb{R}$ is the functional noise such that $\mathbb{E}(\epsilon_t(r)) = 0$ and $\mathrm{Var}(\epsilon_t(r)) < \infty$ for all $r \in [0,1]$. The index $t \in \{1,2,\dots,T\}$ denotes T independent and identically distributed samples. The vector dimension p is allowed to diverge as the sample size T grows unbounded. Furthermore, instead of assuming the functional variables are fully observed, we consider a more realistic setting where the functional covariates $\{X_t\}_{t=1}^T$ and responses $\{Y_t\}_{t=1}^T$ are only observed on discrete sample points $\{s_i\}_{i=1}^{n_1}$ and $\{r_j\}_{j=1}^{n_2}$, respectively. The two collections of sample points do not need to coincide.

As an important real-world example, consider a dataset collected from the popular crowdfunding platform kickstarter.com (see more details in Section 5.4). The website provides a platform for start-ups to create fundraising campaigns and charges a 5% service fee from the final fund raised by each campaign over its 30-day campaign duration. Denote $N_t(r)$ as the pledged fund for the campaign indexed by t at time r. Note that $\{N_t(r),\, r \in [0,30]\}$ forms a fundraising curve. For both the platform and the campaign creators, it is of vital interest to generate an accurate prediction of the future fundraising path $\{N_t(r), r \in (s,30]\}$ at an early time $s \in (0,30)$, as the knowledge of $\{N_t(r),\, r \in (s,30]\}$ helps the platform better assess its future revenue and further suggests timing along $(s,30]$ for potential intervention by the creators and the platform to boost the fundraising campaign and achieve better outcome.

At time s, to predict the functional response Y_t, i.e. $\{N_t(r),\, r \in (s,30]\}$, a functional regression as proposed in model (1) can be built based on the functional covariate X_t, i.e. $\{N_t(r),\, r \in [0,s]\}$, and vector covariates Z_t such as the number of creators and product updates of campaign t. Figure 1 plots the normalized fundraising curves of six representative campaigns and the functional predictions given by our proposed method (RKHS) and two competitors, FDA in Ramsay and Silverman (2005) and PFFR in Ivanescu et al. (2015).

In general, the proposed method achieves more favorable performance. More detailed real data analysis is presented in Section 5.4.

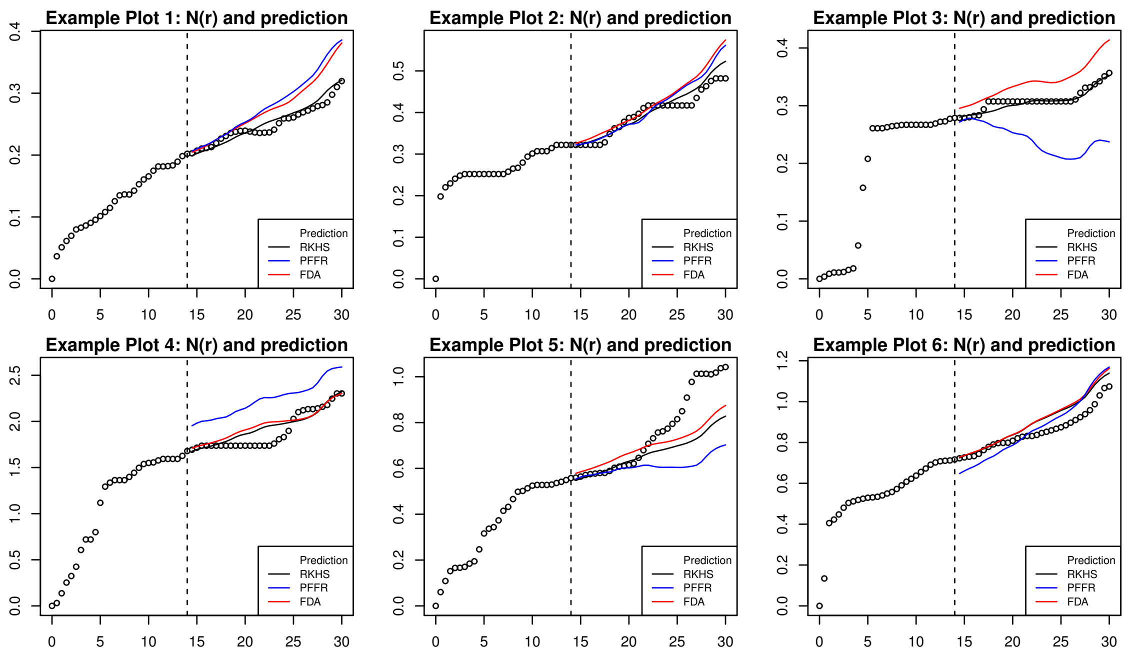

Figure 1: Observed (normalized) fundraising curves $\{N_t(r)\}$ (dots) of six representative campaigns and functional predictions (solid lines) given by RKHS, FDA and PFFR at $s = 14$th day (dashed vertical line) based on the functional covariate $\{N_t(r), r \in [0, s]\}$.

1.1 Literature review

As mentioned in recent reviews, e.g. Wang et al. (2016) and Morris (2015), there are in general three types of functional linear regression models: 1) functional covariates and scalar or vector responses (e.g. Cardot et al., 2003; Cai and Hall, 2006; Hall and Horowitz, 2007; Yuan and Cai, 2010; Raskutti et al., 2012; Cai and Yuan, 2012; Jiang et al., 2014; Fan et al., 2015); 2) functional covariates and functional responses (e.g. Wu et al., 1998; Liang et al., 2003; Yao et al., 2005; Fan et al., 2014; Ivanescu et al., 2015; Sun et al., 2018) and 3) vector covariates and functional responses (e.g. Laird and Ware, 1982; Li and Hsing, 2010; Faraway, 1997; Wu and Chiang, 2000). Another closely related area is functional time series. For instance, functional autoregressive models also preserve the regression format. The literature on functional time series is also abundant, including van Delft et al. (2017), Aue et al. (2015), Bathia et al. (2010), Wang et al. (2020a) and many others.

Regarding the sampling scheme, to establish theoretical guarantees, it is often assumed in the literature that functions are fully observed, which is generally untrue in practice. To study the properties of functional regression under the more realistic scenario where functions are observed only on discrete sample points, one usually imposes certain smoothness conditions on the underlying functions. The errors introduced by the discretization of functions can therefore be controlled as a function of the smoothness level and n, the number of discretized observations available for each function. Some fundamental tools regarding this aspect were established in Mendelson (2002) and Bartlett et al. (2005), and were further applied to various regression analysis problems (e.g. Raskutti et al., 2012; Koltchinskii and Yuan, 2010; Cai and Yuan, 2012).

Regarding the statistical inference task, in the regression context, estimation and prediction are two indispensable pillars. In the functional regression literature, Valencia and Yuan (2013), Cai and Yuan (2011), Yuan and Cai (2010), Lin and Yuan (2006), Park et al. (2018), Fan et al. (2014), Fan et al. (2015), and Reimherr et al. (2019), among many others, have studied different aspects of the estimation problem. As for prediction, the existing literature includes Cai and Yuan (2012), Cai and Hall (2006), Ferraty and Vieu (2009), Sun et al. (2018), Reimherr et al. (2018) to name but a few.

In contrast to the aforementioned three paradigms of functional regression, in this paper, we study the functional linear regression problem in model (1) with functional responses and mixed predictors that consist of both functional and vector covariates. We assume the functions are only observed on discrete sample points and aim to derive optimal upper bounds on the excess prediction risk (see Theorem 2). More detailed comparisons with existing literature are deferred till we present the main results in Section 3.

1.2 Main contributions

The main contributions of this paper are summarized as follows.

Firstly, to the best of our knowledge, the model we study in this paper, as defined in (1), is among the most flexible ones in the literature of functional linear regression. In terms of predictors, we allow for both functional and vector covariates. In terms of model dimensionality, we allow for the dimension p of vector covariates to grow exponentially as the sample size T diverges. In terms of function spaces, which will be elaborated later, we allow for the coefficients of the functional and vector covariates to be from different reproducing kernel Hilbert spaces (RKHS). In terms of dependence between the functional and vector covariates, we allow the correlation to be of order up to $O(1)$, which is a rather weak condition, especially considering that the vector covariates are of high dimensions. In terms of the sampling scheme, we allow for the functional covariate and response to be observed on (possibly different) discretized sample points. This general and unified framework imposes new challenges on both theory and optimization, as we elaborate in detail later.

Secondly, we develop new peeling techniques in our theoretical analysis, which is crucial and fundamental in dealing with the potentially exponentially growing vector covariate dimension. Existing asymptotic analysis techniques and results in the functional regression literature are insufficient to provide Lasso-type guarantees with the presence of high-dimensional covariates. See Theorem 5 for more details.

Thirdly, we demonstrate an interesting phase transition phenomenon in terms of the smoothness of the functional covariate and the dimensionality of the vector covariate. To be specific, let $\mathfrak{s}$ be the sparsity of $\{\beta_j^*\}_{j=1}^p$, i.e. the number of non-zero univariate coefficient functions. Let δ_T be a quantity jointly determined by the complexities and the alignment of the two RKHS's where the functional covariates $\{X_t\}_{t=1}^T$ and the bivariate coefficient function A^* reside; see Theorem 4 for the detailed definition of δ_T. We show that the excess prediction risk of our proposed estimator is upper bounded by

$$O_{\mathrm{p}}\left\{\mathfrak{s}\log(p \vee T)T^{-1} + \delta_T\right\} + \text{discretization error.}$$

- When $\delta_T \lesssim \mathfrak{s}\log(p \vee T)T^{-1}$, the excess risk is dominated by $\mathfrak{s}\log(p \vee T)T^{-1}$, which is the standard excess risk rate in the high-dimensional parametric statistics literature (e.g. Bühlmann and van de Geer, 2011). Therefore, the difficulty is dominated by estimating $\{\beta_j^*\}_{j=1}^p$, the p univariate coefficient functions.

- When $\delta_T \gtrsim \mathfrak{s}\log(p \vee T)T^{-1}$, the excess risk is dominated by δ_T, which suggests that the difficulty is dominated by estimating the bivariate coefficient function A^*. We show that in this regime, the optimal excess risk of model (1) is of order δ_T. As a result, in this regime the difficulty is dominated by estimating the bivariate coefficient function A^*.

- We further develop matching lower bounds to justify that this phase transition between high-dimensional parametric rate and the non-parametric rate is indeed minimax optimal; see Section 3.3. To the best of our knowledge, this is the first time such phenomenon is observed in the functional regression literature.

Note that, this phase transition and its associated optimality are not with respect to the discretization error, i.e. n - the minimum number of observations obtained for each function.

Lastly, we derive a representer theorem and further propose a novel optimization algorithm via iterative coordinate descent, which efficiently solves a sophisticated penalized least squares problem with the presence of both a ridge-type smoothing penalty and a group Lasso-type sparsity penalty.

The rest of the paper is organized as follows. In Section 2, we introduce relevant definitions and quantities. The main theoretical results are presented in Section 3. The optimization procedure is discussed in Section 4. Numerical experiments and real data applications are conducted in Section 5 to illustrate the favorable performance of the proposed estimator over existing approaches in the literature. Section 6 concludes with a discussion.

Notation

Let $\mathbb{N}_*$ denote the collection of positive integers. Let $a(n)$ and $b(n)$ be two quantities depending on n. We say that $a(n) \lesssim b(n)$ if there exists $n_0 \in \mathbb{N}_*$ and an absolute constant $C > 0$ such that for any $n \geq n_0$, $a(n) \leq Cb(n)$. We denote $a(n) \asymp b(n)$, if $a(n) \lesssim b(n)$ and $b(n) \lesssim a(n)$. For any function $g : [0,1] \to \mathbb{R}$, denote

$$\|g\|_{\mathcal{L}^2} = \sqrt{\int_{[0,1]} g^2(r)\,\mathrm{d}r} \quad \text{and} \quad \|g\|_\infty = \sup_{r\in[0,1]} |g(r)|.$$

For any $p \in \mathbb{N}_*$ and $a, b \in \mathbb{R}^p$, let $\langle a, b\rangle_p = \sum_{i=1}^p a_i b_i$.

Let $\mathcal{L}^2 = \mathcal{L}^2([0,1])$ be the space of all square integrable functions with respect to the uniform distribution on $[0,1]$, i.e. $\mathcal{L}^2 = \{f : [0,1] \to \mathbb{R}, \|f\|_{\mathcal{L}^2} < \infty\}$. For any $\alpha > 0$, let $W^{\alpha,2}$ be the Sobolev space of order α. For a positive integer α, we have

$$W^{\alpha,2} = \{f \in \mathcal{L}^2[0,1],\ f^{(\alpha-1)} \text{ is absolutely continuous and } \|f^{(\alpha)}\|_{\mathcal{L}^2} < \infty\},$$

where $f^{(k)}$ is the k-th weak derivative of f, $k \in \mathbb{N}$. In this case the Sobolev norm of f is defined as

$$\|f\|^2_{W^{\alpha,2}} = \|f\|^2_{\mathcal{L}^2} + \|f^{(\alpha)}\|^2_{\mathcal{L}^2}.$$

For non-integer valued $\alpha > 0$, the Sobolev space $W^{\alpha,2}$ and the norm $\|\cdot\|_{W^{\alpha,2}}$ are also well-defined (e.g. see Brezis, 2010). Note that if $0 < \alpha_1 < \alpha_2$, then $W^{\alpha_2,2} \subset W^{\alpha_1,2}$.

2. Background

In this section, we provide some background on the fundamental tools used in this paper, with RKHS and compact linear operators studied in Section 2.1 and Appendix A.1, respectively.

2.1 Reproducing kernel Hilbert spaces (RKHS)

Consider a Hilbert space $\mathcal{H} \subset \mathcal{L}^2$ and its associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, under which $\mathcal{H}$ is complete. We assume that there exists a continuous symmetric nonnegative-definite kernel function $\mathbb{K} : [0,1] \times [0,1] \to \mathbb{R}_+$ such that the space $\mathcal{H}$ is an RKHS, in the sense that for each $r \in [0,1]$, the function $\mathbb{K}(\cdot, r) \in \mathcal{H}$ and $g(r) = \langle g(\cdot), \mathbb{K}(\cdot, r)\rangle_{\mathcal{H}}$, for all $g \in \mathcal{H}$. To emphasize this relationship, we write $\mathcal{H}(\mathbb{K})$ as an RKHS with $\mathbb{K}$ as the associated kernel. For any $\mathbb{K}$, we say it is a bounded kernel if $\sup_{r\in[0,1]} \mathbb{K}(r,r) < \infty$.

It follows from Mercer's theorem (Mercer, 1909) that there exists an orthonormal basis of $\mathcal{L}^2$, $\{\phi_k\}_{k=1}^\infty \subset \mathcal{L}^2$, such that a non-negative definite kernel function $\mathbb{K}(\cdot,\cdot)$ has the representation $\mathbb{K}(s,r) = \sum_{k=1}^\infty \mu_k \phi_k(s)\phi_k(r)$, $s, r \in [0,1]$, where $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ are the eigenvalues of $\mathbb{K}$ and $\{\phi_k\}_{k=1}^\infty$ are the corresponding eigenfunctions. In the rest of this paper, when there is no ambiguity, we drop the dependence of μ_k's and ϕ_k's on their associated kernel function $\mathbb{K}$ for ease of notation.

Note that any function $f \in \mathcal{H}(\mathbb{K})$ can be written as

$$f(r) = \sum_{k=1}^\infty \left\{ \int_{[0,1]} f(r)\phi_k(r)\,\mathrm{d}r \right\} \phi_k(r) = \sum_{k=1}^\infty a_k \phi_k(r), \quad r \in [0,1],$$

and its RKHS norm is defined as $\|f\|_{\mathcal{H}(\mathbb{K})} = \sqrt{\sum_{k=1}^\infty a_k^2/\mu_k}$. Thus, for the eigenfunctions, we have $\|\phi_k\|^2_{\mathcal{H}(\mathbb{K})} = \mu_k^{-1}$. Throughout this section, we further denote $\psi_k = \sqrt{\mu_k}\phi_k$ and note that $\|\psi_k\|_{\mathcal{H}(\mathbb{K})} = 1$ for $k \in \mathbb{N}_*$.

Define the linear map $L_\mathbb{K} : \mathcal{L}^2 \to \mathcal{L}^2$, associated with $\mathbb{K}$, as $L_\mathbb{K}(f)(\cdot) = \int_{[0,1]} \mathbb{K}(\cdot, r)f(r)\,\mathrm{d}r$. It holds that $L_\mathbb{K}(\phi_k) = \mu_k\phi_k$, for $k \in \mathbb{N}_*$. Furthermore, define $L_{\mathbb{K}^{1/2}} : \mathcal{L}^2 \to \mathcal{H}(\mathbb{K})$ such that $L_{\mathbb{K}^{1/2}}(\phi_k) = \sqrt{\mu_k}\phi_k$, and define $L_{\mathbb{K}^{-1/2}} : \mathcal{H}(\mathbb{K}) \to \mathcal{L}^2$ such that $L_{\mathbb{K}^{-1/2}}(\phi_k) = \mu_k^{-1/2}\phi_k$, for $k \in \mathbb{N}_*$. For any two bivariate functions $R_1(\cdot,\cdot), R_2(\cdot,\cdot) : [0,1] \times [0,1] \to \mathbb{R}$, define $R_1R_2(r,s) = \int_{[0,1]} R_1(r,u)R_2(u,s)\,\mathrm{d}u$, $r, s \in [0,1]$. It holds that $L_{R_1R_2} = L_{R_1} \circ L_{R_2}$, where $L_{R_1} \circ L_{R_2}$ represents the composition of R_1 and R_2, which is a linear map from $\mathcal{L}^2$ to $\mathcal{L}^2$ given that both L_{R_1} and L_{R_2} are linear maps from $\mathcal{L}^2$ to $\mathcal{L}^2$.

2.2 Bivariate functions and compact linear operators

To regulate the bivariate coefficient function A^* in model (1), we consider a class of compact linear operators in $\mathcal{H}(\mathbb{K})$. Specifically, for any compact linear operator $A_2 : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$, denote

$$A_2[f, g] = \langle A_2[g], f\rangle_{\mathcal{H}(\mathbb{K})}, \quad f, g \in \mathcal{H}(\mathbb{K}).$$

Note that $A_2[f, g]$ is well defined for any $f, g \in \mathcal{H}(\mathbb{K})$ due to the compactness of A_2. Let $\{\psi_i\}_{i=1}^{\infty}$ be the eigenbasis of $\mathcal{H}(\mathbb{K})$ and $a_{ij} = A_2[\psi_i, \psi_j] = \langle A_2[\psi_j], \psi_i\rangle_{\mathcal{H}(\mathbb{K})}$, $i, j \in \mathbb{N}_*$. We thus have for any $f, g \in \mathcal{H}(\mathbb{K})$, it follows from Theorem 9 in the appendix that

$$A_2[f, g] = \sum_{i,j=1}^{\infty} a_{ij}\langle f, \psi_i\rangle_{\mathcal{H}(\mathbb{K})}\langle g, \psi_j\rangle_{\mathcal{H}(\mathbb{K})}. \tag{2}$$

Note that via (2), we can define a bivariate function $A_1(r, s)$, $r, s \in [0, 1]$, affiliated with the compact operator A_2. Specifically, plugging $f = \mathbb{K}(r, \cdot)$ and $g = \mathbb{K}(s, \cdot)$ into (2), we have that

$$A_2[\mathbb{K}(r, \cdot), \mathbb{K}(s, \cdot)] = \sum_{i,j=1}^{\infty} a_{ij}\langle \mathbb{K}(r, \cdot), \psi_i\rangle_{\mathcal{H}(\mathbb{K})}\langle \mathbb{K}(s, \cdot), \psi_j\rangle_{\mathcal{H}(\mathbb{K})} = \sum_{i,j=1}^{\infty} a_{ij}\psi_i(r)\psi_j(s) = A_1(r, s).$$

From the above set up, it holds that for all $v \in \mathcal{H}(\mathbb{K})$, we have

$$A_2[v](r) = \langle A_1(r, \cdot), v(\cdot)\rangle_{\mathcal{H}(\mathbb{K})}, \quad r \in [0, 1].$$

We have established an equivalence between a compact linear operator A_2 and its corresponding bivariate function $A_1(r, s)$, therefore any compact linear operator $A_2 : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$ can be viewed as a bivariate function $A_1 : [0, 1] \times [0, 1] \to \mathbb{R}$.

To facilitate the later formulation of penalized convex optimization and the derivation of the representer theorem, we further focus on the Hilbert–Schmidt operator, which is an important subclass of compact operators. A compact operator A_2 is Hilbert–Schmidt if

$$\|A_2\|^2_{\mathrm{F}(\mathbb{K})} = \sum_{i,j=1}^{\infty} \langle A_2[\psi_j], \psi_i\rangle^2_{\mathcal{H}(\mathbb{K})} = \sum_{i,j=1}^{\infty} a_{ij}^2 < \infty. \tag{3}$$

In the rest of this paper, for ease of presentation, we adopt some abuse of notation and refer to compact linear operators and the corresponding bivariate functions by the same notation A.

3. Main results

3.1 The constrained/penalized estimator and the representer theorem

Recall that in model (1), the response is the function Y, the covariates include the function X and vector Z, and the unknown parameters are the bivariate coefficient function A^* and univariate coefficient functions $\beta^* = \{\beta_l^*\}_{l=1}^p$. Given the observations $\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=1,i=1,j=1}^{T,n_1,n_2}$, our main task is to estimate A^* and β^*.

Define the weight functions $w_s(i) = (n_1+1)(s_i - s_{i-1})$ for $1 \le i \le n_1$ and $w_r(j) = (n_2+1)(r_j - r_{j-1})$ for $1 \le j \le n_2$, where by convention we set $s_0 = r_0 = 0$. In addition, for any $1 \le l \le p$, define

$$\|\beta_l\|_{n_2} = \sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2} w_r(j)\beta_l^2(r_j)}.$$

We propose the following constrained/penalized least squares estimator

$$(\widehat{A}, \widehat{\beta}) = \mathop{\arg\min}_{\substack{A \in \mathcal{C}_A, \\ \beta = \{\beta_l\}_{l=1}^p \in \mathcal{C}_\beta}} \left[\frac{1}{Tn_2}\sum_{t=1}^{T}\sum_{j=1}^{n_2} w_r(j)\left\{Y_t(r_j) - \frac{1}{n_1}\sum_{i=1}^{n_1} w_s(i)A(r_j, s_i)X_t(s_i) - \langle \beta(r_j), Z_t\rangle_p\right\}^2 + \lambda \sum_{l=1}^{p}\|\beta_l\|_{n_2}\right], \tag{4}$$

where $\lambda > 0$ is a tuning parameter that controls the group Lasso penalty, $\mathcal{C}_A$ and $\mathcal{C}_\beta$ characterize the spaces of coefficient functions A and β respectively such that

$$\mathcal{C}_A = \{A : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K}),\ \|A\|^2_{\mathrm{F}(\mathbb{K})} \le C_A\},\ \mathcal{C}_\beta = \left\{\{\beta_l\}_{l=1}^p \subset \mathcal{H}(\mathbb{K}_\beta) : \sum_{l=1}^{p}\|\beta_l\|_{\mathcal{H}(\mathbb{K}_\beta)} \le C_\beta\right\}. \tag{5}$$

Here $C_A, C_\beta > 0$ are two absolute constants and $\|\cdot\|_{\mathrm{F}(\mathbb{K})}$ is defined in (3). Note that we allow the RKHS's $\mathcal{H}(\mathbb{K})$ and $\mathcal{H}(\mathbb{K}_\beta)$ to be generated by two different kernels $\mathbb{K}$ and $\mathbb{K}_\beta$. The detailed optimization scheme is deferred to Section 4.

The above optimization problem essentially constrains three tuning parameters. The tuning parameters C_A and C_β control the smoothness of the regression coefficient functions. These constrains are standard and ubiquitous in the functional data analysis and non-parametric statistical literature. The tuning parameter λ controls the group Lasso penalty, which is used to encourage sparsity when the dimensionality p diverges faster than the sample size T.

The optimization problem in (4) makes use of the weight functions $w_s(i)$ and $w_r(j)$, which are determined by the discrete sample points $\{s_i\}_{i=1}^{n_1}$ and $\{r_j\}_{j=1}^{n_2}$ respectively. This is designed to handle the scenario where the functional variables are observed on unevenly spaced sample points. Indeed, for evenly spaced sample points $\{s_i\}_{i=1}^{n_1}$ and $\{r_j\}_{j=1}^{n_2}$, we have that $w_s(i) = w_r(j) = 1$ for all i, j. We remark that it is well known that the integral of a regular function can be well approximated by its weighted sum evaluated at discrete sample points.

Note that (4) is an infinite dimensional optimization problem. Fortunately, Theorem 1 states that the estimator $(\widehat{A}, \widehat{\beta})$ can in fact be written as linear combinations of their corresponding kernel functions evaluated at the discrete sample points $\{s_i\}_{i=1}^{n_1}$ and $\{r_j\}_{j=1}^{n_2}$.

Proposition 1 *Denote $\mathbb{K}$ and $\mathbb{K}_\beta$ as the RKHS kernels of $\mathcal{H}(\mathbb{K})$ and $\mathcal{H}(\mathbb{K}_\beta)$ respectively. There always exists a minimizer $(\widehat{A}, \widehat{\beta})$ of (4) such that,*

$$\widehat{A}(r, s) = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \widehat{a}_{ij}\mathbb{K}(r, r_j)\mathbb{K}(s, s_i), \quad (r, s) \in [0,1]\times[0,1],\ \{\widehat{a}_{ij}\}_{i,j=1}^{n_1,n_2} \subset \mathbb{R},$$

and

$$\widehat{\beta}_l(r) = \sum_{j=1}^{n_2} \widehat{b}_{lj} \mathbb{K}_\beta(r_j, r), \quad r \in [0,1],\, l \in \{1,\dots,p\},\, \{\widehat{b}_{lj}\}_{l=1,j=1}^{p,n_2} \subset \mathbb{R}.$$

Theorem 1 is a generalization of the well-known representer theorem for RKHS (Wahba, 1990). Various versions of representer theorems are derived and used in the functional data analysis literature (e.g. Yuan and Cai, 2010).

3.2 Model assumptions

To establish the optimal theoretical guarantees for the estimator $(\widehat{A}, \widehat{\beta})$, we impose some mild model assumptions, on the coefficient functions (Assumption 1), functional and vector covariates (Assumption 2) and sampling scheme (Assumption 3).

Assumption 1 (Coefficient functions)

(a) The bivariate coefficient function A^* belongs to $\mathcal{C}_A$ defined in (5)*, where $\mathcal{H}(\mathbb{K})$ is the RKHS associated with a bounded kernel $\mathbb{K}$ and $C_A > 0$ is an absolute constant.*

(b) The univariate coefficient functions $\{\beta_l^*\}_{l=1}^p$ belong to $\mathcal{C}_\beta$ defined in (5)*, where $\mathcal{H}(\mathbb{K}_\beta)$ is the RKHS associated with a bounded kernel $\mathbb{K}_\beta$ and $C_\beta > 0$ is an absolute constant.*

In addition, there exists a set $S \subset \{1,\dots,p\}$ such that $\beta_j^* = 0$ for all $j \in \{1,\dots,p\} \setminus S$ and there exists a sufficiently large absolute constant $C_{\mathrm{snr}} > 0$ such that

$$T \geq C_{\mathrm{snr}} \mathfrak{s} \log(p \vee T), \tag{6}$$

where $\mathfrak{s} = |S|$ denotes the cardinality of the set S.

Assumption 1(a) requires the bivariate coefficient function A^* to be a Hilbert–Schmidt operator mapping from $\mathcal{H}(\mathbb{K})$ to $\mathcal{H}(\mathbb{K})$. Assumption 1(b) imposes smoothness and sparsity conditions on the univariate coefficient functions $\{\beta_l^*\}_{l=1}^p$ to handle the potential high-dimensionality of the vector covariate. Note that we allow $\{\beta_l^*\}_{l=1}^p$ to be from a possibly different RKHS $\mathcal{H}(\mathbb{K}_\beta)$ than $\mathcal{H}(\mathbb{K})$ and thus would allow users to choose kernels based on their practical needs.

Assumption 2 (Covariates)

(a) The functional noise $\{\epsilon_t\}_{t=1}^T$ is a collection of independent and identically distributed Gaussian processes such that $\mathbb{E}(\epsilon_1(r)) = 0$ and $\mathrm{Var}(\epsilon_1(r)) < C_\epsilon$ for all $r \in [0,1]$. In addition, $\{\epsilon_t\}_{t=1}^T$ are independent of $\{X_t, Z_t\}_{t=1}^T$.

(b) The functional covariate $\{X_t\}_{t=1}^T$ is a collection of independent and identically distributed centered Gaussian processes with covariance operator Σ_X and $\mathbb{E}\big(\|X_1\|_{\mathcal{L}^2}^2\big) = C_X < \infty$.

(c) The vector covariate $\{Z_t\}_{t=1}^T \subset \mathbb{R}^p$ is a collection of independent and identically distributed Gaussian random vectors from $\mathcal{N}(0, \Sigma_Z)$, where $\Sigma_Z \in \mathbb{R}^{p \times p}$ is a positive definite matrix such that

$$c_z \|v\|_2^2 \leq v^\top \Sigma_Z v \leq C_z \|v\|_2^2, \quad v \in \mathbb{R}^p,$$

and $c_z, C_Z > 0$ are absolute constants.

(d) *For any deterministic $f \in \mathcal{H}(\mathbb{K})$ and deterministic $v \in \mathbb{R}^p$, it holds that*

$$\mathbb{E}(\langle X_1, f\rangle_{\mathcal{L}^2} Z_1^\top v) \leq \frac{3}{4}\sqrt{\mathbb{E}\{\langle X_1, f\rangle_{\mathcal{L}^2}^2\}(v^\top \Sigma_Z v)} = \frac{3}{4}\sqrt{\Sigma_X[f,f](v^\top \Sigma_Z v)}.$$

Assumption 2(a) and (b) state that both the functional noise $\{\epsilon_t\}_{t=1}^T$ and functional covariates $\{X_t\}_{t=1}^T$ are Gaussian processes. In fact, one could further relax them to be sub-Gaussian processes. Such assumptions are frequently used in high-dimensional functional literature such as Kneip et al. (2016) and Wang et al. (2020b). Assumption 2(d) allows that the functional and vector covariates to be correlated up to 3/4, which means that the correlation, despite the functional and high-dimensional nature of the problem, can be of order $O(1)$. We do not claim the optimality of the constant 3/4 but emphasize that this correlation cannot be equal to one, as detailed in Appendix A.2.

Assumptions 1 and 2 are sufficient for establishing theoretical guarantees if we require all functions (i.e. the functional responses and functional covariates) to be fully observed, which is typically not realistic in practice. To allow for discretized observations, we further introduce assumptions on the sampling scheme and on the smoothness of the functional covariates.

Assumption 3 (Sampling scheme)

(a) *The discrete sample points $\{s_i\}_{i=1}^{n_1}$ and $\{r_j\}_{j=1}^{n_2}$ are collections of points with $0 \leq s_1 < s_2 \ldots < s_{n_1} = 1$ and $0 \leq r_1 < r_2 \ldots < r_{n_2} = 1$, and there exists an absolute constant C_d such that*

$$s_i - s_{i-1} \leq \frac{C_d}{n_1} \quad \text{for all } 1 \leq i \leq n_1 + 1 \quad \text{and} \quad r_j - r_{j-1} \leq \frac{C_d}{n_2} \quad \text{for all } 1 \leq j \leq n_2 + 1,$$

where by convention we set $s_0 = r_0 = 0$.

(b) *Suppose that $\mathcal{H}(\mathbb{K}), \mathcal{H}(\mathbb{K}_\beta) \subset W^{\alpha,2}$ for some $\alpha > 1/2$. In addition, suppose that*

$$\mathbb{E}(\|X_1\|_{W^{\alpha,2}}^2) < \infty. \tag{7}$$

Assumption 3(a) allows the functional variables to be partially observed on discrete sample points. Importantly, Assumption 3(a) can accommodate both fixed and random sampling schemes. In particular, suppose the sample points are randomly generated from an unknown distribution on $[0,1]$ with a density function $\mu: [0,1] \to \mathbb{R}$ such that $\inf_{r\in[0,1]} \mu(r) > 0$, we have Assumption 3(a) holds with high probability. We refer to Theorem 1 of Wang et al. (2014) for more details.

To handle the partially observed functional variables, Assumption 3(b) imposes the smoothness assumption that X, A^* and $\{\beta_j^*\}_{j=1}^p$ can be enclosed by a common superset, the Sobolev space $W^{\alpha,2}$. In particular, (7) is a commonly used assumption for bivariate function estimation in the functional data analysis literature. See for example, Cai and Yuan (2010) and Wang et al. (2020b) and references therein. We emphasize that Assumption 3(b) indeed allows different smoothness levels for X, A^* and $\{\beta_j^*\}_{j=1}^p$, and only requires that the least smooth space among X, A^* and $\{\beta_j^*\}_{j=1}^p$ is covered by $W^{\alpha,2}$.

The condition (7) requires that the second moment of $\|X\|_{W^{\alpha,2}}$ is finite, which implies $X \in W^{\alpha,2}$ almost surely. Due to the fact that the functional covariates $\{X_t\}_{t=1}^T$ are partially observed, to derive finite-sample guarantees, we need to establish uniform control over the approximation error

$$\left| \int_{[0,1]} A(r,s)X_t(s)\,\mathrm{d}s - \frac{1}{n_1}\sum_{i=1}^{n_1} w_s(i)A(r,s_i)X_t(s_i) \right| \quad \text{for all } t \in \{1,\ldots,T\}.$$

This requires the realized sample paths $\{X_t\}_{t=1}^T$ to be regular. In particular, the second moment condition in (7) can be used to show that $\{\|X_t\|_{W^{\alpha,2}}\}_{t=1}^T$ are bounded with high probability, which implies $\{X_t\}_{t=1}^T$ are Hölder smooth with Hölder parameter $\alpha - 1/2 > 0$ by the Morrey inequality (see Theorem 39 in Appendix E.1 for more details). In Example 1 in Appendix A, we further provide a concrete example to illustrate a sufficient condition for (7).

3.3 Theoretical guarantees

With the assumptions in hand, we establish theoretical guarantees and investigate the minimax optimality of the proposed estimator (4), via the lens of excess risk defined in Theorem 2. The derivation of (8) is collected in Theorem 37.

Definition 2 (Excess prediction risk) *Let $(\widehat{A},\widehat{\beta})$ be any estimator of (A^*,β^*) in model* (1). *The excess risk of $(\widehat{A},\widehat{\beta})$ is defined as*

$$\begin{aligned}
\mathcal{E}^*(\widehat{A},\widehat{\beta}) =& \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(Y^*(r) - \int_{[0,1]}\widehat{A}(r,s)X^*(s)\,\mathrm{d}s - \langle Z^*, \widehat{\beta}(r)\rangle_p\right)^2 \mathrm{d}r\right\} \\
&- \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(Y^*(r) - \int_{[0,1]}A^*(r,s)X^*(s)\,\mathrm{d}s - \langle Z^*, \beta^*(r)\rangle_p\right)^2 \mathrm{d}r\right\} \\
=& \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s + \langle Z^*, \Delta_\beta(r)\rangle_p\right)^2 \mathrm{d}r\right\}, \qquad (8)
\end{aligned}$$

where

- $\Delta_A(r,s) = \widehat{A}(r,s) - A^*(r,s)$ *and* $\Delta_\beta(r) = \widehat{\beta}(r) - \beta^*(r)$ *for* $s,r \in [0,1]$;
- *the random objects* (X^*,Z^*,Y^*) *are independent from and identically distributed as the observed data generated from model* (1).

Remark 3 *Definition 2 is ubiquitously used to measure prediction accuracy in regression settings. Excess risks are usually considered to provide better quantification of prediction accuracy for the estimators of interest than the ℓ_2 error bounds. Consequently, throughout our paper, we evaluate our proposed estimators using excess risks. We remark that as pointed out by Cai and Hall (2006) and Cai and Yuan (2012), much of the practical interest in the functional regression parameters is centered around applications for the purpose of prediction.*

We are now ready to present our main results, which provide upper and lower bounds on the excess risk $\mathcal{E}^*(\widehat{A},\widehat{\beta})$ of the proposed estimator $(\widehat{A},\widehat{\beta})$ defined in (4). For notational simplicity, in the following, we assume without loss of generality that $n_1 = n_2 = n$. For $n_1 \neq n_2$, all the statistical guarantees continue to hold by setting $n = \min\{n_1, n_2\}$.

Upper bounds

Theorem 4 *Under Assumptions 1, 2 and 3, suppose that the eigenvalues* $\{\xi_k\}_{k=1}^{\infty}$ *of the linear operator* $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ *satisfy*

$$\xi_k \asymp k^{-2r}, \tag{9}$$

for some $r > 1/2$. *Let* $(\widehat{A},\widehat{\beta})$ *be any solution to* (4) *with the tuning parameter* $\lambda = C_\lambda\sqrt{\log(p\vee T)T^{-1}}$ *for some sufficiently large constant* C_λ. *For any* $T \gtrsim \log(n)$, *there exists an absolute constant* $C > 0$ *such that with probability at least* $1 - 8T^{-4}$, *it holds that*

$$\mathcal{E}^*(\widehat{A},\widehat{\beta}) \le C\log(T)\big\{\delta_T + \mathfrak{s}\log(p\vee T)T^{-1} + \zeta_n\big\}, \tag{10}$$

where $\zeta_n = n^{-\alpha+1/2}$, $\mathfrak{s}$ *is the sparsity parameter defined in Assumption 1(***b***) and* $\delta_T = T^{-2r/(2r+1)}$.

Theorem 4 provides a high-probability upper bound on the excess risk $\mathcal{E}^*(\widehat{A},\widehat{\beta})$. The result is stated in a general way for any RKHS's $\mathcal{H}(\mathbb{K}_\beta)$ and $\mathcal{H}(\mathbb{K})$, provided that the spectrum of $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ satisfies the polynomial decay rate in (9). There are three components in the upper bound in (10).

- The term δ_T is an upper bound on the error associated with the estimation of the bivariate coefficient function A^* in the presence of functional noise. As formally stated in (9), δ_T is determined by the alignment between the kernel $\mathbb{K}$ of the RKHS that A^* resides in and the covariance operator of X. This is the well-known nonparametric rate frequently seen in the functional regression literature, see e.g. Cai and Yuan (2012).
- The term $\mathfrak{s}\log(p)T^{-1}$ is an upper bound on the error associated with the estimation of the high-dimensional sparse univariate coefficient functions β^* and is a parametric rate frequently seen in the high-dimensional linear regression settings (e.g. Bühlmann and van de Geer, 2011).
- The term ζ_n is an upper bound on the error due to the fact that the functional variables $\{X_t, Y_t\}_{t=1}^T$ are only observed on discrete sample points. Recall model (1) consists of two components $\int_{[0,1]} A^*(\cdot,s)X_t(s)\,\mathrm{d}s$ and $\langle Z_t, \beta^*(\cdot)\rangle$. The discretization errors are captured through $\zeta_n = n^{-\alpha+1/2}$, which only depends on the smoothness of $W^{\alpha,2}$ with $\alpha > 1/2$.

We show later in Theorem 7 that, provided the sample points are dense enough, e.g. $n \gg T$, up to a logarithmic factor of T, this upper bound achieves minimax optimality. In Appendix L, we further provide numerical illustration for the nonparametric rate δ_T and the high-dimensional parametric rate $\mathfrak{s}\log(p\vee T)T^{-1}$ in Theorem 4.

Remark 5 (New peeling techniques) *To prove Theorem 4, we develop new peeling techniques to obtain new exponential tail bounds, which are crucial in dealing with the potentially exponentially growing dimension p.*

Remark 6 (Phase transition) *For sufficiently many samples, i.e. large n, the upper bound in* (10) *implies that, there exists an absolute constant* $C > 0$ *such that with large probability*

$$\mathcal{E}^*(\widehat{A}, \widehat{\beta}) \leq C \log(T)\big\{\delta_T + \mathfrak{s} \log(p \vee T) T^{-1}\big\}.$$

This unveils a phase transition between the nonparametric regime and the high-dimensional parametric regime, governed by the eigen-decay of the linear operator $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ *and the sparsity of the univariate coefficient functions* β^**. To be specific, if* $\delta_T \gtrsim \mathfrak{s}\log(p \vee T)T^{-1}$*, the high-probability upper bound on the excess risk is determined by the nonparametric rate* δ_T*; otherwise, the high-dimensional parametric rate dominates.*

Theorem 16 in Appendix B.2 further presents a formal theoretical guarantee which quantifies a discretized version of the excess risk defined in Theorem 2 for the proposed estimators $(\widehat{A}, \widehat{\beta})$. Due to the fact that the functional variables are only partially observed on discrete sample points, the discretized version of the excess risk can be more relevant in certain practical applications.

Lower bounds

In this section, we derive a matching lower bound on the excess risk $\mathcal{E}^*(\widehat{A}, \widehat{\beta})$ and thus show that the upper bound provided in Theorem 4 is nearly minimax optimal in terms of the sample size T, the dimension p and the sparsity parameter $\mathfrak{s}$, saving for a logarithmic factor. We establish the lower bound under the assumption that the functional variables $\{X_t, Y_t\}_{t=1}^T$ are fully observed, which is equivalent to setting $n = \infty$. Thus, we do not claim optimality in terms of the number of sample points n for the result in Theorem 4.

Proposition 7 *Under Assumptions 1 and 2, suppose that the functional variables* $\{X_t, Y_t\}_{t=1}^T$ *are fully observed and the eigenvalues* $\{\xi_k\}_{k=1}^\infty$ *of the linear operator* $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ *satisfy* $\xi_k \asymp k^{-2r}$*, for some* $r > 1/2$*. There exists a sufficiently small constant* $c > 0$ *such that*

$$\inf_{\widehat{A}, \widehat{\beta}} \sup_{A^* \in \mathcal{C}_A, \beta^* \in \mathcal{C}_\beta} \mathbb{E}\{\mathcal{E}^*(\widehat{A}, \widehat{\beta})\} \geq c\{T^{-\frac{2r}{2r+1}} + \mathfrak{s}\log(p)T^{-1}\},$$

where the infimum is taken over all possible estimators of A^* *and* β^* *based on the observations* $\{X_t, Z_t, Y_t\}_{t=1}^T$.

3.4 Extension to functional responses with measurement errors

In this subsection, we consider the setting where the functional responses are corrupted with measurement errors. More precisely, let the functional response $Y_t(r)$ be generated as in (1). Assume we observe

$$y_{t,j} = Y_t(r_j) + \mathfrak{E}_{t,j}, \quad t \in \{1, \dots, T\},\, j \in \{1, \dots, n_2\},$$

where $\{\mathfrak{E}_{t,j}\}_{t=1,j=1}^{T,n_2}$ is a collection of independent and identically distributed sub-Gaussian measurement errors with mean zero and $\mathrm{Var}(\mathfrak{E}_{t,j}) \leq C_{\mathfrak{E}}$.

Given the observations $\{X_t(s_i), Z_t, y_{t,j}\}_{t=1,i=1,j=1}^{T,n_1,n_2}$, consider

$$(\widehat{A}, \widehat{\beta}) = \mathop{\arg\min}_{\substack{A \in \mathcal{C}_A, \\ \beta = \{\beta_l\}_{l=1}^p \in \mathcal{C}_\beta}} \left[\frac{1}{Tn_2} \sum_{t=1}^T \sum_{j=1}^{n_2} w_r(j) \left\{ y_{t,j} - \frac{1}{n_1} \sum_{i=1}^{n_1} w_s(i) A(r_j, s_i) X_t(s_i) - \langle \beta(r_j), Z_t \rangle_p \right\}^2 + \lambda \sum_{l=1}^p \|\beta_l\|_{n_2} \right], \tag{11}$$

where $\mathcal{C}_A$ and $\mathcal{C}_\beta$ are defined in (5). Note that without the presence of measurement errors, i.e. $\mathfrak{E}_{t,j} = 0$ for all t and j, the estimator (11) is identical to (4) proposed for the setting with only functional noise. In what follows, we show that the excess risk of the estimator (11) also achieves the same convergence rate as that in Theorem 4.

Theorem 8 *Suppose all the assumptions in Theorem 4 hold. Let $(\widehat{A}, \widehat{\beta})$ be any solution to (11) with the tuning parameter $\lambda = C_\lambda \sqrt{\log(p \vee T) T^{-1}}$ for some sufficiently large constant C_λ. For any $T \gtrsim \log(n)$, there exists an absolute constant $C > 0$ such that with probability at least $1 - 8T^{-4}$, it holds that*

$$\mathcal{E}^*(\widehat{A}, \widehat{\beta}) \leq C \log(T) \big\{ \delta_T + \mathfrak{s} \log(p \vee T) T^{-1} + \zeta_n \big\}, \tag{12}$$

*where $\zeta_n = n^{-\alpha+1/2}$, $\mathfrak{s}$ is the sparsity parameter defined in Assumption 1(***b***) and $\delta_T = T^{-2r/(2r+1)}$.*

4. Optimization

In this section, we propose an efficient convex optimization algorithm for solving (4) given the observations $\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=1,i=1,j=1}^{T,n_1,n_2}$. We remark that identical optimization can be applied to (11). To ease presentation, in the following we assume without loss of generality that $\mathcal{H}(\mathbb{K}) = \mathcal{H}(\mathbb{K}_\beta)$, and thus $\mathbb{K} = \mathbb{K}_\beta$. The general case where $\mathbb{K} \neq \mathbb{K}_\beta$ can be handled in exactly the same way with more tedious notation. Section 4.1 formulates (4) as a convex optimization problem and Section 4.2 further proposes a novel iterative coordinate descent algorithm to efficiently solve the formulated convex optimization.

4.1 Formulation of the convex optimization

By the equivalence between constrained and penalized optimization (see e.g. Hastie et al., 2009), we can reformulate (4) into a penalized optimization such that

$$(\widehat{A}, \widehat{\beta}) = \mathop{\arg\min}_{A, \beta} \left[\sum_{t=1}^T \sum_{j=1}^{n_2} w_r(j) \left\{ Y_t(r_j) - \frac{1}{n_1} \sum_{i=1}^{n_1} w_s(i) A(r_j, s_i) X_t(s_i) - \langle \beta(r_j), Z_t \rangle_p \right\}^2 + \lambda_1 \|A\|^2_{\mathrm{F}(\mathbb{K})} + \lambda_2 \sum_{l=1}^p \|\beta_l\|_{\mathcal{H}(\mathbb{K})} + \lambda_3 \sum_{l=1}^p \|\beta_l\|_{n_2} \right], \tag{13}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ denote tuning parameters. Note that for notational simplicity, we drop the factor $1/(Tn_2)$ of the squared loss. In the following, we show that (13) can be solved via convex optimization.

We first define some necessary notation. For $t = 1, \dots, T$, denote the functional curves observed on discrete grids as

$$Y_t = [Y_t(r_1), Y_t(r_2), \cdots, Y_t(r_{n_2})]^\top \in \mathbb{R}^{n_2} \quad \text{and} \quad X_t = [X_t(s_1), X_t(s_2), \cdots, X_t(s_{n_1})]^\top \in \mathbb{R}^{n_1}.$$

Define $Y = [Y_1, \cdots, Y_T] \in \mathbb{R}^{n_2 \times T}$, $X = [X_1, \cdots, X_T] \in \mathbb{R}^{n_1 \times T}$ and $Z = [Z_1, \cdots, Z_T] \in \mathbb{R}^{p \times T}$.

For any $r, s \in [0, 1]$, denote the RKHS kernels as

$$k_1(r) = [\mathbb{K}(r, r_1), \mathbb{K}(r, r_2), \cdots, \mathbb{K}(r, r_{n_2})]^\top \in \mathbb{R}^{n_2} \text{ and } k_2(s) = [\mathbb{K}(s, s_1), \mathbb{K}(s, s_2), \cdots, \mathbb{K}(s, s_{n_1})]^\top \in \mathbb{R}^{n_1}.$$

Denote $K_1 = [k_1(r_1), k_1(r_2), \cdots, k_1(r_{n_2})] \in \mathbb{R}^{n_2 \times n_2}$ and $K_2 = [k_2(s_1), k_2(s_2), \cdots, k_2(s_{n_1})] \in \mathbb{R}^{n_1 \times n_1}$. Note that $K_1 = \langle k_1(r), k_1(r) \rangle_{\mathcal{H}(\mathbb{K})}$ and $K_2 = \langle k_2(s), k_2(s) \rangle_{\mathcal{H}(\mathbb{K})}$, thus both are symmetric and positive definite matrices. Furthermore, denote $W_S = \text{diag}(w_s(1), w_s(2), \cdots, w_s(n_1)) \in \mathbb{R}^{n_1 \times n_1}$ and $W_R = \text{diag}(\sqrt{w_r(1)}, \sqrt{w_r(2)}, \cdots, \sqrt{w_r(n_2)}) \in \mathbb{R}^{n_2 \times n_2}$ as the diagonal weight matrices. Define $Y^* = W_R Y$, $K_1^* = W_R K_1$ and $K_2^* = K_2 W_S$.

By the representer theorem (Theorem 1), we have the minimizer of (13) taking the form

$$A(r, s) = k_1(r)^\top R k_2(s) \quad \text{and} \quad \beta_l(r) = k_1(r)^\top \mathbf{b}_l, \quad l = 1, \cdots, p,$$

where $R \in \mathbb{R}^{n_2 \times n_1}$ is an $n_2 \times n_1$ matrix and $\mathbf{b}_l = [b_{l1}, b_{l2}, \cdots, b_{ln_2}]^\top \in \mathbb{R}^{n_2}$ is an n_2-dimensional vector for $l = 1, \cdots, p$. Denote $\beta(r) = [\beta_1(r), \beta_2(r), \cdots, \beta_p(r)]^\top = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_p]^\top k_1(r) = B^\top k_1(r)$, where $B = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_p] \in \mathbb{R}^{n_2 \times p}$.

By straightforward algebra, we can rewrite the optimization problem in (13) as

$$\left\| Y^* - \frac{1}{n_1} K_1^* R K_2^* X - K_1^* B Z \right\|_{\text{F}}^2 + \lambda_1 \text{tr}(R^\top K_1 R K_2) + \lambda_2 \sum_{l=1}^p \sqrt{\mathbf{b}_l^\top K_1 \mathbf{b}_l} + \lambda_3 \sum_{l=1}^p \sqrt{\frac{1}{n_2} \mathbf{b}_l^\top K_1^{*\top} K_1^* \mathbf{b}_l}, \tag{14}$$

where $\|\cdot\|_{\text{F}}$ and $\text{tr}(\cdot)$ are the Frobenius norm and trace of a matrix. We refer to Appendix F.1 for the detailed derivation.

It is easy to see that (14) is a convex function of R and B. Note that the first two terms of (14) are quadratic functions and can be handled easily, while the main difficulty of the optimization lies in the group Lasso-type penalty $\lambda_2 \sum_{l=1}^p \sqrt{\mathbf{b}_l^\top K_1 \mathbf{b}_l} + \lambda_3 \sum_{l=1}^p \sqrt{\mathbf{b}_l^\top K_1^{*\top} K_1^* \mathbf{b}_l / n_2}$.

4.2 Iterative coordinate descent

In this section, we propose an efficient iterative coordinate descent algorithm which solves (14) by iterating between the optimization of R and B.

Optimization of R (i.e. the bivariate coefficient function A): Given a fixed $B = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_p]$, denote $\widetilde{Y} = Y^* - K_1^* B Z \in \mathbb{R}^{n_2 \times T}$. We have that (14) reduces to a function of R that

$$\|\widetilde{Y} - n_1^{-1} K_1^* R K_2^* X\|_{\text{F}}^2 + \lambda_1 \text{tr}(R^\top K_1 R K_2). \tag{15}$$

Define $E = K_1^{1/2} R K_2^{1/2} \in \mathbb{R}^{n_2 \times n_1}$ and $\mathbf{e} = \mathrm{vec}(E)$. Denote $\widetilde{\mathbf{y}} = \mathrm{vec}(\widetilde{Y}) = [\widetilde{Y}_1^\top, \widetilde{Y}_2^\top, \cdots, \widetilde{Y}_T^\top]^\top$ and denote $S_1 = n_1^{-1}(X^\top W_S K_2^{1/2}) \otimes (W_R K_1^{1/2}) \in \mathbb{R}^{Tn_1 \times n_1 n_2}$. We can rewrite (15) as

$$\|\widetilde{\mathbf{y}} - S_1 \mathbf{e}\|_2^2 + \lambda_1 \|\mathbf{e}\|_2^2, \tag{16}$$

which can be seen as a classical ridge regression with a structured design matrix S_1. This ridge regression has a closed-form solution and can be solved efficiently by exploiting the Kronecker structure of S_1, see Appendix F.2 for more details. Thus, given B, R can be updated efficiently.

Optimization of B (i.e. the univariate coefficient functions β): Given a fixed R, with some abuse of notation, denote $\widetilde{Y} = Y^* - n_1^{-1} K_1^* R K_2^* X \in \mathbb{R}^{n_2 \times T}$. For $l = 1, \ldots, p$, define $\mathbf{h}_l = K_1^* \mathbf{b}_l \in \mathbb{R}^{n_2}$ and $H = K_1^* B = [\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_p] \in \mathbb{R}^{n_2 \times p}$. Further define $K_3 = (K_1^*)^{-1} K_1 (K_1^*)^{-1}$. We have that (14) reduces to a function of H (and thus B), i.e.

$$\begin{aligned} &\left\|\widetilde{Y} - HZ\right\|_{\mathrm{F}}^2 + \lambda_2 \sum_{l=1}^p \sqrt{\mathbf{h}_l^\top K_3 \mathbf{h}_l} + \lambda_3 \sum_{l=1}^p \sqrt{\frac{1}{n_2}\mathbf{h}_l^\top \mathbf{h}_l} \\ =& \sum_{t=1}^T \left\|\widetilde{Y}_t - \sum_{l=1}^p Z_{tl}\mathbf{h}_l\right\|_2^2 + \lambda_2 \sum_{l=1}^p \sqrt{\mathbf{h}_l^\top K_3 \mathbf{h}_l} + \frac{\lambda_3}{\sqrt{n_2}} \sum_{l=1}^p \|\mathbf{h}_l\|_2, \end{aligned} \tag{17}$$

which can be seen as a linear regression with two group penalties: a weighted group Lasso penalty and a standard group Lasso penalty. We solve the optimization of (17) by performing coordinate descent on $\mathbf{h}_l, l = 1, 2 \cdots, p$. See Friedman et al. (2010) for a similar strategy used to solve the sparse group Lasso problem for a linear regression with both a Lasso and a group Lasso penalty.

Specifically, for each $l = 1, 2, \cdots, p$, the Karush–Kuhn–Tucker condition for $\mathbf{h}_l$ is

$$-2\sum_{t=1}^T Z_{tl}\left(\widetilde{Y}_t - \sum_{l=1}^p Z_{tl}\mathbf{h}_l\right) + \lambda_2 K_4 \mathbf{s}_l^{(1)} + \frac{\lambda_3}{\sqrt{n_2}}\mathbf{s}_l^{(2)} = 0.$$

Define K_4 as the root of K_3, i.e. $K_4 K_4 = K_3$. We have that $K_4 \mathbf{s}_l^{(1)}$ and $\mathbf{s}_l^{(2)}$ are the subgradients of $\sqrt{\mathbf{h}_l^\top K_3 \mathbf{h}_l}$ and $\|\mathbf{h}_l\|_2$ respectively, such that

$$\mathbf{s}_l^{(1)} = \begin{cases} K_4\mathbf{h}_l / \|K_4 \mathbf{h}_l\|_2, & \mathbf{h}_l \neq 0, \\ \text{a vector with } \|\mathbf{s}_l^{(1)}\|_2 \leq 1, & \mathbf{h}_l = 0, \end{cases} \quad \text{and} \quad \mathbf{s}_l^{(2)} = \begin{cases} \mathbf{h}_l / \|\mathbf{h}_l\|_2, & \mathbf{h}_l \neq 0, \\ \text{a vector with } \|\mathbf{s}_l^{(2)}\|_2 \leq 1, & \mathbf{h}_l = 0. \end{cases}$$

Define $\widetilde{Y}_t^l = \widetilde{Y}_t - \sum_{j \neq l} Z_{tj}\mathbf{h}_j$. Thus, given $\{\mathbf{h}_j, j \neq l\}$, we have that $\mathbf{h}_l = 0$ is the optimal solution if there exist two vectors $\mathbf{s}_l^{(1)}$ and $\mathbf{s}_l^{(2)}$ with $\|\mathbf{s}_l^{(1)}\|_2 \leq 1$ and $\|\mathbf{s}_l^{(2)}\|_2 \leq 1$, and

$$\mathbf{s}_l^{(1)} = \frac{1}{\lambda_2} K_4^{-1}\left(2\sum_{t=1}^T Z_{tl}\widetilde{Y}_t^l - \frac{\lambda_3}{\sqrt{n}}\mathbf{s}_l^{(2)}\right).$$

This is equivalent to checking

$$\min_{\|\mathbf{s}\|_2 \leq 1} \left\|2\sum_{t=1}^T Z_{tl}\widetilde{Y}_t^l - \frac{\lambda_3}{\sqrt{n}}\mathbf{s}\right\|_2 \leq 1,$$

which is a standard constrained optimization problem and can be solved efficiently.

Otherwise, $\mathbf{h}_l \neq 0$ and to update $\mathbf{h}_l$, we need to optimize

$$\sum_{t=1}^{T} \|\widetilde{Y}_t^l - Z_{tl}\mathbf{h}_l\|_2^2 + \lambda_2 \sqrt{\mathbf{h}_l^\top K_3 \mathbf{h}_l} + \frac{\lambda_3}{\sqrt{n_2}}\|\mathbf{h}_l\|_2. \tag{18}$$

We again solve this optimization by performing coordinate descent on $\mathbf{h}_{lk}, k = 1, 2, \cdots, n_2$, where for each k, given $\{\mathbf{h}_{lj}, j \neq k\}$, we can update $\mathbf{h}_{lk}$ by solving a simple one-dimensional optimization

$$\min\left\{\sum_{t=1}^{T}(\widetilde{Y}_{tk}^l - Z_{tl}\mathbf{h}_{lk})^2 + \lambda_2\sqrt{K_{3,kk}\mathbf{h}_{lk}^2 + 2\mathbf{h}_{lk}\sum_{j\neq k} K_{3,kj}\mathbf{h}_{lj} + \sum_{i,j\neq k}\mathbf{h}_{li}K_{3,ij}\mathbf{h}_{lj}} + \frac{\lambda_3}{\sqrt{n_2}}\sqrt{\mathbf{h}_{lk}^2 + \sum_{j\neq k}\mathbf{h}_{lj}^2}\right\}. \tag{19}$$

Thus, given R, B can also be updated efficiently.

The iterative coordinate descent algorithm: Algorithm 1 formalizes the above discussion and outlines the proposed iterative coordinate descent algorithm. The convergence of coordinate descent for convex optimization is guaranteed under mild conditions, see e.g. Wright (2015). Empirically, the proposed algorithm is found to be efficient and stable, and typically reaches a reasonable convergence tolerance within a few iterations.

5. Numerical results

In this section, we conduct extensive numerical experiments to investigate the performance of the proposed RKHS-based penalized estimator (hereafter RKHS) for the functional linear regression with mixed predictors. Sections 5.1-5.3 compare RKHS with popular methods in the literature via simulation studies. Section 5.4 presents a real data application on crowdfunding prediction to further illustrate the potential utility of the proposed method. The implementations of our numerical experiments can be found at `https://github.com/darenwang/functional_regression`.

5.1 Simulation settings

Data generating process: We simulate data from the functional linear regression model

$$Y_t(r) = \int_{[0,1]} A^*(r,s)X_t(s)\,\mathrm{d}s + \sum_{j=1}^{p}\beta_j^*(r)Z_{tj} + \epsilon_t(r), \quad r \in [0,1], \tag{20}$$

for $t = 1, 2, \ldots, T$. Note that for $p = 0$, (20) reduces to the classical function-on-function regression.

We generate the functional covariate $\{X_t\}$ and the functional noise $\{\epsilon_t\}$ from a q-dimensional subspace spanned by basis functions $\{u_i(s)\}_{i=1}^q$, where $\{u_i(s)\}_{i=1}^q$ consists of orthonormal basis of $\mathcal{L}^2[0,1]$. Following Yuan and Cai (2010), we set $u_i(s) = 1$ if $i = 1$ and $u_i(s) = \sqrt{2}\cos((i-1)\pi s)$ for $i = 2, \ldots, q$. Thus, we have $X_t(r) = \sum_{i=1}^q x_{ti}u_i(r)$ and $\epsilon_t(r) = \sum_{i=1}^q e_{ti}u_i(r)$. For $x_t = (x_{t1}, \ldots, x_{tq})^\top$, we simulate $x_{ti} \overset{\text{i.i.d.}}{\sim} \text{Unif}[-1/i, 1/i]$,

Algorithm 1 Iterative coordinate descent

input: Observations $\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=1,i=1,j=1}^{T,n_1,n_2}$, tuning parameters $(\lambda_1, \lambda_2, \lambda_3)$, the maximum iteration L_{max} and tolerance ϵ.
initialization: $L = 1, B_0 = R_0 = 0$.
repeat ▷ First level block coordinate descent
  Given $B = B_{L-1}$, update R_L via the ridge regression formulation (16).
  Given $R = R_L$, set $\widetilde{Y} = Y^* - 1/n_1 K_1^* R K_2^* X$ and initialize $H = K_1^* B_{L-1}$.
  repeat ▷ Second level coordinate descent
    for $l = 1, 2, \cdots, p$ **do**
      Given $\{\mathbf{h}_j, j \neq l\}$, set $\widetilde{Y}_t^l = \widetilde{Y}_t - \sum_{j \neq l} Z_{tj} \mathbf{h}_j$, for $t = 1, \cdots, T$.
      if $\min_{\|\mathbf{s}\|_2 \leq 1} \left\| 2 \sum_{t=1}^T Z_{tl} \widetilde{Y}_t^l - \frac{\lambda_3}{\sqrt{n}} \mathbf{s} \right\|_2 \leq 1$ **then**
        Update $\mathbf{h}_l = 0$.
      else
        repeat ▷ Third level coordinate descent
          for $k = 1, 2, \cdots, n_2$ **do**
            Given $\{\mathbf{h}_{lj}, j \neq k\}$, update $\mathbf{h}_{lk}$ via the one-dimensional optimization (19).
          end for
        until Decrease of function value (18) $< \epsilon$.
      end if
    end for
  until Decrease of function value (17) $< \epsilon$.
  Update $B_L = K_1^{*-1} H$ and set $L \leftarrow L + 1$.
until Decrease of function value (14) $< \epsilon$ or $L \geq L_{max}$.
output: $\widehat{R} = R_L$ and $\widehat{B} = B_L$.

$i = 1, \ldots, q$. For $e_t = (e_{t1}, \ldots, e_{tq})^\top$, we simulate $e_{ti} \overset{\text{i.i.d.}}{\sim} \text{Unif}[-0.2/i, 0.2/i]$, $i = 1, \ldots, q$. For the vector covariate $\{Z_{tj}\}$, we simulate $Z_{tj} \overset{\text{i.i.d.}}{\sim} \text{Unif}[-1/\sqrt{3}, 1/\sqrt{3}]$, $j = 1, \ldots, p$.

For the coefficient functions $A^*(r,s)$ and $\{\beta_j^*(r)\}_{j=1}^p$, we consider Scenarios A and B.

- Scenario A (Exponential): We set $A^*(r,s) = \kappa\sqrt{3}e^{-(r+s)}$, $\beta_1^*(r) = \kappa\sqrt{3}e^{-r}$ and $\beta_j^*(r) \equiv 0$, $j = 2, \ldots, p$.
- Scenario B (Random): We set $A^*(r,s) = \kappa\sum_{i,j=1}^q \lambda_{ij}u_i(r)u_j(s)$, $\beta_1^*(r) = \kappa\sum_{i=1}^q b_i u_i(r)$ and $\beta_j^*(r) \equiv 0$, $j = 2, \ldots, p$. Define matrix $\Lambda = (\lambda_{ij})$ and $b = (b_1, \ldots, b_q)^\top$. We simulate $\lambda_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and rescale Λ such that its spectral norm $\|\Lambda\|_{\text{op}} = 1$. For b, we simulate $b_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and rescale b such that $\|b\|_2 = 1$.

Note that (A^*, β^*) in Scenario B is more complex than that in Scenario A, especially when q is large, while for Scenario A, its complexity is insensitive to q. The parameter κ is later used to control the signal-to-noise ratio (SNR). Here, we define the SNR for $A^*(r,s)$ as

$$\sqrt{\mathbb{E}\int_{[0,1]}\left[\int_{[0,1]} A^*(r,s)X_t(s)\,\mathrm{d}s\right]^2 dr} \Big/ \sqrt{\mathbb{E}\int_{[0,1]} \epsilon_t(r)^2\,\mathrm{d}r},$$

which roughly equals to 1, 2 and 4 as we vary $\kappa = 0.5, 1, 2$. Similarly, we define SNR for β^* (note that only β_1^* is a non-zero function) as

$$\sqrt{\mathbb{E}\int_{[0,1]}\left[\beta_1^*(r)Z_{t1}\right]^2 \mathrm{d}r} \Big/ \sqrt{\mathbb{E}\int_{[0,1]} \epsilon_t(r)^2\,\mathrm{d}r},$$

which also roughly equals to 1, 2 and 4 as we vary $\kappa = 0.5, 1, 2$.

For simplicity, we set the discrete sample points $\{r_j\}_{j=1}^{n_2}$ for Y_t and $\{s_i\}_{i=1}^{n_1}$ for X_t to be evenly spaced grids on $[0,1]$ with the same number of grids $n = n_1 = n_2$. The simulation result for random sample points where $\{r_j\}_{j=1}^{n_2}$ and $\{s_i\}_{i=1}^{n_1}$ are generated independently via the uniform distribution on [0,1] is similar and thus omitted.

Evaluation criteria: We evaluate the performance of the estimator by its excess risk. Specifically, given the sample size (n, T), we simulate observations $\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=1,i=1,j=1}^{T+0.5T,n,n}$, which are then split into the training data $\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=1,i=1,j=1}^{T,n,n}$ for constructing the estimator $(\widehat{A}, \widehat{\beta})$ and the test data $\{X_t(s_i), Z_t, Y_t(r_j)\}_{t=T+1,i=1,j=1}^{T+0.5T,n,n}$ for the evaluation of the excess risk. Based on $(\widehat{A}, \widehat{\beta})$ and the predictors $\{X_t(s_i), Z_t\}_{t=T+1,i=1}^{T+0.5T,n}$, we generate the prediction $\{\widehat{Y}_t(r)\}_{t=T+1}^{T+0.5T}$ and define

$$\text{RMISE}(\widehat{A}, \widehat{\beta}) = \sqrt{\frac{1}{0.5T}\sum_{t=T+1}^{T+0.5T}\int_{[0,1]}\left(Y_t^{\text{oracle}}(r) - \widehat{Y}_t(r)\right)^2 \mathrm{d}r}, \tag{21}$$

$$\text{nRMISE}(\widehat{A}, \widehat{\beta}) = \sqrt{\frac{\frac{1}{0.5T}\sum_{t=T+1}^{T+0.5T}\int_{[0,1]}\left(Y_t^{\text{oracle}}(r) - \widehat{Y}_t(r)\right)^2 \mathrm{d}r}{\frac{1}{0.5T}\sum_{t=T+1}^{T+0.5T}\int_{[0,1]}\left(Y_t^{\text{oracle}}(r) - 0\right)^2 \mathrm{d}r}}, \tag{22}$$

where $Y_t^{\text{oracle}}(r) = \int_{[0,1]} A(r,s)X_t(s)\,\mathrm{d}s + \beta_1(r)Z_{t1}$ is the oracle prediction of $Y_t(r)$. Note that nRMISE is a normalized RMISE, which can be viewed as a percentage error and thus is easy to assess and interpret. Smaller RMISE and nRMISE indicate a better recovery of the signal.

Simulation settings: We consider three simulation settings for (n, T, q) where $(n, T, q) \in \{(5, 50, 5), (20, 100, 20), (40, 200, 50)\}$. For each setting, we vary $\kappa \in \{0.5, 1, 2\}$, which roughly corresponds to SNR $\in \{1, 2, 4\}$. As for the number of scalar predictors p, Section 5.2 considers the classical function-on-function regression, which is a special case of (20) with $p = 0$ and Section 5.3 considers functional regression with mixed predictors and sets $p = 3, 10, 50, 100, 200$. For each setting, we conduct 500 experiments.

Implementation details of the RKHS estimator: We set $\mathbb{K} = \mathbb{K}_\beta$ and use the rescaled Bernoulli polynomial as the reproducing kernel such that

$$\mathbb{K}(x, y) = 1 + k_1(x)k_1(y) + k_2(x)k_2(y) - k_4(x - y),$$

where $k_1(x) = x - 0.5$, $k_2(x) = 2^{-1}\{k_1^2(x) - 1/12\}$, $k_4(x) = 1/24\{k_1^4(x) - k_1^2(x)/2 + 7/240\}$, $x \in [0, 1]$, and $k_4(x - y) = k_4(|x - y|)$, $x, y \in [0, 1]$. Such $\mathbb{K}$ is the reproducing kernel for $W^{2,2}$. See Chapter 2.3.3 of Gu (2013) for more details. In Algorithm 1, we set the tolerance parameter $\epsilon = 10^{-8}$ and the maximum iterations $L_{\max} = 10^4$. A standard 5-fold cross-validation (CV) on the training data is used to select the tuning parameters $(\lambda_1, \lambda_2, \lambda_3)$. Note that for $p = 0$, we explicitly set $\beta \equiv 0$ in the penalized optimization (13), which reduces to the structured ridge regression in (16) and can be solved efficiently with a closed-form solution as detailed in Section 4.2.

5.2 Function-on-function regression

In this subsection, we set $p = 0$ and compare the proposed RKHS estimator with two popular methods in the literature for function-on-function regression.

The first competitor (Ramsay and Silverman, 2005) estimates the coefficient function $A(r, s)$ based on a penalized B-spline basis function expansion. We denote this approach as FDA and it is implemented via the R package `fda.usc` (`fregre.basis.fr` function). The second competitor (Ivanescu et al., 2015) is the penalized flexible functional regression (PFFR) in and is implemented via the R package `refund` (`pffr` function). Unlike our proposed RKHS method, neither FDA or PFFR has optimal theoretical guarantees.

Both FDA and PFFR are based on penalized basis function expansions and require a hyper-parameter N_b, which is the number of basis. Intuitively, the choice of N_b is related to the complexity of $A^*(r, s)$, which is unknown in practice. In the literature, it is recommended to set N_b at a large number to guard against the underfitting of $A^*(r, s)$. On the other hand, a larger N_b can potentially incur higher estimation variance and will also increase the computational cost. For a fair comparison, for both FDA and PFFR we set $N_b = 20$, which is sufficient to accommodate the model complexity across all simulation settings except for Scenario B with $q = 50$. See more discussions later. We use a standard 5-fold CV to select the tuning parameter λ_1 of RKHS and the roughness penalty of FDA from the range $10^{(-15:0.5:2)}$. PFFR uses a restricted maximum likelihood (REML) approach to automatically select the roughness penalty.

Besides the penalized method based on basis function expansions (i.e. FDA and PFFR), another popular method for function-on-function regression in the literature is based on functional PCA (i.e. the Karhunen–Loéve expansion, FPCA), see for example Yao et al. (2005). However, FPCA does not seem to perform competitively in our simulation and real data analysis, see similar observations in Cai and Yuan (2012) and Sun et al. (2018). For completeness, we report the performance of FPCA in Appendix H and Appendix K.

Numerical results: For each method, Table 1 reports its average nRMISE ($\text{nRMISE}_{\text{avg}}$) across 500 experiments under all simulation settings. For each simulation setting, we further report R_{avg}, which is the percentage improvement of excess risk given by RKHS defined as

$$\text{R}_{\text{avg}}=[\min\{\text{nRMISE}_{\text{avg}}(\text{FDA}),\ \text{nRMISE}_{\text{avg}}(\text{PFFR})\}\ /\ \text{nRMISE}_{\text{avg}}(\text{RKHS}) - 1] \times 100\%.$$

In addition, we report R_w, which is the percentage of experiments (among 500 experiments) where RKHS gives the lowest RMISE. We further give the boxplot of RMISE in Figure 2 under $\kappa = 1$. The boxplots of RMISE under $\kappa \in \{0.5, 2\}$ can be found in Appendix G.

As can be seen, RKHS in general delivers the best performance across almost all simulation settings. For all methods, the estimation performance improves with a larger SNR (reflected in κ). Since the complexity of Scenario A is insensitive to q, the performance of all methods improve as (n, T) increases under Scenario A. However, this is not the case for Scenario B, as a larger q increases the estimation difficulty. Compared to Scenario A, the excess risk RMISE of Scenario B is larger, especially for $q \in \{20, 50\}$, due to the more complex nature of the operator A^*. In general, the improvement of RKHS is more notable under high model complexity and SNR.

Note that for Scenario B with $q = 50$, both FDA and PFFR incur significantly higher excess risks due to the underfitting bias caused by the insufficient basis dimension N_b, indicating the potential sensitivity of the penalized basis function approaches to the hyperparameter. In comparison, RKHS automatically adapts to different levels of model complexity. In Appendix J, we further conduct the same simulation but with a larger number of basis dimension $N_b = 50$ for FDA and $N_b = 30$ for PFFR. For Scenario A, where the bivariate function $A^*(r, s)$ is a simple exponential function, FDA and PFFR give essentially the same performance, while for Scenario B with $q = 50$, FDA and PFFR give much improved performance due to lower underfitting bias, though still having a notable performance gap compared to RKHS. In addition, note that a larger N_b can significantly increase the computational cost of the penalized basis function approaches. We refer to Appendix J for more details.

In addition, Appendix I collects the results of the same simulation but with the observed functional responses $Y_t(r)$ being additionally corrupted with measurement errors, i.e. the scenario discussed in Section 3.4 (see Theorem 8). It is seen that the performance of RKHS only worsens slightly with the additional measurement errors, providing numerical support for Theorem 8 that measurement errors do not affect the convergence rate of RKHS.

5.3 Functional regression with mixed predictors

In this subsection, we set $p = 3$ and compare the performance of RKHS with PFFR, as FDA cannot handle mixed predictors. For RKHS, we use the standard 5-fold CV to select the tuning parameter $(\lambda_1, \lambda_2, \lambda_3)$ from the range $10^{(-4:1:0)} \times 10^{(-4:1:0)} \times 10^{(-1:1:3)}$ for Scenario A and from the range $10^{(-17:1:-13)} \times 10^{(-17:1:-13)} \times 10^{(-1:1:3)}$ for Scenario B. PFFR uses REML to automatically select the roughness penalty and we set $N_b = 20$ for PFFR as before.

	Scenario A: $n = 5, T = 50, q = 5$					Scenario B: $n = 5, T = 50, q = 5$				
κ	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)
0.5	15.98	17.97	19.80	12.46	64	20.86	22.98	22.28	6.79	63
1	9.14	10.31	10.98	12.81	68	10.42	11.50	11.12	6.63	71
2	4.99	6.18	5.70	14.37	72	5.21	5.75	5.56	6.68	74
	Scenario A: $n = 20, T = 100, q = 20$					Scenario B: $n = 20, T = 100, q = 20$				
κ	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)
0.5	10.61	12.15	24.84	14.57	80	37.41	45.91	36.51	−2.41	47
1	5.89	6.75	12.84	14.66	81	18.39	32.41	22.32	21.37	92
2	3.60	4.26	6.89	18.45	84	9.19	27.71	12.56	36.58	100
	Scenario A: $n = 40, T = 200, q = 50$					Scenario B: $n = 40, T = 200, q = 50$				
κ	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)
0.5	7.37	8.53	18.93	15.72	81	39.22	79.37	78.92	101.21	100
1	3.98	4.41	9.68	10.96	75	20.75	76.74	77.40	269.73	100
2	2.34	2.36	5.05	0.71	49	12.59	75.95	77.09	503.19	100

Table 1: Numerical performance of RKHS, FDA and PFFR under function-on-function regression. The reported $\text{nRMISE}_{\text{avg}}$ is multiplied by 100 in scale. R_{avg} reflects the percent improvement of RKHS over the best performing competitor, and R_w reflects the percentage of experiments in which RKHS achieves the lowest RMISE.

Numerical results: Table 2 reports the average nRMISE ($\text{nRMISE}_{\text{avg}}$) across 500 experiments for RKHS and PFFR under all simulation settings. Table 2 further reports R_{avg}, the percentage improvement of RKHS over PFFR, and R_w, the percentage of experiments (among 500 experiments) where RKHS returns lower RMISE. We further give the boxplot of RMISE in Figure 3 under $\kappa = 1$. The boxplots of RMISE under $\kappa = 0.5, 2$ can be found in Appendix G. The result is consistent with the one for function-on-function regression, where RKHS delivers the best performance across almost all simulation settings with notable improvement.

In the following, we further examine the performance of RKHS for high-dimensional scalar predictors. Specifically, keep the simulation setting identical as above, we increase the number of scalar predictors to $p = 10, 50, 100, 200$. As a reminder, the sparsity of $\beta^* = (\beta_1^*, \beta_2^*, \dots, \beta_p^*)$ is 1 as only β_1^* is a non-zero function. In other words, the additionally introduced scalar predictors $(Z_{t2}, Z_{t3}, \dots, Z_{tp})$ are purely noise. Thus, compared to the case of $p = 3$, the performance of RKHS and PFFR are expected to worsen with an increasing dimension p. However, thanks to the group Lasso-type penalty, which induces sparsity of the estimated coefficient functions $\widehat{\beta}$, we expect the excess risk of RKHS to grow at the rate of $O\big(\log(p)\big)$, as suggested in Theorem 4.

To conserve space, we present the result under the simulation setting with $(n, T, q) = (20, 100, 20)$ and $\kappa = 1$. Results under other settings are similar and thus omitted. Due to the lack of sparsity penalty, PFFR may not be suitable for the setting of high-dimensional predictors. For comparison, we only implement PFFR for $p = 10$. Figure 4 (A) and (B) give the boxplot of RMISE for RKHS and PFFR across 500 experiments. As expected, the RMISE of RKHS increases as the dimension p increases though at a rate slower than

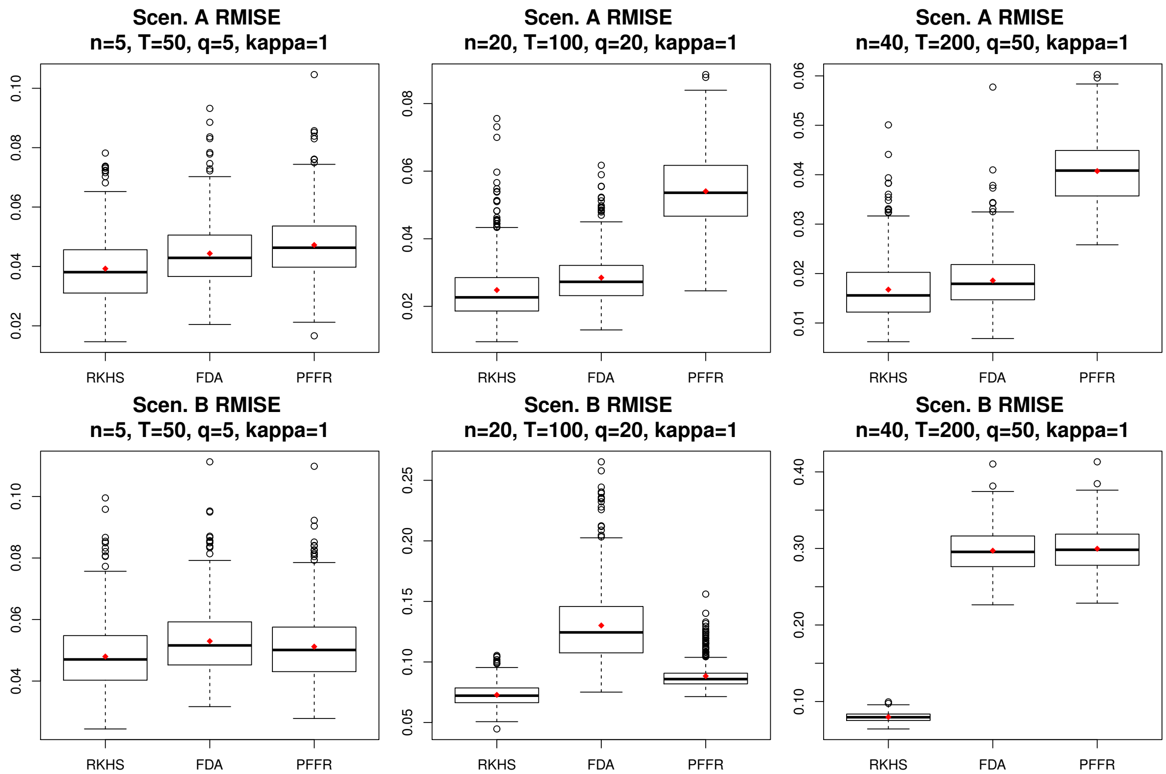

Figure 2: Boxplots of RMISE of RKHS, FDA and PFFR across 500 experiments under function-on-function regression with $\kappa = 1$. Red points denote the average RMISE.

p. Indeed, RKHS at $p = 200$ still gives better performance than PFFR at $p = 10$. Figure 4 (C) gives the plot between average MISE[1] across 500 experiments and $\log(p)$, where the relationship is seen to be roughly linear. This confirms that the excess risk of RKHS increases in the order of $O\big(\log(p)\big)$ as suggested in Theorem 4.

5.4 Real data applications

In this section, we conduct real data analysis to further demonstrate the promising utility of the proposed RKHS-based functional regression in the context of crowdfunding. In recent years, crowdfunding has become a flexible and cost-effective financing channel for start-ups and entrepreneurs, which helps expedite product development and diffusion of innovations.

We consider a novel dataset collected from one of the largest crowdfunding websites, kickstarter.com, which provides an online platform for creators, e.g. start-ups, to launch fundraising campaigns for developing a new product such as electronic devices and card games. The fundraising campaign takes place on a webpage set up by the creators, where information of the new product is provided, and typically has a 30-day duration with a goal amount G preset by the creators. Over the course of the 30 days, backers can pledge funds to the campaign, resulting in a functional curve of pledged funds $\{P(r), r \in [0, 30]\}$. At

1. To match the result in Theorem 4, we compute MISE, which is the squared RMISE with MISE = RMISE^2.

	Scenario A: $n = 5, T = 50, q = 5$				Scenario B: $n = 5, T = 50, q = 5$			
κ	RKHS	PFFR	R_{avg}(%)	R_w (%)	RKHS	PFFR	R_{avg}(%)	R_w (%)
0.5	17.16	20.14	17.35	77	14.81	15.99	8.02	70
1	9.43	10.89	15.53	76	7.42	7.97	7.46	75
2	5.03	5.60	11.35	74	3.72	3.99	7.29	75
	Scenario A: $n = 20, T = 100, q = 20$				Scenario B: $n = 20, T = 100, q = 20$			
κ	RKHS	PFFR	R_{avg}(%)	R_w (%)	RKHS	PFFR	R_{avg}(%)	R_w (%)
0.5	11.25	21.95	95.11	100	23.95	25.74	7.48	63
1	5.95	11.25	89.31	100	11.83	14.44	22.03	90
2	3.38	5.94	75.77	100	5.92	7.85	32.47	99
	Scenario A: $n = 40, T = 200, q = 50$				Scenario B: $n = 40, T = 200, q = 50$			
κ	RKHS	PFFR	R_{avg}(%)	R_w (%)	RKHS	PFFR	R_{avg}(%)	R_w (%)
0.5	8.50	15.76	85.46	99	27.52	77.13	180.22	100
1	4.36	8.02	84.03	100	14.23	76.19	435.38	100
2	2.33	4.16	78.36	100	8.20	75.99	827.09	100

Table 2: Numerical performance of RKHS and PFFR under functional regression with mixed predictors. The reported $\text{nRMISE}_{\text{avg}}$ is multiplied by 100 in scale. R_{avg} reflects the percent improvement of RKHS over PFFR and R_w reflects the percentage of experiments in which RKHS achieves the lower RMISE.

time r, the webpage displays real-time $P(r)$ along with other information of the campaign, such as its number of creators Z_1, number of product updates Z_2 and number of product FAQs Z_3.

A fundraising campaign succeeds if $P(30) \geq G$. Importantly, only creators of successful campaigns can be awarded the final raised funds $P(30)$ and the platform kickstarter.com charges a service fee ($5\% \times P(30)$) of successful campaigns. Thus, for both the platform and the creators, an accurate prediction of the future curve $\{P(r), r \in (s, 30]\}$ at an early time s is valuable, as it not only reveals whether the campaign will succeed but more importantly suggests timing along $(s, 30]$ for potential intervention by the creators and the platform to boost the fundraising campaign and achieve greater revenue.

The dataset consists of $T = 454$ campaigns launched between Dec-01-2019 and Dec-31-2019. For each campaign $t = 1, \ldots, T$, we observe its normalized curve $N_t(r) = P_t(r)/G_t$[2] at 60 evenly spaced time points over its 30-day duration and denote it as $\{N_t(r_i)\}_{i=1}^{60}$. See Figure 1 for normalized curves of six representative campaigns. At a time $s \in (0, 30)$, to predict $\{N_t(r), r \in (s, 30]\}$ for campaign t, we employ functional regression, where we treat $\{N_t(r_i), r_i \in (s, 30]\}$ as the functional response Y_t, use $\{N_t(r_i), r_i \in [0, s]\}$ as the functional covariate X_t and (Z_1, Z_2, Z_3) as the vector covariate. We compare the performances of RKHS, FDA and PFFR. Note that FDA only allows for one functional covariate, thus for RKHS and PFFR, we implement both the function-on-function regression and the functional regression with mixed predictors (denoted by $\text{RKHS}_{\text{mixed}}$ and $\text{PFFR}_{\text{mixed}}$). The implementation of each method is the same as that in Sections 5.2 and 5.3.

2. Note that the normalized curve $N_t(r)$ is as sufficient as the original curve $P_t(r)$ for monitoring the fundraising process of each campaign.

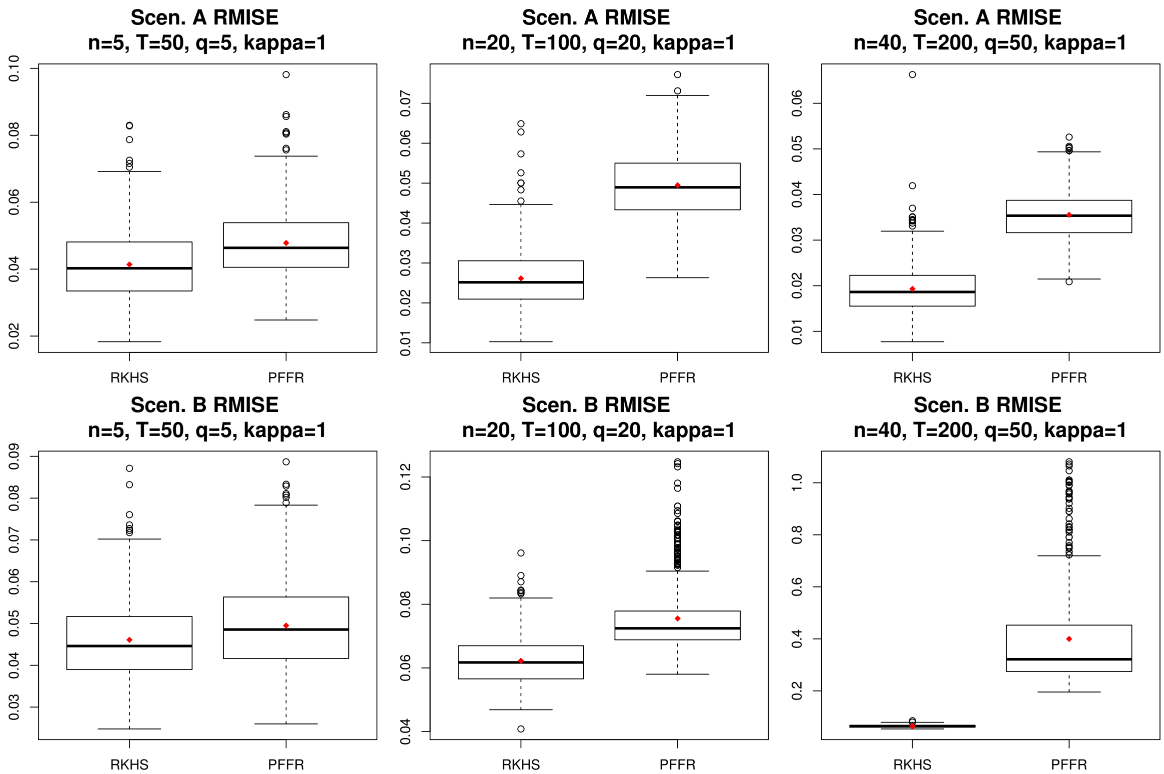

Figure 3: Boxplots of RMISE of RKHS and PFFR across 500 experiments under functional regression with mixed predictors with $\kappa = 1$. Red points denote the average RMISE.

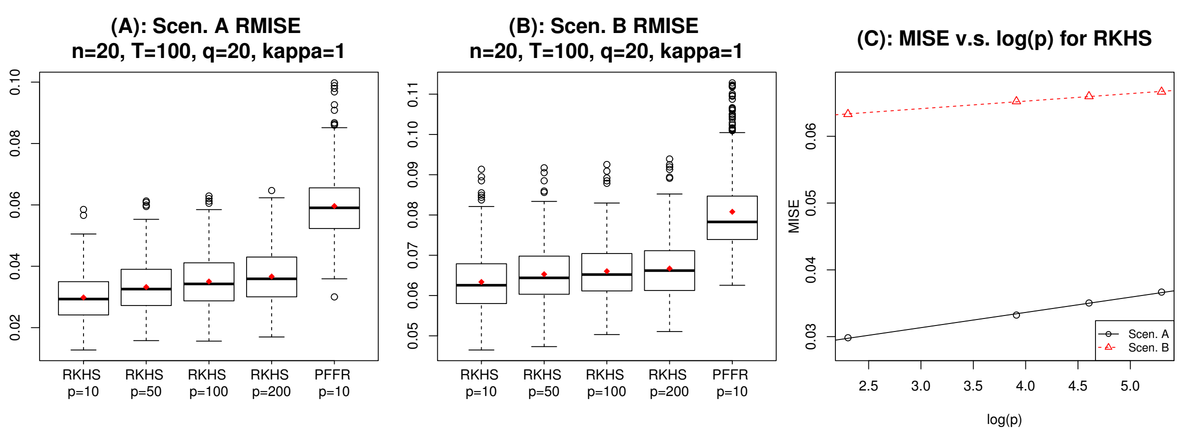

Figure 4: Boxplots of RMISE of RKHS and PFFR across 500 experiments under functional regression with mixed predictors with $\kappa = 1$ for $p = 10, 50, 100, 200$. Red points denote the average RMISE. For plot (C), MISE = RMISE^2 and the two lines are fitted via OLS.

We vary the prediction time s such that $s = 7\text{th}, 8\text{th}, \ldots, 20\text{th}$ day of a campaign. Note that we stop at $s = 20$ as prediction made in the late stage of a campaign is not as useful as early forecasts. To assess the out-of-sample performance of each method, we use a 2-fold

CV, where we partition the 454 campaigns into two equal-sized sets and use one set to train the functional regression and the other to test the prediction performance, and then switch the role of the two sets. For each campaign t in the test set, given its prediction $\{\widehat{N}_t(r), r \in (s, 30]\}$ generated at time s, we calculate its RMSE and MAE with respect to the true value $\{N_t(r_i), r_i \in (s, 30]\}$, where

$$\mathrm{RMSE}_{t,s} = \sqrt{\frac{1}{\#\{r_i \in (s, 30]\}} \sum_{r_i \in (s,30]} (\widehat{N}_t(r_i) - N_t(r_i))^2}$$

and

$$\mathrm{MAE}_{t,s} = \frac{1}{\#\{r_i \in (s, 30]\}} \sum_{r_i \in (s,30]} \left| \widehat{N}_t(r_i) - N_t(r_i) \right|.$$

Figure 5 visualizes $\mathrm{RMSE}_s = T^{-1} \sum_{t=1}^{T} \mathrm{RMSE}_{t,s}$ and $\mathrm{MAE}_s = T^{-1} \sum_{t=1}^{T} \mathrm{MAE}_{t,s}$ achieved by different functional regression methods across $s \in \{7, 8, \dots, 20\}$. In general, the two RKHS-based estimators consistently achieve the best prediction accuracy. As expected, the performance of all methods improve with s approaching 20. Interestingly, the functional regression with mixed predictors does not seem to improve the prediction performance, which is especially evident for PFFR. Thanks to the group Lasso-type penalty, RKHS can perform variable selection on the vector covariate. Indeed, among the 28 (2 folds $\times$ 14 days) estimated functional regression models based on $\mathrm{RKHS}_{\mathrm{mixed}}$, 19 models select no vector covariate and thus reduce to the function-on-function regression, suggesting the potential irrelevance of the vector covariate. We further provide a robustness check of the above analysis in Appendix H, where we repeat the 2-fold CV procedure 100 times for RKHS, FDA and PFFR. It is seen that RKHS consistently provides the best performance, confirming the robustness of our findings. We refer to Appendix H for more details.

For more intuition, Figure 1 plots the normalized fundraising curves $\{N_t(r_i)\}_{i=1}^{60}$ of six representative campaigns and further visualizes the functional predictions given by RKHS, FDA and PFFR at $s = 14$th day. Note that FDA and RKHS provide more similar prediction while the prediction of PFFR seems to be more volatile. This is indeed consistent with the estimated bivariate coefficient functions visualized in Figure 6, where $\widehat{A}(r, s)$ of RKHS and FDA resembles each other while PFFR seems to suffer from under-smoothing.

Figure 11 in Appendix G further gives the prediction performance of the classical time series model ARIMA, which is significantly worse than the predictions given by RKHS, FDA and PFFR, suggesting the advantage of functional regression for handling the current application.

6. Conclusion

In this paper, we study a functional linear regression model with functional responses and accommodating both functional and high-dimensional vector covariates. We provide a minimax lower bound on its excess risk. To match the lower bound, we propose an RKHS-based penalized least squares estimator, which is based on a generalization of the representer lemma and achieves the optimal upper bound on the excess risk. Our framework allows for partially observed functional variables and provides finite sample probability bounds.

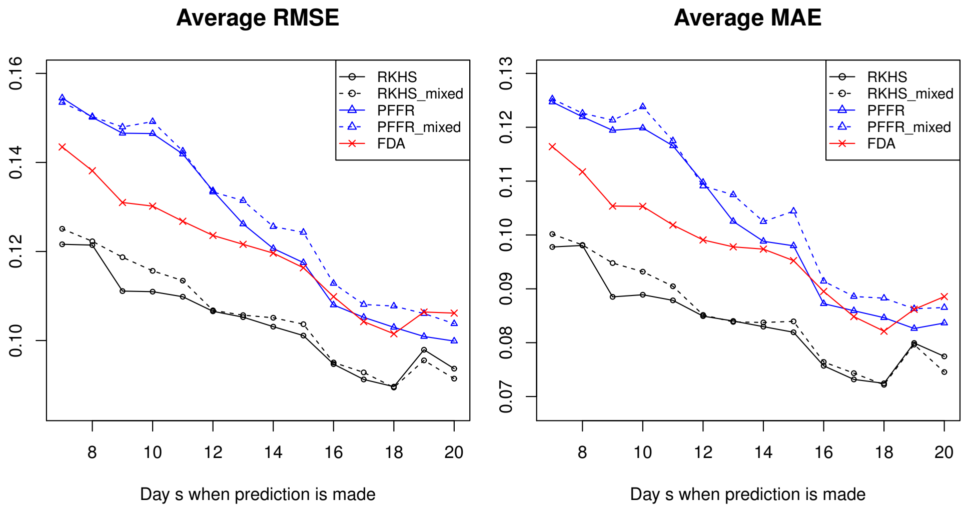

Figure 5: Average RMSE and MAE achieved by different functional regression methods.

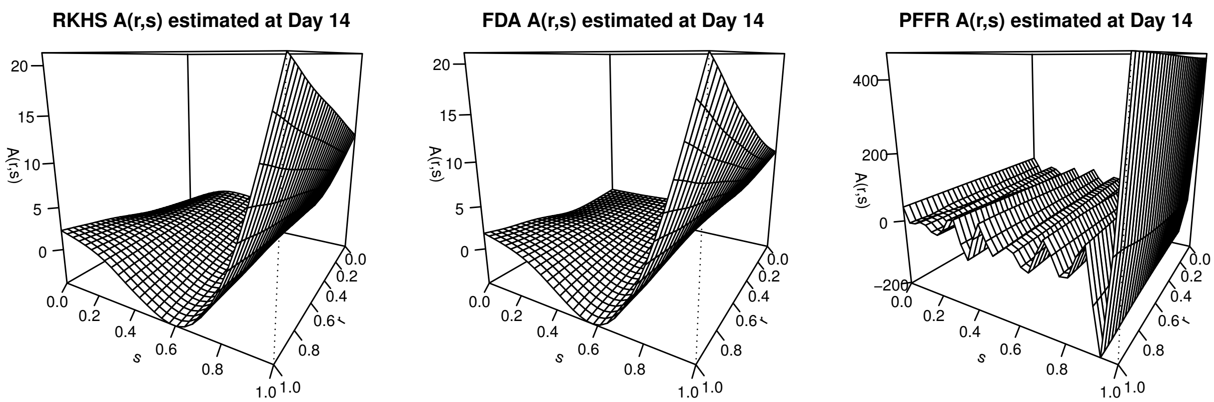

Figure 6: Estimated bivariate coefficient functions $\widehat{A}(r,s)$ by RKHS, FDA and PFFR at Day 14 ((r,s) is rescaled to $[0,1]^2$).

Furthermore, the result unveils an interesting phase transition between a high-dimensional parametric rate and a nonparametric rate in terms of the excess risk. A novel iterative coordinate descent based algorithm is proposed to efficiently solve the penalized least squares problem. Simulation studies and real data applications are further conducted, where the proposed estimator is seen to provide favorable performance compared to existing methods in the literature.

Throughout the paper, we assume knowledge of the kernel functions $\mathbb{K}$ and $\mathbb{K}_\beta$. In practice, it would be ideal if one would be able to learn the kernel functions from data. However, to our best knowledge, assuming knowing the true kernel functions is adopted in all RKHS-based functional data analysis literature. Learning kernels is beyond the scope of our paper and we would like to pursue this direction in the future research. We would also like to point out that our estimator $(\widehat{A}, \widehat{\beta})$ is a constrained estimator, with the constraints related with the true kernels. A mis-specification might lead to an over/under penalization, which will affect the final prediction error rate.

Finally, we briefly discuss important distinctions between non-parametric regression and functional regression for interested readers. To make the comparison easier, we consider a simpler setting: the scalar response functional linear regression and the classical non-parametric regression.

In non-parametric regression, we have

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

where $\{x_i\}_{i=1}^n$ are assumed to be random or fixed grid points in $[0, 1]$. In scalar response functional linear regression with fully observed functional data, we have

$$y_i = \int X_i(s)\beta(s)\,\mathrm{d}s + \epsilon_i, \tag{23}$$

where $\{X_i\}_{i=1}^n$ are assumed to be i.i.d. stochastic processes in $\mathcal{L}^2$. Note that we can treat X_i and β as infinite-dimensional vectors w.r.t. some $\mathcal{L}^2$ basis system, thus model (23) can be rewritten as

$$y_i = \widetilde{X}_i^\top \widetilde{\beta} + \epsilon_i,$$

where $\widetilde{X}_i \in \mathbb{R}^\infty, \widetilde{\beta} \in \mathbb{R}^\infty$. Under the standard assumption that $\mathbb{E}(\|X_i\|_{\mathcal{L}^2}^2) < \infty$, the eigenvalue sequence of the covariance matrix $\widetilde{\Sigma}_X = \mathbb{E}(\widetilde{X}_i \widetilde{X}_i^\top)$ converges to 0. As such, model (23) was described as a high-dimensional or infinitely-dimensional "ill-posed" linear regression model in Hall and Horowitz (2007), which is fundamentally different from the standard non-parametric regression.

Acknowledgments

We would like to thank the action editor, Dr. Pradeep Ravikumar, as well as the four anonymous reviewers for their thoughtful assessment and constructive comments which helped us to improve the quality and the presentation of our paper. Z.Z. would like to acknowledge support from NSF DMS-2014053. Y.Y. would like to acknowledge support from DMS-EPSRC EP/V013432/1 and EPSRC EP/W003716/1. R.W. would like to acknowledge support from AFOSR FA9550-18-1-0166, DOE DE-AC02-06CH113575, NSF DMS-1925101, ARO W911NF-17-1-0357, and NGA HM0476-17-1-2003.

References

Alexander Aue, Diogo Dubart Norinho, and Siegfried Hörmann. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110 (509):378–392, 2015.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Neil Bathia, Qiwei Yao, and Flavio Ziegelmann. Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, 38(6):3352–3386, 2010.

Michal Benko. *Functional data analysis with applications in finance*. Humboldt-Universität zu Berlin, Wirtschaftswissenschaftliche Fakultät, 2007.

Simon J Bonner, Nathaniel K Newlands, and Nancy E Heckman. Modeling regional impacts of climate teleconnections using functional data analysis. *Environmental and ecological statistics*, 21(1):1–26, 2014.

Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

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

T Tony Cai and Ming Yuan. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The annals of statistics*, 39(5):2330–2355, 2011.

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

Tony Cai and Ming Yuan. Nonparametric covariance function estimation for functional and longitudinal data. *University of Pennsylvania and Georgia inistitute of technology*, 2010.

Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, pages 571–591, 2003.

Kehui Chen, Pedro Francisco Delicado Useros, and Hans-Georg Müller. Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(1):177–196, 2017.

Jeng-Min Chiou, Yi-Chen Zhang, Wan-Hui Chen, and Chiung-Wen Chang. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics*, 2(2):106–129, 2014.

Xiongtao Dai, Hans-Georg Müller, Jane-Ling Wang, and Sean CL Deoni. Age-dynamic networks and functional correlation for early white matter myelination. *Brain Structure and Function*, 224(2):535–551, 2019.

Lawrence C Evans. *Partial differential equations. second. vol. 19.* Graduate Studies in Mathematics. American Mathematical Society, 2010.

Yingying Fan, Natasha Foutz, Gareth M James, and Wolfgang Jank. Functional response additive model estimation with online virtual stock markets. *The Annals of Applied Statistics*, 8(4):2435–2460, 2014.

Yingying Fan, Gareth M James, and Peter Radchenko. Functional additive regression. *The Annals of Statistics*, 43(5):2296–2325, 2015.

Julian J Faraway. Regression analysis for a functional response. *Technometrics*, 39(3): 254–261, 1997.

Frédéric Ferraty and Philippe Vieu. Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis*, 53(4):1400–1413, 2009.

Ricardo Fraiman, Ana Justel, Regina Liu, and Pamela Llop. Detecting trends in time series of functional data: A study of antarctic climate change. *Canadian Journal of Statistics*, 42(4):597–609, 2014.

Jerome H. Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv:1001.0736*, 2010.

Chong Gu. *Smoothing Spline ANOVA Models*. Springer-Verlag, New York, 2 edition, 2013.

Pantelis Z Hadjipantelis, John AD Aston, Hans-Georg Müller, and Jonathan P Evans. Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese. *Journal of the American Statistical Association*, 110(510):545–559, 2015.

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

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning*. Springer-Verlag New York, 2 edition, 2009.

Andrada E Ivanescu, Ana-Maria Staicu, Fabian Scheipl, and Sonja Greven. Penalized function-on-function regression. *Computational Statistics*, 30(2):539–568, 2015.

Ci-Ren Jiang, Wei Yu, Jane-Ling Wang, et al. Inverse regression for longitudinal data. *The Annals of Statistics*, 42(2):563–591, 2014.

Alois Kneip, Dominik Poß, and Pascal Sarda. Functional linear regression with points of impact. *The Annals of Statistics*, 44(1):1–30, 2016.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6):3660–3695, 2010.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Yehua Li and Tailen Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351, 2010.

Hua Liang, Hulin Wu, and Raymond J Carroll. The relationship between virologic and immunologic responses in aids clinical research using mixed-effects varying-coefficient models with measurement error. *Biostatistics*, 4(2):297–312, 2003.

Yi Lin and Ming Yuan. Convergence rates of compactly supported radial basis function regularization. *Statistica Sinica*, pages 425–439, 2006.

Po-Ling Loh and Martin J Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer, 2002.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2: 321–359, 2015.

Byeong U Park, Chun-Jui Chen, Wenwen Tao, and Hans-Georg Müller. Singular additive models for function to function regression. *Statistica Sinica*, 28(4):2497–2520, 2018.

Alexander Petersen, Sean Deoni, and Hans-Georg Müller. Fréchet estimation of time-varying covariance matrices from sparse data, with application to the regional co-evolution of myelination in the developing brain. *The Annals of Applied Statistics*, 13(1):393–419, 2019.

James O Ramsay and James B Ramsey. Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics*, 107(1-2):327–344, 2002.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer-Verlag, New York, 2005.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over lq -balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of machine learning research*, 13(2), 2012.

Sarah J Ratcliffe, Gillian Z Heller, and Leo R Leader. Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in medicine*, 21(8):1115–1127, 2002.

Matthew Reimherr, Bharath Sriperumbudur, Bahaeddine Taoufik, et al. Optimal prediction for additive function-on-function regression. *Electronic Journal of Statistics*, 12(2):4571–4601, 2018.

Matthew Reimherr, Bharath Sriperumbudur, and Hyun Bin Kang. Optimal function-on-scalar regression over complex domains. *arXiv preprint arXiv:1902.07284*, 2019.

Xiaoxiao Sun, Pang Du, Xiao Wang, and Ping Ma. Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association*, 113(524):1601–1611, 2018.

Shahin Tavakoli, Davide Pigoli, John AD Aston, and John S Coleman. A spatial modeling approach for linguistic object data: Analyzing dialect sound variations across great britain. *Journal of the American Statistical Association*, 114(527):1081–1096, 2019.

Carlos Valencia and Ming Yuan. Radial basis function regularization for linear inverse problems with random noise. *Journal of Multivariate Analysis*, 116:92–108, 2013.

Anne van Delft, Vaidotas Characiejus, and Holger Dette. A nonparametric test for stationarity in functional time series. *arXiv preprint arXiv:1708.05248*, 2017.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Isaac Michael Wagner-Muns, Ivan G Guardiola, VA Samaranayke, and Wasim Irshad Kayani. A functional data analysis approach to traffic volume forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):878–888, 2017.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Daren Wang, Zifeng Zhao, Rebecca Willett, and Chun Yip Yau. Functional autoregressive processes in reproducing kernel hilbert spaces. *arXiv preprint arXiv:2011.13993*, 2020a.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Jiayi Wang, Raymond KW Wong, and Xiaoke Zhang. Low-rank covariance function estimation for multidimensional functional data. *Journal of the American Statistical Association*, pages 1–14, 2020b.

Yu-Xiang Wang, Alex Smola, and Ryan Tibshirani. The falling factorial basis and its statistical applications. In *International Conference on Machine Learning*, pages 730–738. PMLR, 2014.

Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.

Colin O Wu and Chin-Tsang Chiang. Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, pages 433–456, 2000.

Colin O Wu, Chin-Tsang Chiang, and Donald R Hoover. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association*, 93(444):1388–1402, 1998.

Ming-Yi Yan. Economic data analysis: an approach based on functional viewpoint of data [j]. *Modern Economic Science*, 1, 2007.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, pages 2873–2903, 2005.

Bin Yu. Assouad, fano, and le cam. In Yang G.L. Pollard D., Torgersen E., editor, *Festschrift for Lucien Le Cam*, pages 423–435. Springer, New York, NY, 1997.

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

Appendices

Appendix A provides more discussions on the connection between bivariate functions and compact linear operators, Assumptions 2 and 3. Appendix B collects proofs of results in Section 3. Appendix C contains a series of deviation bounds, which are interesting *per se*. Appendix D contains the proof of Theorem 7. Appendix E includes additional details on technical proofs. Appendix F collects additional details on optimization. Additional numerical details are exhibited in Appendices G, H, I, J, K and L.

Appendix A. More discussions on bivariate functions, Assumptions 2 and 3

A.1 Bivariate functions and compact linear operators

Remark 9 *Fro any compact linear operator $A_2 : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$ denote*

$$A_2[f, g] = \langle A_2[g], f\rangle_{\mathcal{H}(\mathbb{K})}, \quad f, g \in \mathcal{H}(\mathbb{K}).$$

Note that $A_2[f, g]$ is well defined for any $f, g \in \mathcal{H}(\mathbb{K})$ due to the compactness of A_2. Define $a_{ij} = A_2[\psi_i, \psi_j] = \langle A_2[\psi_j], \psi_i\rangle_{\mathcal{H}(\mathbb{K})}$, $i, j \in \mathbb{N}_*$. We thus have for any $f, g \in \mathcal{H}(\mathbb{K})$, it holds that

$$\begin{aligned}
A_2[f, g] &= \langle A_2[g], f\rangle_{\mathcal{H}(\mathbb{K})} = \sum_{i=1}^{\infty}\langle f, \psi_i\rangle_{\mathcal{H}(\mathbb{K})}\langle A_2[g], \psi_i\rangle_{\mathcal{H}(\mathbb{K})} \\
&= \sum_{i=1}^{\infty}\langle f, \psi_i\rangle_{\mathcal{H}(\mathbb{K})}\left\langle A_2\left[\sum_{j=1}^{\infty}\langle g, \psi_j\rangle_{\mathcal{H}(\mathbb{K})}\psi_j\right], \psi_i\right\rangle_{\mathcal{H}(\mathbb{K})} \\
&= \sum_{i,j=1}^{\infty}\langle f, \psi_i\rangle_{\mathcal{H}(\mathbb{K})}\langle g, \psi_j\rangle_{\mathcal{H}(\mathbb{K})}\langle A_2[\psi_j], \psi_i\rangle_{\mathcal{H}(\mathbb{K})} = \sum_{i,j=1}^{\infty} a_{ij}\langle f, \psi_i\rangle_{\mathcal{H}(\mathbb{K})}\langle g, \psi_j\rangle_{\mathcal{H}(\mathbb{K})},
\end{aligned}$$

where the fourth identity follows from the compactness of A_2. This justifies Equation (2).

A.2 Assumption 2(d)

In Assumption 2(d), we assume that

$$\mathbb{E}(\langle X_1, f\rangle_{\mathcal{L}^2} Z_1^\top v) \leq \frac{3}{4}\sqrt{\mathbb{E}\{\langle X_1, f\rangle_{\mathcal{L}^2}^2\}(v^\top \Sigma_Z v)},$$

i.e. the functional and vector covariates are allowed to be correlated up to $O(1)$. In this subsection, we emphasize that the correlation cannot be equal to one. We first note that

$$\mathbb{E}(\langle X_1, f\rangle_{\mathcal{L}^2} Z_1^\top v) \leq \sqrt{\mathbb{E}\{\langle X_1, f\rangle_{\mathcal{L}^2}^2\}(v^\top \Sigma_Z v)}. \tag{24}$$

For model identification, we require the inequality in (24) to be strict. To see this, we suppose that there exist $g \in \mathcal{H}(\mathbb{K})$ and $u \in \mathbb{R}^p$ such that

$$\mathbb{E}(\langle X_1, g\rangle_{\mathcal{L}^2} Z_1^\top u) = \sqrt{\mathbb{E}\{\langle X_1, g\rangle_{\mathcal{L}^2}^2\}(u^\top \Sigma_Z u)}.$$

Since $\langle X_1, g\rangle_{\mathcal{L}^2}$ and $Z_1^\top u$ are both normal random variables, the above equality implies that the correlation between $\langle X_1, g\rangle_{\mathcal{L}^2}$ and $Z_1^\top u$ is 1 and thus $\langle X_1, g\rangle_{\mathcal{L}^2} = aZ_1^\top u$ for some constant $a \in \mathbb{R}$. This means that the model defined in (1) is no longer identifiable, because the covariates are perfectly correlated. Even in finite-dimensional regression problems, if the covariates are perfectly correlated, the solution to the linear system is ill-defined.

A.3 Assumption 3(b)

Recall that Assumption 3 is required for handling the case where the functional observations are on discrete sample points. Assumption 3(a) formalizes the sampling scheme and Assumption 3(b) requires that the second moment of the random variable $\|X\|_{W^{\alpha,2}}$ is finite.

In the following, via a concrete example, we show that the second moment assumption in (7) holds under mild conditions. Specifically in Example 1, we provide a simple sufficient condition on the eigen-decay of the covariance operator Σ_X, under which (7) holds for $W^{\alpha,2}$ with $\alpha > 1/2$.

Example 1 *Let $\alpha > 1/2$ and suppose the Sobolev space $W^{\alpha,2}$ is generated by the kernel*

$$\mathbb{K}_\alpha(r,s) = \sum_{k=1}^{\infty} \omega_k \psi_k^\alpha(r)\psi_k^\alpha(s),$$

where due to the property of the Sobolev space, we have $\omega_k \asymp k^{-2\alpha}$ and $\{\psi_k^\alpha\}_{k=1}^\infty$ can be taken as the Fourier basis such that $\|\psi_k^\alpha\|^2_{W^{\alpha,2}} \asymp k^{2\alpha}$. To better understand the second moment condition

$$\mathbb{E}(\|X_1\|^2_{W^{\alpha,2}}) < \infty,$$

we proceed by following the same strategy as in Cai and Hall (2006) and Yuan and Cai (2010) and assume that $\mathbb{K}_\alpha$ and Σ_X are perfectly aligned, i.e., they share the same set of eigen-functions.

Let $z_k = \langle X_1, \psi_k^\alpha\rangle_{\mathcal{L}^2}$. We have that $\{z_k\}_{k=1}^\infty$ is a collection of independent Gaussian random variables such that $z_k \sim N(0, \sigma_k^2)$, where $\{\sigma_k^2\}_{k=1}^\infty$ are the eigenvalues of Σ_X. Thus we have

$$\|X_1\|^2_{W^{\alpha,2}} = \sum_{k=1}^{\infty} z_k^2 \|\psi_k^\alpha\|^2_{W^{\alpha,2}} \asymp \sum_{k=1}^{\infty} z_k^2 k^{2\alpha},$$

and

$$\mathbb{E}(\|X_1\|^2_{W^{\alpha,2}}) \asymp \sum_{k=1}^{\infty} \mathbb{E}(z_k^2) k^{2\alpha} = \sum_{k=1}^{\infty} \sigma_k^2 k^{2\alpha}.$$

To assure that $\mathbb{E}(\|X_1\|^2_{W^{\alpha,2}}) < \infty$, it suffices to have that

$$\sigma_k^2 \asymp k^{-2\alpha-1-\nu}$$

for any constant $\nu > 0$. Note that when α is close to $1/2$, this means that the eigenvalues of Σ_X decay at a polynomial rate slightly faster than 2.

Appendix B. Proofs of the results in Section 3

B.1 Proof of Theorem 1

Proof [Proof of Theorem 1]

For any linear subspaces $R, S \subset \mathcal{H}(\mathbb{K})$, let $\mathcal{P}_R$ and $\mathcal{P}_S$ be the projection mappings of the spaces R and S, respectively with respect to $\|\cdot\|_{\mathcal{H}(\mathbb{K})}$. For any $f, g \in \mathcal{H}(\mathbb{K})$ and any compact linear operator $A : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$, denote

$$A|_{R\times S}[f, g] = A[\mathcal{P}_R f, \mathcal{P}_S g].$$

Let $(\widehat{B}, \widehat{\alpha})$ be any solution to (4). Let $Q = \text{span}\{\mathbb{K}_\beta(r_j, \cdot)\}_{j=1}^n \subset \mathcal{H}(\mathbb{K}_\beta)$, $R = \text{span}\{\mathbb{K}(r_j, \cdot)\}_{j=1}^n \subset \mathcal{H}(\mathbb{K})$ and $S = \text{span}\{\mathbb{K}(s_j, \cdot)\}_{j=1}^n \subset \mathcal{H}(\mathbb{K})$. Denote $\widehat{\beta}_l = \mathcal{P}_Q(\widehat{\alpha}_l)$, $l \in \{1, \ldots, p\}$, and $\widehat{A}[\cdot, \cdot] = \widehat{B}|_{R\times S}[\cdot, \cdot] = \widehat{B}[\mathcal{P}_R\cdot, \mathcal{P}_S\cdot]$.

Let $S^\perp$ and $R^\perp$ be the orthogonal complements of S and R in $\mathcal{H}$, respectively. Then for any compact linear operator A, we have the decomposition

$$A = A|_{R\times S} + A|_{R\times S^\perp} + A|_{R^\perp\times S} + A|_{R^\perp\times S^\perp}.$$

Due to Theorem 10, there exist $\{a_{ij}\}_{i,j=1}^n \subset \mathbb{R}$ such that

$$A|_{R\times S}[f, g] = \sum_{i,j=1}^n a_{ij} \langle \mathbb{K}(r_i, \cdot), f\rangle_{\mathcal{H}(\mathbb{K})} \langle \mathbb{K}(s_j, \cdot), g\rangle_{\mathcal{H}(\mathbb{K})}.$$

Step 1. In this step, it is shown that for any compact linear operator A, its associated bivariate function $A(\cdot, \cdot)$ satisfies that $\{A(r_i, s_j)\}_{i,j=1}^n$ only depend on $A|_{R\times S}$. The details of the bivariate function is explained in Appendix A.1.
Observe that

$$A|_{R\times S^\perp}(r_i, s_j) = A|_{R\times S^\perp}[\mathbb{K}(r_i, \cdot), \mathbb{K}(s_j, \cdot)] = A[\mathcal{P}_R\mathbb{K}(r_i, \cdot), \mathcal{P}_{S^\perp}\mathbb{K}(s_j, \cdot)] = 0.$$

Similar arguments also lead to that $A|_{R^\perp\times S}[\mathbb{K}(r_i, \cdot), \mathbb{K}(s_j, \cdot)] = 0$ and $A|_{R^\perp\times S^\perp}[\mathbb{K}(r_i, \cdot), \mathbb{K}(s_j, \cdot)] = 0$, $i, j = 1, \ldots, n$.

Step 2. By **Step 1**, it holds that $\widehat{A}(r_i, s_j) = \widehat{B}(r_i, s_j)$, $i, j = 1, \ldots, n$. By Theorem 10, we have that there exist $\{\widehat{a}_{ij}\}_{i,j=1}^n \subset \mathbb{R}$ such that for any $f, g \in \mathcal{H}(\mathbb{K})$,

$$\widehat{A}[f, g] = \sum_{i,j=1}^n \widehat{a}_{ij} \langle \mathbb{K}(r_i, \cdot), f\rangle_{\mathcal{H}(\mathbb{K})} \langle \mathbb{K}(s_j, \cdot), g\rangle_{\mathcal{H}(\mathbb{K})}.$$

Therefore the associated bivariate function satisfies that, for all $r, s \in [0, 1]$,

$$\begin{aligned}\widehat{A}(r, s) &= \sum_{i,j=1}^\infty \widehat{A}[\psi_i, \psi_j]\psi_i(r)\psi_j(s) = \sum_{i,j=1}^\infty \sum_{k,l=1}^n \widehat{a}_{kl}\psi_i(r_k)\psi_j(s_l)\psi_i(r)\psi_j(s) \\ &= \sum_{k,l=1}^n \widehat{a}_{kl} \left\{\sum_{i=1}^\infty \psi_i(r)\psi_i(r_k)\right\} \left\{\sum_{i=1}^\infty \psi_i(s)\psi_i(s_l)\right\} = \sum_{k,l=1}^n \widehat{a}_{kl}\mathbb{K}(r, r_k)\mathbb{K}(s, s_l),\end{aligned}$$

where $K(r,s) = \sum_{i=1}^{\infty} \psi(s)\psi(r)$ is used in the last inequality.

Step 3. For any $j \in \{1, \dots, n\}$ and $l \in \{1, \dots, p\}$, by the definition of $\widehat{\beta}_l$, we have that

$$\widehat{\alpha}_l(r_j) = \langle \widehat{\alpha}_l, \mathbb{K}_\beta(r_j, \cdot) \rangle_{\mathcal{H}} = \langle \widehat{\beta}_l, \mathbb{K}_\beta(r_j, \cdot) \rangle_{\mathcal{H}} = \widehat{\beta}_l(r_j).$$

Therefore

$$\begin{aligned} &\frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} w_r(i) \left\{ Y_t(r_i) - \frac{1}{n} \sum_{j=1}^{n} w_s(j) \widehat{A}(r_i, s_j) X_t(s_j) - \langle \widehat{\beta}(r_i), Z_t \rangle_p \right\}^2 + \lambda \sum_{l=1}^{p} \|\widehat{\beta}_l\|_n \\ =&\frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} w_r(i) \left\{ Y_t(r_i) - \frac{1}{n} \sum_{j=1}^{n} w_s(j) \widehat{B}(r_i, s_j) X_t(s_j) - \langle \widehat{\alpha}(r_i), Z_t \rangle_p \right\}^2 + \lambda \sum_{l=1}^{p} \|\widehat{\alpha}_l\|_n. \end{aligned}$$

In addition, by Theorem 10 we have that

$$\|\widehat{A}\|_{\mathrm{F}(\mathbb{K})} \le \|\widehat{B}\|_{\mathrm{F}(\mathbb{K})} \quad \text{and} \quad \|\widehat{\beta}_l\|_{\mathcal{H}(\mathbb{K}_\beta)} \le \|\widehat{\alpha}_l\|_{\mathcal{H}(\mathbb{K}_\beta)} \quad \text{for all } l \in \{1, \dots, p\},$$

which completes the proof. ■

Lemma 10 *Let $A : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$ be any Hilbert–Schmidt operator. Let $\{v_i, v_i'\}_{i=1}^m \subset \mathcal{H}(\mathbb{K})$ be a collection of functions in $\mathcal{H}(\mathbb{K})$. Let R and S be the subspaces of $\mathcal{H}(\mathbb{K})$ spanned by $\{v_i\}_{i=1}^m$ and $\{v_i'\}_{i=1}^m$, respectively. Let $\mathcal{P}_R$ and $\mathcal{P}_S$ be the projection mappings of the spaces R and S, respectively. Then there exist $\{a_{ij}\}_{i,j=1}^m \subset \mathbb{R}$, which are not necessarily unique, such that for any $f, g \in \mathcal{H}(\mathbb{K})$,*

$$A|_{R \times S}[f, g] = A[\mathcal{P}_R f, \mathcal{P}_S g] = \sum_{i,j=1}^{m} a_{ij} \langle f, v_i \rangle_{\mathcal{H}(\mathbb{K})} \langle g, v_i \rangle_{\mathcal{H}(\mathbb{K})},$$

and that $\|A|_{R\times S}[f, g]\|_{\mathrm{F}(\mathbb{K})} \le \|A\|_{\mathrm{F}(\mathbb{K})}$.

Proof Let $\{u_k\}_{k=1}^K$ and $\{u_l'\}_{l=1}^L$ be orthogonal basis of subspaces R and S of $\mathcal{L}^2$, respectively, with $K, L \le m$. Since each $u_i(u_i')$ can be written as a linear combination of $\{v_k\}_{k=1}^m(\{v_k'\}_{k=1}^m)$, it suffices to show that there exist $\{b_{ij}\}_{i=1,j=1}^{K,L} \subset \mathbb{R}$, such that for any $f, g \in \mathcal{H}(\mathbb{K})$,

$$A|_{R\times S}[f, g] = \sum_{k=1}^{K} \sum_{l=1}^{L} b_{ij} \langle u_i, f \rangle_{\mathcal{H}(\mathbb{K})} \langle u_j', g \rangle_{\mathcal{H}(\mathbb{K})}.$$

Since R and S are linear subspaces of $\mathcal{H}(\mathbb{K})$, there exist $\{u_k\}_{k=K+1}^{\infty}, \{u_l'\}_{l=L+1}^{\infty} \subset \mathcal{H}(\mathbb{K})$, such that $\{u_k\}_{k=1}^{\infty}$ and $\{u_l'\}_{l=1}^{\infty}$ are two orthogonal basis of $\mathcal{H}(\mathbb{K})$, and that

$$A[f, g] = \sum_{i,j=1}^{\infty} b_{ij} \langle u_i, f \rangle_{\mathcal{H}(\mathbb{K})} \langle u_j', g \rangle_{\mathcal{H}(\mathbb{K})},$$

where $b_{ij} = A[u_i, u'_j] = \langle A[u'_j], u_i \rangle_{\mathcal{H}(\mathbb{K})}$. Therefore

$$\begin{aligned} A|_{R\times S}[f,g] &= \sum_{i,j=1}^{\infty} b_{ij} \langle u_i, \mathcal{P}_R f\rangle_{\mathcal{H}(\mathbb{K})} \langle u'_j, \mathcal{P}_S g\rangle_{\mathcal{H}(\mathbb{K})} = \sum_{i,j=1}^{\infty} b_{ij} \langle \mathcal{P}_R u_i, f\rangle_{\mathcal{H}(\mathbb{K})} \langle \mathcal{P}_S u'_j, g\rangle_{\mathcal{H}(\mathbb{K})} \\ &= \sum_{k=1}^{K}\sum_{l=1}^{L} b_{kl} \langle u_k, f\rangle_{\mathcal{H}(\mathbb{K})} \langle u'_l, g\rangle_{\mathcal{H}(\mathbb{K})}. \end{aligned}$$

Moreover, we have that

$$\|A\|^2_{\mathrm{F}(\mathbb{K})} = \sum_{k,l=1}^{\infty} b_{kl}^2 \geq \sum_{k=1}^{K}\sum_{l=1}^{L} b_{kl}^2 = \|A|_{R\times S}\|^2_{\mathrm{F}(\mathbb{K})},$$

which concludes the proof. ■

B.2 Proof of Theorem 4

Proof [Proof of Theorem 4]

First note that $0 < \delta_T, \zeta_n < 1$, then $\delta_T^2 \leq \delta_T$ and $\zeta_n^2 \leq \zeta_n$. These inequalities will be used repeatedly in the rest of this proof. Recall the notation that $\Delta_\beta(r) = \widehat{\beta}(r) - \beta^*(r)$ and $\Delta_A(r,s) = \widehat{A}(r,s) - A^*(r,s)$, $r, s \in [0,1]$. Note that for any $g \in \mathcal{L}^2$, it holds that

$$\mathbb{E}(\langle g, X^*\rangle^2_{\mathcal{L}^2}) = \iint_{[0,1]^2} g(r)\Sigma_X(r,s)g(s)\,\mathrm{d}r\,\mathrm{d}s.$$

The above expression will also be used repeatedly in the rest of the proof. Observe that for any $r \in [0,1]$, it holds that

$$\|\Delta_A(r,\cdot)\|_{W^{\alpha,2}} \leq \|\Delta(r,\cdot)\|_{\mathcal{H}(\mathbb{K})} \leq C_{\mathbb{K}}\|\Delta\|_{\mathrm{F}(\mathbb{K})} \leq C_{\mathbb{K}} C_A,$$

where the first inequality follows from Assumption 3 **b**, which assumes that $\mathcal{H}(\mathbb{K}) \subset W^{\alpha,2}$, the second inequality follows from Theorem 36 and the last inequality follows from the assumption that $A^*, \widehat{A} \in \mathcal{C}_A$. Throughout the proof, we assume without loss of generality that $C_{\mathbb{K}} = 1$. In addition, observe that by Theorem 33, it holds that

$$\mathbb{P}\big(\|X_t\|_{W^{\alpha,2}} \geq 4C_X\sqrt{\log(T)}\big) \leq T^{-6}.$$

By a union bound argument,

$$\mathbb{P}(\mathcal{D}) = \mathbb{P}\big(\|X_t\|_{W^{\alpha,2}} \geq 4C_X\sqrt{\log(T)} \text{ for all } 1 \leq t \leq T\big) \leq T^{-5}.$$

The rest of the proof is shown in the event of $\mathcal{D}$.

Let $\lambda = C_\lambda\sqrt{\frac{\log(p\vee T)}{T}}$ for sufficiently large constant C_λ. From the minimizer property, we have that

$$\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{Y_t(r_i) - \frac{1}{n}\sum_{j=1}^{n} w_s(j)\widehat{A}(r_i,s_j)X_t(s_j) - \langle Z_t, \widehat{\beta}(r_i)\rangle_p\right\}^2 + \lambda\sum_{j=1}^{p}\|\widehat{\beta}_j\|_n$$

$$\leq \frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{Y_t(r_i) - \frac{1}{n}\sum_{j=1}^{n} w_s(j)A^*(r_i,s_j)X_t(s_j) - \langle Z_t, \beta^*(r_i)\rangle_p\right\}^2 + \lambda\sum_{j=1}^{p}\|\beta_j^*\|_n,$$

which implies that

$$\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j)\right\}^2 \tag{25}$$

$$+\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\langle\Delta_\beta(r_i), Z_t\rangle_p^2 \tag{26}$$

$$+\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j)\langle\Delta_\beta(r_i), Z_t\rangle_p\right\} \tag{27}$$

$$\leq\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left(\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j) + \langle\Delta_\beta(r_i), Z_t\rangle_p\right)\epsilon_t(r_i) \tag{28}$$

$$+\lambda\sum_{j=1}^{p}\|\beta_j^*\|_n - \lambda\sum_{j=1}^{p}\|\widehat{\beta}_j\|_n. \tag{29}$$

In the rest of the proof, for readability, we will assume that

$$w_r(i) = 1 \text{ for all } 1 \leq i \leq n \quad \text{and} \quad w_s(j) = 1 \text{ for all } 1 \leq j \leq n.$$

This assumption is equivalent to the case that $\{r_i\}_{i=1}^n$ and $\{s_j\}_{j=1}^n$ are equally spaced grid. We note that in view of Theorem 40, the general case follows straightforwardly from the same argument and thus will be omitted for readability.

Step 1. Observe that

$$(27) = \frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j) - \int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s\right\}\langle\Delta_\beta(r_i), Z_t\rangle_p \tag{30}$$

$$+\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s\right\}\langle\Delta_\beta(r_i), Z_t\rangle_p. \tag{31}$$

As for the term (30), we have that for any $j \in \{1,\ldots,n\}$ and $t \in \{1,\ldots,T\}$,

$$\|\Delta_A(r_j,\cdot)X_t(\cdot)\|_{W^{\alpha,2}} \leq \|\Delta_A(r_j,\cdot)\|_{W^{\alpha,2}}\|X_t\|_{W^{\alpha,2}} \leq 8C_AC_X\sqrt{\log(T)},$$

where the first inequality follows from Theorem 41 and event $\mathcal{D}$. By Theorem 40, it holds that for all $1 \leq i \leq n, 1 \leq t \leq T$,

$$\left|\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j) - \int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s\right| \leq C_1\zeta_n\sqrt{\log(T)}$$

where $\zeta_n = n^{-\alpha+1/2}$ and C_1 depends on C_A and C_X only. So,

$$\begin{aligned}
(30) \geq & -\frac{C_1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\zeta_n\sqrt{\log(T)}\,|\langle\Delta_\beta(r_i), Z_t\rangle_p| \\
\geq & -C_1\zeta_n\sqrt{\log(T)}\sqrt{\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\Delta_\beta(r_i), Z_t\rangle_p^2} \\
\geq & -\frac{1}{640Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\Delta_\beta(r_i), Z_t\rangle_p^2 - 640C_1^2\zeta_n^2\log(T) \\
= & -\frac{1}{640Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle\Delta_\beta(r_i), Z_t\rangle_p^2 - C_1'\zeta_n^2\log(T),
\end{aligned}$$

for some absolute constant $C_1' > 0$.

In addition, for Equation (31), note that with probability at least $1-(p\vee T)^{-4}$,

$$\begin{aligned}
&\left\{\int_{[0,1]}\Delta_A(r_i,s)\frac{1}{T}\sum_{t=1}^{T}X_t(s)Z_t^\top\Delta_\beta(r_i)\,\mathrm{d}s\right\} - \left\{\int_{[0,1]}\Delta_A(r_i,s)\mathbb{E}\{X_t(s)Z_t^\top\}\Delta_\beta(r_i)\,\mathrm{d}s\right\} \\
=&\int_{[0,1]}\Delta_A(r_i,s)\left(\frac{1}{T}\sum_{t=1}^{T}X_t(s)Z_t^\top - \mathbb{E}\{X_t(s)Z_t^\top\}\right)\Delta_\beta(r_i)\,\mathrm{d}s \\
\geq& -\|\Delta_A(r_i,\cdot)\|_{\mathcal{L}^2}\left\|\left(\frac{1}{T}\sum_{t=1}^{T}X_t(s)Z_t^\top - \mathbb{E}\{X_t(s)Z_t^\top\}\right)\Delta_\beta(r_i)\right\|_{\mathcal{L}^2} \\
\geq& -2C_A\sum_{1\leq j\leq p}\left\|\frac{1}{T}\sum_{t=1}^{T}X_t(s)Z_{t,j}^\top - \mathbb{E}\{X_t(s)Z_{t,j}^\top\}\right\|_{\mathcal{L}^2}|\Delta_{\beta_j}(r_i)| \\
\geq& -2C_A\sum_{1\leq j\leq p}|\Delta_{\beta_j}(r_i)|\max_{1\leq j\leq p}\left\|\frac{1}{T}\sum_{t=1}^{T}X_t(s)Z_{t,j}^\top - \mathbb{E}\{X_t(s)Z_{t,j}^\top\}\right\|_{\mathcal{L}^2} \\
\geq& -C_1'\sum_{1\leq j\leq p}|\Delta_{\beta_j}(r_i)|\sqrt{\frac{\log(p)}{T}} \geq -\frac{\lambda}{200}\sum_{1\leq j\leq p}|\Delta_{\beta_j}(r_i)|, \qquad (32)
\end{aligned}$$

where the second inequality holds because $\|\Delta_A(r_i,\cdot)\|_{\mathcal{L}^2} \leq \|\Delta_A(r_i,\cdot)\|_{\mathcal{H}(\mathbb{K})} \leq 2C_A$, the fourth inequality follows from Theorem 34, and the last inequality follows from the assumption that $\lambda = C_\lambda\sqrt{\frac{\log(p\vee T)}{T}}$ with sufficiently large C_λ. In addition, by Assumption 2(c),

$$\begin{aligned}
&\int_{[0,1]} -\Delta_A(r_i,s)\mathbb{E}\{X_t(s)Z_t^\top\}\Delta_\beta(r_i)\,\mathrm{d}s \\
\leq&\frac{3}{4}\sqrt{\Sigma_X[\Delta_A(r_i,\cdot),\Delta_A(r_i,\cdot)]}\sqrt{\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)} \\
\leq&\frac{3}{8}\left(\frac{6}{5}\Sigma_X[\Delta_A(r_i,\cdot),\Delta_A(r_i,\cdot)] + \frac{5}{6}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)\right)
\end{aligned}$$

$$
\begin{aligned}
=&\frac{9}{20}\Sigma_X[\Delta_A(r_i,\cdot),\Delta_A(r_i,\cdot)] + \frac{5}{16}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)\\
=&\frac{9}{20}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + \frac{5}{16}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i).
\end{aligned}
$$

So

$$
\begin{aligned}
&(31)\\
\geq & -\frac{\lambda}{100}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p|\Delta_{\beta_j}(r_i)| - \frac{1}{n}\sum_{i=1}^n\frac{9}{10}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s - \frac{1}{n}\sum_{i=1}^n\frac{5}{8}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)\\
\geq & -\frac{\lambda}{100}\sum_{j=1}^p\|\Delta_{\beta_j}\|_n - \frac{1}{n}\sum_{i=1}^n\frac{9}{10}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s - \frac{1}{n}\sum_{i=1}^n\frac{5}{8}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)\\
= & -\frac{\lambda}{100}\big(\sum_{j\in S}\|\Delta_{\beta_j}\|_n + \sum_{j\in S^c}\|\widehat{\beta}_j\|_n\big)\\
- & \frac{1}{n}\sum_{i=1}^n\frac{9}{10}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s - \frac{1}{n}\sum_{i=1}^n\frac{5}{8}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)
\end{aligned}
$$

Putting calculations for Equation (30) and Equation (31) together, we have that

$$
\begin{aligned}
(27) \geq & -\frac{1}{640Tn}\sum_{t=1}^T\sum_{i=1}^n\langle\Delta_\beta(r_i),Z_t\rangle_p^2 - C_1'\zeta_n^2\log(T) - \frac{\lambda}{100}\big(\sum_{j\in S}\|\Delta_{\beta_j}\|_n + \sum_{j\in S^c}\|\widehat{\beta}_j\|_n\big)\\
&-\frac{1}{n}\sum_{i=1}^n\frac{9}{10}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s - \frac{1}{n}\sum_{i=1}^n\frac{5}{8}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i). \quad (33)
\end{aligned}
$$

Step 2. It follows from (43) that with probability at least $1-T^{-4}$, it holds that

$$
(25) \geq \frac{159}{160n}\sum_{i=1}^n\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s - C_2\left(\log(T)\zeta_n^2 + \log(T)\delta_T\right), \quad (34)
$$

where $C_2 > 0$ is an absolute constant.

It follows from (47) and (51), we have that with probability at least $1-2T^{-4}$, it holds that

$$
\begin{aligned}
(28) \leq & \frac{1}{320n}\sum_{i=1}^n\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + C_2'\left\{\zeta_n + \log(T)\delta_T\right\}\\
&+\frac{\lambda}{320}\sum_{j\in S}\|\Delta_{\beta_j}\|_n + \frac{\lambda}{320}\sum_{j\in S^c}\|\widehat{\beta}_j\|_n
\end{aligned}
$$

where $C_2' > 0$ is an absolute constant.

Step 3. Note that for Equation (29),

$$
\lambda\sum_{j=1}^p\|\beta_j^*\|_n - \sum_{j=1}^p\lambda\|\widehat{\beta}_j\|_n = \lambda\left\{\sum_{j\in S}\|\beta_j^*\|_n - \sum_{j\in S}\|\widehat{\beta}_j\|_n - \sum_{j\in S^c}\|\widehat{\beta}_j\|_n\right\}
$$

$$\leq \lambda \sum_{j \in S} \|\Delta_{\beta_j}\|_n - \lambda \sum_{j \in S^c} \|\widehat{\beta}_j\|_n.$$

Step 4. Putting all previous calculations together leads to that

$$\begin{aligned}
&\frac{31}{320n} \sum_{i=1}^{n} \iint_{[0,1]^2} \Delta_A(r_i, r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s \\
&+\frac{639}{640Tn} \sum_{t=1}^{T}\sum_{i=1}^{n} \langle \Delta_\beta(r_i), Z_t\rangle_p^2 - \frac{5}{8n}\sum_{i=1}^{n} \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) \qquad (35)\\
&\leq \big(1 + \frac{1}{100} + \frac{1}{320}\big)\lambda \sum_{j\in S} \|\Delta_{\beta_j}\|_n - \lambda\big(1 - \frac{1}{100} - \frac{1}{320}\big) \sum_{j \in S^c} \|\widehat{\beta}_j\|_n + C_3\big\{ \log(T)\zeta_n + \log(T)\delta_T\big\},\\
&\leq 2\lambda \sum_{j\in S} \|\Delta_{\beta_j}\|_n + C_3\big\{ \log(T)\zeta_n + \log(T)\delta_T\big\} \qquad (36)
\end{aligned}$$

where $C_3 > 0$ is some absolute constant.

For Equation (35), note that by constrain set $\mathcal{C}_\beta$, we have that

$$\sum_{1\leq j \leq p} \|\widehat{\beta}_j\|_n \leq \sum_{1\leq j \leq p} \|\widehat{\beta}_j\|_\infty \leq \sum_{1\leq j \leq p} \|\widehat{\beta}_j\|_{\mathcal{H}(\mathbb{K}_\beta)} \leq C_\beta.$$

Therefore

$$\sum_{1\leq j \leq p} \|\Delta_{\beta_j}\|_n \leq \sum_{1\leq j \leq p} \|\widehat{\beta}_j\|_n + \sum_{1\leq j \leq p} \|\beta^*_j\|_n \leq 2C_\beta.$$

From Theorem 27, it holds that with probability at least $1 - \exp(-cT)$,

$$\begin{aligned}
\frac{639}{640T} \sum_{t=1}^{T} \langle \Delta_\beta(r_i), Z_t\rangle_p^2 \geq& \frac{639}{640} \cdot \frac{2}{3} \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) - C_3' \frac{\log(p)}{T}\Big(\sum_{1\leq j\leq p} \|\Delta_{\beta_j}\|_n\Big)^2\\
\geq& \frac{213}{320} \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) - 4C_3' C_\beta^2 \frac{\log(p)}{T}.
\end{aligned}$$

Substitute the above inequality into Equation (35) gives

$$\begin{aligned}
&\frac{31}{320n} \sum_{i=1}^{n} \iint_{[0,1]^2} \Delta_A(r_i, r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s \\
&+\frac{13}{320n}\sum_{i=1}^{n} \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i)\\
&\leq 2\lambda \sum_{j\in S} \|\Delta_{\beta_j}\|_n \qquad (37)\\
&+C_3''\big\{ \log(T)\zeta_n + \log(T)\delta_T + \frac{\log(p)}{T}\big\}.
\end{aligned}$$

Step 5. Note that by Assumption 2**c**,

$$\left(\lambda\sum_{j\in S}\|\Delta_{\beta_j}\|_n\right)^2 \leq \lambda^2\mathfrak{s}\sum_{j=1}^p\|\Delta_{\beta_j}\|_n^2 = \lambda^2\mathfrak{s}\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p\{\Delta_{\beta_j}(r_i)\}^2 \leq \frac{\lambda^2\mathfrak{s}}{c_z}\frac{1}{n}\sum_{i=1}^n\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i),$$

where $\mathfrak{s} = |S|$, the cardinality of the support set S. This gives

$$\sum_{j\in S}\|\Delta_{\beta_j}\|_n \leq \sqrt{\frac{\mathfrak{s}}{c_z}\frac{1}{n}\sum_{i=1}^n\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)}, \tag{38}$$

which implies that

$$\begin{aligned}\frac{1}{160}\lambda\sum_{j\in S}\|\Delta_{\beta_j}\|_n \leq& \frac{1}{320n}\sum_{i=1}^n\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i) + C_4\mathfrak{s}\lambda^2\\ \leq& \frac{1}{320n}\sum_{i=1}^n\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i) + C_4C_\lambda^2\frac{\mathfrak{s}\log(p\vee T)}{T}.\end{aligned}$$

Substituting the above inequality into (37) gives

$$\begin{aligned}&\frac{31}{320n}\sum_{i=1}^n\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + \frac{12}{320n}\sum_{i=1}^n\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i)\\ \leq& C_4'\Big\{\log(T)\zeta_n + \log(T)\delta_T + \frac{\log(p)}{T} + \frac{\mathfrak{s}\log(p\vee T)}{T}\Big\}\end{aligned} \tag{39}$$

Step 6. Note that by definition, $\mathcal{E}^*(\widehat{A},\widehat{\beta}) \geq 0$ for any estimators $\widehat{A}$ and $\widehat{\beta}$, and that

$$\begin{aligned}\mathcal{E}^*(\widehat{A},\widehat{\beta}) \leq& 4\int_{[0,1]}\mathbb{E}\left\{\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s\right\}^2\mathrm{d}r + 4\mathbb{E}\int_{[0,1]}\left\{(Z_i^*)^\top\Delta_\beta(r)\right\}^2\mathrm{d}r\\ =& 4\int_{[0,1]}\left\{\iint_{[0,1]^2}\Delta_A(r,s)\Sigma_X(s,u)\Delta_A(r,u)\,\mathrm{d}s\,\mathrm{d}u + \Delta_\beta^\top(r)\Sigma_Z\Delta_\beta(r)\right\}\mathrm{d}r,\end{aligned} \tag{40}$$

where Σ_X and Σ_Z are the covariance of X_t and Z_t, as defined in Assumption 2.

It follows from (54) that with probability at least $1 - T^{-4}$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\iint_{[0,1]^2}\Delta_A(r_i,r)^\top\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s\\ =& \mathbb{E}_{X^*}\left\{\frac{1}{n}\sum_{i=1}^n\left(\int_{[0,1]}\Delta_A(r_i,s)X^*(s)\,\mathrm{d}s\right)^2\right\}\end{aligned}$$

$$\geq c_5 \mathbb{E}_{X^*} \left\{ \int_{[0,1]} \left(\int_{[0,1]} \Delta_A(s,r) X^*(r) \, \mathrm{d}r \right)^2 \mathrm{d}s \right\} - C_5 \log(T) \zeta_n, \tag{41}$$

with $c_5, C_5 > 0$ being absolute constants. In addition, with absolute constants $c_5', C_5' > 0$, due to (55), we have that

$$\frac{1}{n} \sum_{i=1}^n \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) \geq c_5' \int_{[0,1]} \Delta_\beta(r)^\top \Sigma_Z \Delta_\beta(r) \, \mathrm{d}r - C_5' \zeta_n \tag{42}$$

Finally, (39), (40), (41) and (42) together imply that with probability at least $1 - 6T^{-4}$, it holds that

$$\mathcal{E}^*(\widehat{A}, \widehat{\beta}) \leq C_5'' \big\{ \log(T) \zeta_n + \log(T) \delta_T + \frac{\log(p)}{T} + \frac{\mathfrak{s} \log(p \vee T)}{T} \big\}.$$

where C is some absolute constant. This directly leads to the desired result. ■

B.3 Additional proofs related to Theorem 4

In this section, we present the technical results related to Theorem 4. Recall in the proof of Theorem 4, we set that $\Delta_\beta(r) = \widehat{\beta}(r) - \beta^*(r)$ and $\Delta_A(r,s) = \widehat{A}(r,s) - A^*(r,s)$, $r, s \in [0,1]$. Note that for any $g \in \mathcal{L}^2$, it holds that

$$\mathbb{E}(\langle g, X^* \rangle_{\mathcal{L}^2}^2) = \iint_{[0,1]^2} g(r) \Sigma_X(r,s) g(s) \, \mathrm{d}r \, \mathrm{d}s.$$

We also assume that the following good event holds:

$$\|X_t\|_{W^{\alpha,2}} \leq 4 C_X \sqrt{\log(T)} \text{ for all } 1 \leq t \leq T.$$

It was justified in the proof of Theorem 4 that

$$\mathbb{P}(\|X_t\|_{W^{\alpha,2}} \leq 4 C_X \sqrt{\log(T)} \text{ for all } 1 \leq t \leq T) \geq 1 - T^{-5}.$$

Lemma 11 *Let $\zeta_n = n^{-\alpha+1/2}$. Under the same conditions as in Theorem 4, with probability at least $1 - T^{-4}$, it holds that*

$$\begin{aligned} &\frac{1}{Tn} \sum_{t=1}^T \sum_{i=1}^n \left\{ \frac{1}{n} \sum_{j=1}^n \Delta_A(r_i, s_j) X_t(s_j) \right\}^2 \\ \geq &\frac{159}{160n} \sum_{i=1}^n \iint_{[0,1]^2} \Delta_A(r_i, r) \Sigma(r,s) \Delta_A(r_i, s) \, \mathrm{d}r \, \mathrm{d}s - C_2 \left(\log(T) \zeta_n^2 + \log(T) \delta_T \right). \end{aligned} \tag{43}$$

Proof Let $\zeta_n = n^{-\alpha+1/2}$ and $\delta_T = T^{-2r/(2r+1)}$. Observe that for any $i \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, we have

$$\|\Delta_A(r_i, \cdot) X_t(\cdot)\|_{W^{\alpha,2}} \leq \|\Delta_A(r_i, \cdot)\|_{W^{\alpha,2}} \|X_t(\cdot)\|_{W^{\alpha,2}} \leq 8 C_A C_X \sqrt{\log(T)},$$

where the first inequality follows Theorem 41. By Theorem 40, we have that

$$\left|\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j)-\int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s\right|\le C_4\zeta_n\sqrt{\log(T)},$$

where $C_4>0$ is an absolute constant. Therefore, we have that

$$\begin{aligned}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j)\right\}^2 &\ge \frac{319}{320}\left\{\int_{[0,1]}\Delta_A(r_i,s)X_t(s)ds\right\}^2-C_4'\zeta_n^2\log(T)\\ &=\frac{319}{320}\langle\Delta_A(r_i,\cdot),X_t(\cdot)\rangle_{\mathcal{L}^2}^2-C_4'\zeta_n^2\log(T).\end{aligned} \tag{44}$$

By Theorem 19 and the fact that

$$\|\Delta_A(r_i,\cdot)\|_{\mathcal{H}(\mathbb{K})}\le\|A^*\|_{\mathcal{H}(\mathbb{K})}+\|\widehat{A}\|_{\mathcal{H}(\mathbb{K})}\le 2C_A,$$

we have that with probability at least $1-T^{-4}$, it holds that uniformly for all $1\le i\le n$,

$$\begin{aligned}&\left|\frac{1}{T}\sum_{t=1}^{T}\langle\Delta_A(r_i,\cdot),X_t(\cdot)\rangle_{\mathcal{L}^2}^2-\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s\right|\\ \le&\frac{1}{320}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s+C_4'\log(T)\delta_T\end{aligned}$$

where $C_4'>0$ is an absolute constant. Thus the above display implies that

$$\frac{1}{T}\sum_{t=1}^{T}\langle\Delta_A(r_i,\cdot),X_t\rangle_{\mathcal{L}^2}^2\ge\frac{319}{320}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s-C_4''\log(T)\delta_T. \tag{45}$$

Therefore,

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{i=1}^{n}\left(\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j)\right)^2\\ \ge&\sum_{t=1}^{T}\sum_{i=1}^{n}\frac{319}{320}\langle\Delta_A(r_i,\cdot),X_t\rangle_{\mathcal{L}^2}^2-C_4'Tn\log(T)\zeta_n^2\\ \ge&T\left(\frac{319}{320}\right)^2\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s-C_4'''Tn\left(\log(T)\zeta_n^2+\log(T)\delta_T\right),\\ \ge&T\frac{159}{160}\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s-C_4'''Tn\left(\log(T)\zeta_n^2+\log(T)\delta_T\right),\end{aligned} \tag{46}$$

where the first inequality follows from (44) and the second inequality follows from (45). ■

Lemma 12 *Let $\zeta_n = n^{-\alpha+1/2}$ and $\delta_T = T^{-2r/(2r+1)}$. Under the same conditions as in Theorem 4, if in addition, $\delta_T \geq \frac{\log(T)}{T}$, then with probability at least $1 - T^{-4}$, it holds that*

$$\begin{aligned}
&\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j)\right\}\epsilon_t(r_i) \\
\leq&\frac{1}{320n}\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + C\left\{\zeta_n + \log(T)\delta_T\right\}. \qquad (47)
\end{aligned}$$

In addition, let $\{\mathfrak{E}_{t,i}\}_{t=1,i=1}^{T,n}$ be a collection of standard normal random variables independent of $\{X_t\}_{t=1}^{T}$,$\{r_i\}_{i=1}^{n}$ and $\{s_i\}_{i=1}^{n}$. Then with probability at least $1 - T^{-4}$, it holds that

$$\begin{aligned}
&\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j)\right\}\mathfrak{E}_{t,i} \\
\leq&\frac{1}{320n}\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + C\left\{\zeta_n + \log(T)\delta_T\right\}. \qquad (48)
\end{aligned}$$

Proof Let $\zeta_n = n^{-\alpha+1/2}$. For Equation (47), note that

$$\begin{aligned}
&2\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j)\right\}\epsilon_t(r_i) \\
=&2\sum_{t=1}^{T}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j) - \int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s\right\}\epsilon_t(r_i) \qquad (49)\\
+&2\sum_{t=1}^{T}\sum_{i=1}^{n}\epsilon_t(r_i)\int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s. \qquad (50)
\end{aligned}$$

Since for all r_i

$$\|\Delta_A(r_i,\cdot)X_t(\cdot)\|_{W^{\alpha,2}} \leq 8C_AC_X\sqrt{\log(T)},$$

Theorem 17 implies that with probability $1 - T^{-4}$, it holds that uniformly for all $1 \leq i \leq n$,

$$\frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{n}\sum_{j=1}^{n}\Delta_A(r_i,s_j)X_t(s_j) - \int_{[0,1]}\Delta_A(r_i,s)X_t(s)\,\mathrm{d}s\right\}\epsilon_t(r_i) \leq C_2\zeta_n\log(T),$$

where $C_2 > 0$ is an absolute constant. Therefore

$$(49) \leq 2C_2Tn\zeta_n\log(T).$$

To control the term (50), we deploy Theorem 18. Since

$$\|\Delta_A(r_i,\cdot)\|_{\mathcal{H}(\mathbb{K})} \leq \|A^*\|_{\mathcal{H}(\mathbb{K})} + \|\widehat{A}\|_{\mathcal{H}(\mathbb{K})} \leq 2C_A,$$

then with probability at least $1 - T^{-4}$, it holds that that

$$
\begin{aligned}
(50) \leq & C_2' Tn \left\{ \frac{1}{n} \sum_{i=1}^{n} \sqrt{\log(T)\delta_T \iint_{[0,1]^2} \Delta_A(r_i, r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s} + \log(T)\delta_T \right\} \\
\leq & 2C_2' Tn \left\{ 320 C_2' \log(T)\delta_T + \frac{1}{320 C_2' n} \sum_{i=1}^{n} \iint_{[0,1]^2} \Delta_A(r_i, r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s \right\},
\end{aligned}
$$

where $C_2' > 0$ is an absolute constant.

Therefore, we have that with probability at least $1 - 2T^{-4}$ that

$$
\begin{aligned}
& 2 \sum_{t=1}^{T} \sum_{i=1}^{n} \left\{ \frac{1}{n} \sum_{j=1}^{n} \Delta_A(r_i, s_j) X_t(s_j) \right\} \epsilon_t(r_i) \\
\leq & Tn \frac{1}{320n} \sum_{i=1}^{n} \iint_{[0,1]^2} \Delta_A(r_i, r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + C_2'' Tn \left\{ \zeta_n + \frac{\log^2(T)}{T} + \log(T)\delta_T \right\},
\end{aligned}
$$

where $C_3 > 0$ is an absolute constant. Equation (47) follows from the assumption that $\delta_T \geq \log(T)/T$.

The argument of Equation (48) is the same as that of Equation (47) and will be omitted for brevity. ■

Lemma 13 *Let* $\zeta_n = n^{-\alpha+1/2}$ *and* $\delta_T = T^{-2r/(2r+1)}$. *Under the same conditions as in Theorem 4, with probability at least* $1 - (T \vee p)^{-4}$, *it holds that*

$$
\frac{2}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \langle \Delta_\beta(r_i), Z_t \rangle_p \epsilon_t(r_i) \leq \frac{\lambda}{320} \sum_{l \in S} \|\Delta_{\beta_l}\|_n + \frac{\lambda}{320} \sum_{l \in S^c} \|\widehat{\beta}_l\|_n. \tag{51}
$$

In addition, let $\{\mathfrak{E}_{t,i}\}_{t=1,i=1}^{T,n}$ *be a collection of standard normal random variables independent of* $\{X_t\}_{t=1}^{T}$, $\{r_i\}_{i=1}^{n}$ *and* $\{s_i\}_{i=1}^{n}$. *Then with probability at least* $1 - T^{-4}$, *it holds that*

$$
\frac{2}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \langle \Delta_\beta(r_i), Z_t \rangle_p \mathfrak{E}_{t,i} \leq \frac{\lambda}{320} \sum_{l \in S} \|\Delta_{\beta_l}\|_n + \frac{\lambda}{320} \sum_{l \in S^c} \|\widehat{\beta}_l\|_n. \tag{52}
$$

Proof Let $\zeta_n = n^{-\alpha+1/2}$. For Equation (51), note that since $Z_{t,j}$ is centered Gaussian with variance bounded by C_Z and $\epsilon_t(r_i)$ is Gaussian with variance bounded by C_ϵ, $Z_{t,j}\epsilon_t(r_i)$ is sub-exponential with parameter $C_\epsilon C_Z$. Therefore by a union bound argument, there exists an absolute constant $C_1 > 0$ such that, for any $i \in \{1, \ldots, n\}$,

$$
\mathbb{P}\left\{ \left\| \frac{1}{T} \sum_{t=1}^{T} Z_t \epsilon_t(r_i) \right\|_\infty \geq C_1 \sqrt{\frac{\log(p \vee T)}{T}} \right\} \leq (Tp)^{-5}. \tag{53}
$$

Therefore, with probability at least $1 - (T \vee p)^{-5}$, it holds that

$$
\begin{aligned}
&\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\langle \Delta_\beta(r_i), Z_t\rangle_p \epsilon_t(r_i) \\
\le&\frac{2}{n}\sum_{i=1}^{n}\left\{\left\|\frac{1}{T}\sum_{t=1}^{T} Z_t\epsilon_t(r_i)\right\|_\infty \sum_{l=1}^{p}|\Delta_{\beta_l}(r_i)|\right\} \\
\le&\frac{2C_1}{n}\sqrt{\frac{\log(p\vee T)}{T}}\sum_{i=1}^{n}\sum_{l=1}^{p}|\Delta_{\beta_l}(r_i)| \\
\le&C_2\sqrt{\frac{\log(p\vee T)}{T}}\sum_{l=1}^{p}\|\Delta_{\beta_l}\|_n \\
\le&\frac{\lambda}{320}\sum_{l\in S}\|\Delta_{\beta_l}\|_n + \frac{\lambda}{320}\sum_{l\in S^c}\|\widehat{\beta}_l\|_n
\end{aligned}
$$

where the second inequality follows from (53), the third inequality follows from Hölder's inequality, and the last inequality follows from the assumption that $\lambda = C_\lambda\sqrt{\frac{\log(p\vee T)}{T}}$ for sufficiently large constant C_λ.

The argument of Equation (52) is the same as that of Equation (51) and will be omitted for brevity. ■

Lemma 14 *Let $\zeta_n = n^{-\alpha+1/2}$. Under the same conditions as in Theorem 4, with probability at least $1 - T^{-4}$, it holds that*

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s \\
\ge& c\mathbb{E}_{X^*}\left\{\int_{[0,1]}\left(\int_{[0,1]}\Delta_A(s,r)X^*(r)\,\mathrm{d}r\right)^2\mathrm{d}s\right\} - C'\log(T)\zeta_n.
\end{aligned}
\tag{54}
$$

Proof We have

$$
\mathbb{E}_{X^*}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\int_{[0,1]}\Delta_A(r_i,s)X^*(s)\,\mathrm{d}s\right)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s
$$

where X^* is the predictor in the test set. By Theorem 33, we have with probability at least $1 - T^{-4}$ that $\|X^*\|_{W^{\alpha,2}} \le 4C_X\sqrt{\log(T)}$. So it holds that

$$
\begin{aligned}
\|\langle\Delta_A(r_i,\cdot), X^*\rangle_{\mathcal{L}^2}\|_{W^{\alpha,2}} &= \left\|\int_{[0,1]}\Delta_A(r_i,s)X^*(s)\,\mathrm{d}s\right\|_{W^{\alpha,2}} \\
\le \int_{[0,1]}\|\Delta_A(\cdot,s)\|_{W^{\alpha,2}}|X^*(s)|\,\mathrm{d}s &\le 2C_A\|X^*\|_{\mathcal{L}^2} \le 8C_AC_X\sqrt{\log(T)}.
\end{aligned}
$$

Let $f(r) = \langle \Delta_A(r,\cdot), X^*(\cdot)\rangle_{\mathcal{L}^2}$. Then

$$\|f^2\|_{W^{\alpha,2}} \le 64 C_A^2 C_X^2 \log(T).$$

Applying Theorem 40 to f^2, we have that

$$\frac{1}{n}\sum_{i=1}^n \langle \Delta_A(r_i,\cdot), X^*\rangle_{\mathcal{L}^2}^2 \ge \int_{[0,1]} \langle \Delta_A(r,\cdot), X^*\rangle_{\mathcal{L}^2}^2 \,\mathrm{d}r - C_5' \log(T)\zeta_n,$$

and this implies that

$$\begin{aligned}
&\mathbb{E}_{X^*}\left\{\frac{1}{n}\sum_{i=1}^n \left(\int_{[0,1]} \Delta_A(r_i,s)X^*(s)\,\mathrm{d}s\right)^2\right\}\\
\ge&\mathbb{E}_{X^*}\left\{\int_{[0,1]}\left(\int_{[0,1]} \Delta_A(s,r)X^*(r)\,\mathrm{d}r\right)^2 \mathrm{d}s\right\} - C_5'\log(T)\zeta_n
\end{aligned}$$

■

Lemma 15 *Let $\zeta_n = n^{-\alpha+1/2}$. Under the same conditions as in Theorem 4, it holds that*

$$\frac{1}{n}\sum_{i=1}^n \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) \ge c\int_{[0,1]} \Delta_\beta(r)^\top \Sigma_Z \Delta_\beta(r)\,\mathrm{d}r - C\zeta_n. \tag{55}$$

Proof Note that since the minimal eigenvalue of Σ_Z is lower bounded by c_z,

$$\frac{1}{n}\sum_{i=1}^n \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) \ge \frac{c_z}{n}\sum_{i=1}^n\sum_{l=1}^p \Delta_{\beta_l}^2(r_i).$$

By Theorem 40, there exists an absolute constant C_4 such that

$$\frac{1}{n}\sum_{i=1}^n \Delta_{\beta_l}^2(r_i) \ge \int_{[0,1]} \Delta_{\beta_l}^2(r)\,\mathrm{d}r - C_4\zeta_n \|\Delta_{\beta_l}^2\|_{W^{\alpha,2}}, \quad \forall l = 1,\ldots,p. \tag{56}$$

As a result,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \Delta_\beta(r_i)^\top \Sigma_Z \Delta_\beta(r_i) \ge \frac{c_z}{n}\sum_{i=1}^n\sum_{l=1}^p \Delta_{\beta_l}^2(r_i) \ge& c_4 c_z \sum_{l=1}^p \int_{[0,1]} \Delta_{\beta_l}^2(r)\,\mathrm{d}r - C_4 c_z \sum_{l=1}^p \zeta_n \|\Delta_{\beta_l}^2\|_{W^{\alpha,2}}\\
\ge& c_4' \sum_{l=1}^p \int_{[0,1]} \Delta_{\beta_l}^2(r)\,\mathrm{d}r - C_4'\zeta_n,
\end{aligned}$$

where the second inequality follows from (56), and the third inequality follows from

$$\sum_{l=1}^p \|\Delta_{\beta_l}^2\|_{W^{\alpha,2}} \le \sum_{l=1}^p \|\Delta_{\beta_l}\|_{W^{\alpha,2}}^2 \le \sum_{l=1}^p \|\Delta_{\beta_l}\|_{\mathcal{H}(\mathbb{K}_\beta)}^2 \le 2\sum_{l=1}^p \left(\|\beta_l^*\|_{\mathcal{H}(\mathbb{K}_\beta)}^2 + \|\widehat{\beta}_l\|_{\mathcal{H}(\mathbb{K}_\beta)}^2\right)$$

$$\leq 2\Big\{\sum_{l=1}^{p}\|\beta_l^*\|_{\mathcal{H}(\mathbb{K}_\beta)}\Big\}^2 + 2\Big\{\sum_{l=1}^{p}\|\widehat{\beta}_l\|_{\mathcal{H}(\mathbb{K}_\beta)}\Big\}^2 \leq 4C_\beta^2.$$

Therefore

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Delta_\beta(r_i)^\top\Sigma_Z\Delta_\beta(r_i) \geq & c_4'\sum_{l=1}^{p}\int_{[0,1]}\Delta_{\beta_l}^2(r)\,\mathrm{d}r - C_4'\zeta_n \\ \geq & \frac{c_4'}{C_z}\int_{[0,1]}\Delta_\beta(r)^\top\Sigma_Z\Delta_\beta(r)\,\mathrm{d}r - C_4'\zeta_n \end{aligned} \tag{57}$$

where the last inequality follows from the fact that $w^\top\Sigma w \leq C_z\|w\|_2^2$ for all $w \in \mathbb{R}^p$.

■

B.4 Extensions

Corollary 16 *Define the discretized excess risk as*

$$\begin{aligned}\mathcal{E}^*_{\mathrm{dis}}(\widehat{A},\widehat{\beta}) =& \mathbb{E}_{X^*,Z^*,Y^*}\left\{\frac{1}{n_2}\sum_{j=1}^{n_2}\left(Y^*(r_j) - \frac{1}{n_1}\sum_{i=1}^{n_1}\widehat{A}(r_j,s_i)X^*(s_i) - \langle Z^*,\widehat{\beta}(r_j)\rangle_p\right)^2\right\} \\ & - \mathbb{E}_{X^*,Z^*,Y^*}\left\{\frac{1}{n_2}\sum_{j=1}^{n_2}\left(Y^*(r_j) - \frac{1}{n_1}\sum_{i=1}^{n_1}A^*(r_j,s_i)X^*(s_i)\,\mathrm{d}s - \langle Z^*,\beta^*(r_j)\rangle_p\right)^2\right\}. \end{aligned} \tag{58}$$

Suppose that Assumptions 1, 2 and 3 hold. Let $(\widehat{A},\widehat{\beta})$ be any solution to (4) with the tuning parameter $\lambda = C_\lambda\sqrt{\frac{\log(p\vee T)}{T}}$ for some sufficiently large constant C_λ. Define $n = \min\{n_1,n_2\}$. For $T \gtrsim \log(n)$, there exists absolute constants $C > 0$ such that with probability at least $1-8T^{-4}$, it holds that

$$\mathcal{E}^*_{\mathrm{dis}}(\widehat{A},\widehat{\beta}) \leq C\log(T)\big\{\delta_T + \mathfrak{s}\log(p\vee T)/T + \zeta_n\big\} \tag{59}$$

where $\zeta_n = n^{-\alpha+1/2}$ and $\delta_T = T^{\frac{-2r}{2r+1}}$.

Theorem 16 is a direct consequence of Theorem 4 and provides a formal theoretical guarantee for using $\mathcal{E}^*_{\mathrm{dis}}(\widehat{A},\widehat{\beta})$ to evaluate the proposed estimators in practice.

B.5 Proof of Theorem 8

Proof [Proof of Theorem 8] The proof is almost identical to the proof of Theorem 4. As a result, we only point out the difference.

From the minimizer property, we have that

$$\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}w_r(i)\left\{y_{t,i} - \frac{1}{n}\sum_{j=1}^{n}w_s(j)\widehat{A}(r_i,s_j)X_t(s_j) - \langle Z_t,\widehat{\beta}(r_i)\rangle_p\right\}^2 + \lambda\sum_{j=1}^{p}\|\widehat{\beta}_j\|_n$$

$$\leq \frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{y_{t,i} - \frac{1}{n}\sum_{j=1}^{n} w_s(j)A^*(r_i,s_j)X_t(s_j) - \langle Z_t, \beta^*(r_i)\rangle_p\right\}^2 + \lambda\sum_{j=1}^{p}\|\beta_j^*\|_n,$$

which implies that

$$\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j)\right\}^2 \tag{60}$$

$$+\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\langle \Delta_\beta(r_i), Z_t\rangle_p^2 \tag{61}$$

$$+\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left\{\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j)\langle \Delta_\beta(r_i), Z_t\rangle_p\right\} \tag{62}$$

$$\leq\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left(\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j) + \langle \Delta_\beta(r_i), Z_t\rangle_p\right)\epsilon_t(r_i) \tag{63}$$

$$+\lambda\sum_{j=1}^{p}\|\beta_j^*\|_n - \lambda\sum_{j=1}^{p}\|\widehat{\beta}_j\|_n \tag{64}$$

$$+\frac{2}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n} w_r(i)\left(\frac{1}{n}\sum_{j=1}^{n} w_s(j)\Delta_A(r_i,s_j)X_t(s_j) + \langle \Delta_\beta(r_i), Z_t\rangle_p\right)\mathfrak{E}_{t,i}. \tag{65}$$

Note that Equation (60) - (64) are identical to Equation (25) - (29). So it suffices to analyzed Equation (65). Without loss of generality, we will assume that

$$w_r(i) = 1 \text{ for all } 1 \leq i \leq n \quad \text{and} \quad w_s(j) = 1 \text{ for all } 1 \leq j \leq n.$$

It follows from (48) and (52), we have that with probability at least $1 - 2T^{-4}$, it holds that

$$\begin{aligned}(65) \leq &\frac{1}{320n}\sum_{i=1}^{n}\iint_{[0,1]^2}\Delta_A(r_i,r)\Sigma_X(r,s)\Delta_A(r_i,s)\,\mathrm{d}r\,\mathrm{d}s + C_1\left\{\zeta_n + \log(T)\delta_T\right\}\\ &+\frac{\lambda}{320}\sum_{j\in S}\|\Delta_{\beta_j}\|_n + \frac{\lambda}{320}\sum_{j\in S^c}\|\widehat{\beta}_j\|_n\end{aligned}$$

where $C_1 > 0$ is an absolute constant. The rest of the argument is identical to Theorem 4 and therefore is omitted. ■

Appendix C. Deviation bounds

C.1 Functional deviation bounds

Lemma 17 *Let $\{s_j\}_{j=1}^n \subset [0,1]$ be independent uniform random variables. Let $\{\varepsilon_t\}_{t=1}^T$ be a collection of independent standard Gaussian random variables independent of $\{X_t\}_{t=1}^T$ and*

$\{s_i\}_{i=1}^n$. *Under Assumption 3, it holds that*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{j=1}^{n} g(s_j)X_t(s_j) - \int_{[0,1]} g(s)X_t(s)\,\mathrm{d}s\right)\varepsilon_t\right| \geq C\log(T)\zeta_n \text{ for all } \|g\|_{W^{\alpha,2}} \leq 1\right) \leq T^{-5}$$

where $C, c > 0$ are absolute constants depending only on C_X and C_ε and $\zeta_n = n^{-\alpha+1/2}$.

Proof Step 1. Note that by Theorem 33, it holds that

$$\mathbb{P}\big(\|X_t\|_{W^{\alpha,2}} \geq 4C_X\sqrt{\log(T)}\big) \leq T^{-6}.$$

By a union bound argument,

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}\big(\|X_t\|_{W^{\alpha,2}} \geq 4C_X\sqrt{\log(T)} \text{ for all } 1 \leq t \leq T\big) \leq T^{-5}.$$

Under the event $\mathcal{E}$, for any g such that $\|g\|_{W^{\alpha,2}}$, it holds that

$$\|gX_t\|_{W^{\alpha,2}} \leq \|g\|_{W^{\alpha,2}}\|X_t\|_{W^{\alpha,2}} \leq 4C_X\sqrt{\log(T)}.$$

So under the event $\mathcal{E}$, for all $1 \leq t \leq T$ and all $\|g\|_{W^{\alpha,2}} \leq 1$, it holds that from Theorem 40 that

$$\left|\frac{1}{n}\sum_{j=1}^{n} g(s_j)X_t(s_j) - \int_{[0,1]} g(s)X_t(s)\,\mathrm{d}s\right| \leq C_1 n^{-\alpha+1/2}\sqrt{\log(T)} \text{ for all } \|g\|_{W^{\alpha,2}} \leq 1.$$

Since

$$\mathbb{P}(|\varepsilon_t| \geq C_2\sqrt{\log(T)}) \leq T^{-5}$$

the desired result immediately follows. ■

Through out this section, denote the eigen-expansion of linear map $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ as

$$L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}(\Phi_k) = \sum_{k=1}^{\infty}\xi_k\Phi_k.$$

In addition, for any bilinear operator Σ and any $\mathcal{L}^2$ functions f, g, denote

$$\Sigma[f,g] = \iint_{[0,1]^2}\Sigma(r,s)f(r)g(s)\,\mathrm{d}r\,\mathrm{d}s.$$

Lemma 18 *Suppose $\{X_t\}_{t=1}^T$ are independent and identically distributed centered Gaussian random processes and that the eigenvalues $\{\xi_k\}_{k=1}^\infty$ of the linear operator $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ satisfy that*

$$\xi_k \asymp k^{-2r}$$

for some $r > 1/2$*. Let* $\{\varepsilon_t\}_{t=1}^T \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ *and be independent of* $\{X_t\}_{t=1}^T$*. Then with probability at least* $1 - T^{-4}$*,*

$$\left| \frac{1}{T} \sum_{t=1}^T \langle X_t, \beta \rangle_{\mathcal{L}^2} \varepsilon_t \right| \le C\left(\sqrt{\Sigma_X[\beta,\beta] \log(T)\delta_T} + \log(T)\delta_T \right) \quad \textit{for all } \|\beta\|_{\mathcal{H}(\mathbb{K})} \le 1,$$

where $\delta_T = T^{-\frac{2r}{2r+1}}$ *and* C *is some absolute constant independent of* T*.*

Proof

Note that for any deterministic $\alpha \in \mathcal{H}(\mathbb{K})$.

$$\mathbb{E}(\langle X_t, \alpha \rangle_{\mathcal{L}^2} \varepsilon_t) = 0.$$

In addition, since $\langle X_t, \alpha \rangle_{\mathcal{L}^2} \sim \mathcal{N}(0, \Sigma_X[\alpha,\alpha])$ and $\varepsilon_t \sim \mathcal{N}(0,1)$, where

$$\Sigma_X[\alpha,\alpha] = \iint_{[0,1]^2} \alpha(s)\Sigma_X(s,t)\alpha(t)\,\mathrm{d}s\,\mathrm{d}t.$$

Thus $\langle X_t, \alpha \rangle_{\mathcal{L}^2} \varepsilon_t$ is a centered sub-exponential random variable with parameter $\Sigma_X[\alpha,\alpha]$. By Theorem 29, for $\gamma < 1$.

$$\mathbb{P}\left(\left| \frac{1}{T} \sum_{t=1}^T \langle X_t, \alpha \rangle_{\mathcal{L}^2} \varepsilon_t \right| \ge \gamma \sqrt{\Sigma_X[\alpha,\alpha]} \right) \le \exp(-2\gamma^2 T). \tag{66}$$

Denote

$$\mathcal{F} := \operatorname{span}\{L_{\mathbb{K}^{1/2}}(\Phi_k)\}_{k=1}^D \subset \mathcal{H}(\mathbb{K}) \text{ and}$$
$$\mathcal{F}^\perp := \operatorname{span}\{L_{\mathbb{K}^{1/2}}(\Phi_k)\}_{k=D+1}^\infty \subset \mathcal{H}(\mathbb{K}).$$

Denote $\mathcal{P}_{\mathcal{F}}$ to be the projection operator from $\mathcal{H}(\mathbb{K})$ to $\mathcal{F}$ with respect to the $\mathcal{H}(\mathbb{K})$ topology.
By Theorem 21, $L_{\mathbb{K}^{1/2}}(\Phi_k)$ is a collection of $\mathcal{H}(\mathbb{K})$ basis. For any β such that $\|\beta\|_{\mathcal{H}(\mathbb{K})} \le 1$, since $\{L_{\mathbb{K}^{1/2}}(\Phi_k)\}_{k=1}^\infty$ form a $\mathcal{H}(\mathbb{K})$ basis, $\|\mathcal{P}_{\mathcal{F}}(\beta)\|_{\mathcal{H}(\mathbb{K})}^2 + \|\mathcal{P}_{\mathcal{F}^\perp}(\beta)\|_{\mathcal{H}(\mathbb{K})}^2 = \|\beta\|_{\mathcal{H}(\mathbb{K})}^2 \le 1$. Note that

$$\frac{1}{T} \sum_{t=1}^T \langle X_t, \beta \rangle_{\mathcal{L}^2} \varepsilon_t = \frac{1}{T} \sum_{t=1}^T \langle X_t, \mathcal{P}_{\mathcal{F}}(\beta) \rangle_{\mathcal{L}^2} \varepsilon_t + \frac{1}{T} \sum_{t=1}^T \langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta) \rangle_{\mathcal{L}^2} \varepsilon_t.$$

Step 1. Let $D \le T$ to be chosen later. For any $J \in \mathbb{Z}$, $J \ge 1$ and $2^{-J} \ge T^{-6}$, consider the sets

$$\mathcal{F}_J = \left\{ \alpha = \sum_{k=1}^D f_k L_{\mathbb{K}^{1/2}}(\Phi_k) : 2^{-J-1} \le \sqrt{\Sigma_X[\alpha,\alpha]} \le 2^{-J}, \sum_{k=1}^D f_k^2 = \|\alpha\|_{\mathcal{H}(\mathbb{K})}^2 \le 1 \right\}$$

So $\mathcal{F}_J$ can be identified as a unit ball in $\mathbb{R}^D$. This means that for every $\delta > 0$, there exists a collection $\{\alpha_m\}_{m=1}^M$ such that for any $\alpha \in \mathcal{F}_J$,

$$\|\alpha_m - \alpha\|_{\mathcal{H}(\mathbb{K})} \leq \delta$$

and $M \leq \left(\frac{2}{\delta}\right)^D$.

Therefore for given m, $\alpha_m \in \mathcal{F}_J$, and so

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha_m\rangle_{\mathcal{L}^2}\varepsilon_t\right| \geq \gamma 2^{-J}\right) \leq \exp(-2\gamma^2 T).$$

So for any $\alpha \in \mathcal{F}_J$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha\rangle_{\mathcal{L}^2}\varepsilon_t \\
\leq &\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha - \alpha_m\rangle_{\mathcal{L}^2}\varepsilon_t + \sup_{1\leq m\leq M}\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha_m\rangle_{\mathcal{L}^2}\varepsilon_t \\
\leq &\|\alpha - \alpha_m\|_{\mathcal{L}^2}\left\|\frac{1}{T}\sum_{t=1}^T X_t\varepsilon_t\right\|_{\mathcal{L}^2} + \sup_{1\leq m\leq M}\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha_m\rangle_{\mathcal{L}^2}\varepsilon_t \\
\leq &\frac{\delta}{\sqrt{T}} + \sup_{1\leq m\leq M}\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha_m\rangle_{\mathcal{L}^2}\varepsilon_t.
\end{aligned}$$

Therefore

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha\rangle_{\mathcal{L}^2}\varepsilon_t\right| \geq \gamma 2\sqrt{\Sigma_X[\alpha,\alpha]} + \frac{\delta}{\sqrt{T}} \text{ for all } \alpha \in \mathcal{F}_J\right) \leq \exp\left(D\log(2/\delta) - \gamma^2 T\right).$$

Letting $\gamma = C\sqrt{\frac{D\log(T)}{T}}$ and $\delta = T^{-9/2}$ gives

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha\rangle_{\mathcal{L}^2}\varepsilon_t\right| \geq 2C\sqrt{\frac{\Sigma_X[\alpha,\alpha]D\log(T)}{T}} + \frac{1}{T^5} \text{ for all } \alpha \in \mathcal{F}_J\right) \leq T^{-6}.$$

Let

$$\mathcal{E} = \left\{\alpha = \sum_{k=1}^D f_k L_{\mathbb{K}^{1/2}}(\Phi_k) : \sqrt{\Sigma_X[\alpha,\alpha]} \leq \frac{1}{T^5}, \sum_{k=1}^D f_k^2 = \|\alpha\|^2_{\mathcal{H}(\mathbb{K})} \leq 1\right\}.$$

The similar argument shows that

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha\rangle_{\mathcal{L}^2}\varepsilon_t\right| \geq 2C\sqrt{\frac{\Sigma_X[\alpha,\alpha]D\log(T)}{T}} + \frac{1}{T^5} \text{ for all } \alpha \in \mathcal{E}\right) \leq T^{-6}.$$

Since $\Sigma_X[\alpha,\alpha] \le T^{-6}$ and $D \le T$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t,\alpha\rangle_{\mathcal{L}^2}\varepsilon_t\right| \ge \frac{\sqrt{\log(T)}}{T^3}+\frac{1}{T^5} \text{ for all } \alpha\in\mathcal{E}\right) \le T^{-6}.$$

Since

$$\{\alpha\in\mathcal{F}:\|\alpha\|_{\mathcal{H}(\mathbb{K})}\le 1\} = \bigcup_{2^{-J}\ge\frac{1}{T}} \mathcal{F}_J \cup \mathcal{E},$$

by a union bound argument,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t,\alpha\rangle_{\mathcal{L}^2}\varepsilon_t\right| \ge 2C\sqrt{\frac{D\Sigma_X[\alpha,\alpha]\log(T)}{T}}+\frac{\sqrt{\log(T)}}{T^3}+\frac{1}{T^5} \text{ for all } \alpha\in\mathcal{F}, \|\alpha\|_{\mathcal{H}(\mathbb{K})}\le 1\right)$$
$$\le \log(T)T^{-6} \tag{67}$$

Step 2. For any β such that $\|\beta\|_{\mathcal{H}(\mathbb{K})}\le 1$, it holds that $\|\mathcal{P}_{\mathcal{F}}(\beta)\|_{\mathcal{H}(\mathbb{K})}\le 1$. Therefore

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\langle X_t,\mathcal{P}_{\mathcal{F}}(\beta)\rangle_{\mathcal{L}^2}\varepsilon_t\\
\le& 2C\sqrt{\frac{D\Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta),\mathcal{P}_{\mathcal{F}}(\beta)]\log(T)}{T}}+\frac{\sqrt{\log(T)}}{T^3}+\frac{1}{T^5}\\
\le& 2C\sqrt{\frac{D\Sigma_X[\beta,\beta]\log(T)}{T}}+\frac{\sqrt{\log(T)}}{T^3}+\frac{1}{T^5},
\end{aligned}$$

where the first inequality follows from (67), and the second inequality holds because

$$\begin{aligned}
\Sigma_X[\beta,\beta] =& \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta),\mathcal{P}_{\mathcal{F}}(\beta)] + 2\Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta),\mathcal{P}_{\mathcal{F}}(\beta)] + \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta),\mathcal{P}_{\mathcal{F}^\perp}(\beta)]\\
=& \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta),\mathcal{P}_{\mathcal{F}}(\beta)] + \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta),\mathcal{P}_{\mathcal{F}^\perp}(\beta)]\\
\ge& \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta),\mathcal{P}_{\mathcal{F}}(\beta)]
\end{aligned}$$

where the second equality follows from Theorem 25, which implies that $\Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta),\mathcal{P}_{\mathcal{F}}(\beta)] = 0$

Step 3. Denote the eigen-expansion of linear map $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ as

$$L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}(\Phi_k) = \sum_{k=1}^{\infty}\xi_k\Phi_k.$$

Let

$$w_{t,k} = \frac{\langle L_{\mathbb{K}^{-1/2}}(X_t),\Phi_k\rangle_{\mathcal{L}^2}}{\sqrt{\xi_k}}.$$

Then

$$\sup_{\|\beta\|_{\mathcal{H}(\mathbb{K})}\le 1}\frac{1}{T}\sum_{t=1}^{T}\langle X_t,\mathcal{P}_{\mathcal{F}^\perp}(\beta)\rangle_{\mathcal{L}^2}\varepsilon_t$$

$$\le \sup_{\beta_2\in\mathcal{F}^\perp, \|\beta_2\|_{\mathcal{H}(\mathbb{K})}\le 1} \frac{1}{T}\sum_{t=1}^T \langle X_t, \beta_2\rangle_{\mathcal{L}^2}\varepsilon_t$$
$$\le \sup_{\beta_2\in\mathcal{F}^\perp, \|\beta_2\|_{\mathcal{H}(\mathbb{K})}\le 1} \frac{1}{T}\sum_{t=1}^T \langle L_{\mathbb{K}^{1/2}}(X_t), L_{\mathbb{K}^{-1/2}}(\beta_2)\rangle_{\mathcal{L}^2}\varepsilon_t$$
$$= \sup_{\beta_2\in\mathcal{F}^\perp, \|\beta_2\|_{\mathcal{H}(\mathbb{K})}\le 1} \frac{1}{T}\sum_{t=1}^T \sum_{k=1}^\infty \langle L_{\mathbb{K}^{1/2}}(X_t), \Phi_k\rangle_{\mathcal{L}^2}\langle L_{\mathbb{K}^{-1/2}}(\beta_2), \Phi_k\rangle_{\mathcal{L}^2}\varepsilon_t$$
$$= \sup_{\beta_2\in\mathcal{F}^\perp, \|\beta_2\|_{\mathcal{H}(\mathbb{K})}\le 1} \frac{1}{T}\sum_{t=1}^T \sum_{k=D+1}^\infty \langle L_{\mathbb{K}^{1/2}}(X_t), \Phi_k\rangle_{\mathcal{L}^2}\langle L_{\mathbb{K}^{-1/2}}(\beta_2), \Phi_k\rangle_{\mathcal{L}^2}\varepsilon_t$$
$$= \sup_{\sum_{k=D+1}^\infty f_k^2\le 1} \frac{1}{T}\sum_{t=1}^T \sum_{k=D+1}^\infty \sqrt{\xi_k} w_{t,k} f_k \varepsilon_t,$$

where the fourth inequality holds because $\beta_2 \in \mathcal{F}^\perp$ and the last inequality follows from Theorem 22.
Note that since $w_{t,k}$ and ε_t are both centered Gaussian with variance 1,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{i=1}^n w_{t,k}\varepsilon_t\right| \le \sqrt{\frac{4\log(k)+12\log(T)}{T}} + \frac{4\log(k)+12\log(T)}{T} \text{ for all } k\ge 1\right)$$
$$\le \sum_{k=1}^\infty \mathbb{P}\left(\left|\frac{1}{T}\sum_{i=1}^n w_{t,k}\varepsilon_t\right| \le \sqrt{\frac{4\log(k)+12\log(T)}{T}} + \frac{4\log(k)+12\log(T)}{T}\right)$$
$$\le \sum_{k=1}^\infty \frac{1}{T^6 k^2} \le \frac{1}{T^5}.$$

So for any $D' \ge D$,

$$\sup_{\sum_{k=D+1}^\infty f_k^2\le 1} \frac{1}{T}\sum_{t=1}^T \sum_{k=D+1}^\infty \sqrt{\xi_k} w_{t,k} f_k \varepsilon_t$$
$$= \sup_{\sum_{k=D+1}^\infty f_k^2\le 1} \sum_{k=D+1}^\infty \sqrt{\xi_k} f_k \left(\frac{1}{T}\sum_{t=1}^T w_{t,k}\varepsilon_t\right)$$
$$\le \sup_{\sum_{k=D+1}^\infty f_k^2\le 1} \sum_{k=D+1}^\infty k^{-r}\left(\sqrt{\frac{4\log(k)+12\log(T)}{T}} + \frac{4\log(k)+12\log(T)}{T}\right) f_k$$
$$\le \sup_{\sum_{k=D+1}^\infty f_k^2\le 1} \left(\sqrt{\sum_{k=D+1}^\infty \frac{4\log(k)+12\log(T)}{k^{2r}T}} + \sqrt{\sum_{k=D+1}^\infty \frac{4\log^2(k)+12\log^2(T)}{k^{2r}T^2}}\right)\sqrt{\sum_{k=D+1}^\infty f_k^2} \tag{68}$$

Note that

$$\sup_{\sum_{k=D+1}^\infty f_k^2\le 1} \sqrt{\sum_{k=D+1}^\infty \frac{4\log(k)+12\log(T)}{k^{2r}T}}\sqrt{\sum_{k=D+1}^\infty f_k^2}$$

$$
\leq \sqrt{\sum_{k=D+1}^{\infty} \frac{4\log(k)}{k^{2r}T}} + \sqrt{\sum_{k=D+1}^{\infty} \frac{12\log(T)}{k^{2r}T}}
$$

$$
\leq \sqrt{\frac{4\log(p)}{D^{2r-1}T} + \frac{1}{(D')^{r-1/2}T}} + \sqrt{\frac{12\log(T)}{TD^{2r-1}}}.
$$

where the first inequality follows from Theorem 30. The second term in Equation (68) can be bounded in a similar way.

Step 4. Putting **Step 2** and **Step 3** together, it holds that with high probability, for all $\beta \in \mathcal{H}(\mathbb{K})$ such that $\|\beta\|_{\mathcal{H}(\mathbb{K})} \leq 1$,

$$
\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \beta\rangle_{\mathcal{L}^2}\varepsilon_t \leq C'\left(\sqrt{\frac{D\Sigma_X[\alpha,\alpha]\log(T)}{T}} + \frac{\sqrt{\log(T)}}{T^3} + \frac{1}{T^5} + \sqrt{\frac{4\log(p)}{D^{2r-1}T} + \frac{1}{(D')^{r-1/2}T}} + \sqrt{\frac{12\log(T)}{TD^{2r-1}}}\right)
$$

Set $D' = \max\{T^{\frac{10}{r-1/2}}, D\}$ gives

$$
\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \beta\rangle_{\mathcal{L}^2}\varepsilon_t \leq C'\left(\sqrt{\frac{D\Sigma_X[\alpha,\alpha]\log(T)}{T}} + \frac{\sqrt{\log(T)}}{T^3} + \frac{1}{T^5} + \sqrt{\frac{\log(T)}{T}D^{-2r+1}} + \frac{1}{T^5} + \sqrt{\frac{12\log(T)}{TD^{2r-1}}}\right)
$$

Taking $D = T^{\frac{1}{2r+1}}$ gives

$$
\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \beta\rangle_{\mathcal{L}^2}\varepsilon_t \leq C''\left(\sqrt{\Sigma_X[\alpha,\alpha]\frac{\log(T)}{T^{\frac{2r}{2r+1}}}} + \frac{\log(T)}{T^{\frac{2r}{2r+1}}} + \frac{1}{T^2}\right).
$$

■

Lemma 19 *Suppose $\{X_t\}_{t=1}^T$ are independent identically distributed centered Gaussian random processes and that the eigenvalues of $\{\xi_k\}_{k=1}^{\infty}$ of the linear operator $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ satisfy*

$$
\xi_k \asymp k^{-2r}
$$

for some $r > 1/2$. Then with probability at least $1 - T^{-4}$, it holds that for any $0 < \tau < 1$,

$$
\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \beta\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\beta,\beta]\right| \leq \tau\Sigma_X[\beta,\beta] + C_\tau\log(T)\delta_T \quad \textit{for all} \quad \|\beta\|_{\mathcal{H}(\mathbb{K})} \leq 1,
$$

where $\delta_T = T^{-\frac{2r}{2r+1}}$, and C_τ is some constant only depending on τ and independent of T.

Proof

Note that for any deterministic $\alpha \in \mathcal{H}(\mathbb{K})$, $\langle X_t, \alpha\rangle_{\mathcal{L}^2}$ is $SG(\Sigma_X[\alpha,\alpha])$. Thus by Hanson-Wright, it holds that for all $\gamma < 1$,

$$
\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \alpha\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha,\alpha]\right| \geq \gamma\Sigma_X[\alpha,\alpha]\right) \leq \exp(-2\gamma^2 T). \tag{69}
$$

For $D \le T$ to be chosen later, denote

$$\mathcal{F} := \mathrm{span}\left\{L_{\mathbb{K}^{1/2}}(\Phi_k)\right\}_{k=1}^{D} \quad \text{and} \quad \mathcal{F}^{\perp} := \mathrm{span}\left\{L_{\mathbb{K}^{1/2}}(\Phi_k)\right\}_{k=D+1}^{\infty}.$$

Denote $\mathcal{P}_{\mathcal{F}}$ to be the projection operator from $\mathcal{H}(\mathbb{K})$ to $\mathcal{F}$ with respect to the $\mathcal{H}(\mathbb{K})$ topology. Them it holds that for all $\beta \in \mathcal{H}(\mathbb{K})$,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \beta\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\beta, \beta]\right| \\
\le &\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \mathcal{P}_{\mathcal{F}}(\beta)\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)]\right| \\
+2&\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \mathcal{P}_{\mathcal{F}}(\beta)\rangle_{\mathcal{L}^2}\langle X_t, \mathcal{P}_{\mathcal{F}^{\perp}}(\beta)\rangle_{\mathcal{L}^2} - \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}^{\perp}}(\beta)]\right| \\
+&\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \mathcal{P}_{\mathcal{F}^{\perp}}(\beta)\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}^{\perp}}(\beta), \mathcal{P}_{\mathcal{F}^{\perp}}(\beta)]\right|
\end{aligned}$$

Step 1. For any $J \in \mathbb{Z}$, $J \ge 1$ and $2^{-J} \ge T^{-5}$, consider the set

$$\mathcal{F}_J = \left\{\alpha = \sum_{k=1}^{D} f_k L_{\mathbb{K}^{1/2}}(\Phi_k) : 2^{-J-1} \le C[\alpha, \alpha] \le 2^{-J}, \sum_{k=1}^{D} f_k^2 = \|\alpha\|_{\mathcal{H}(\mathbb{K})}^2 \le 1\right\}.$$

So $\mathcal{F}_J$ can be viewed as a subset of unit ball in $\mathbb{R}^D$. This means that for every $\delta > 0$, there exists a collection $\{\alpha_m\}_{m=1}^{M}$ such that for any $\alpha \in \mathcal{F}_J$,

$$\|\alpha_m - \alpha\|_{\mathcal{H}(\mathbb{K})} \le \delta$$

and $M \le \left(\frac{2}{\delta}\right)^D$. Therefore for given m, $\alpha_m \in \mathcal{F}_J$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \alpha_m\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha_m, \alpha_m]\right| \ge \gamma\Sigma_X[\alpha_m, \alpha_m]\right) \le \exp(-2\gamma^2 T). \tag{70}$$

Denote $\widehat{\Sigma}_X(r, s) = \frac{1}{T}\sum_{t=1}^{T} X_t(r)X_t(s)$. We have that for any $\alpha \in \mathcal{F}_J$,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^{T}\langle X_t, \alpha\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha, \alpha]\right| \\
=&\left|\iint \alpha(r)\left(\frac{1}{T}\sum_{t=1}^{T} X_t(r)X_t(s) - \Sigma_X(r, s)\right)\alpha(s)\,\mathrm{d}r\,\mathrm{d}s\right|
\end{aligned}$$

$$
\begin{aligned}
\leq & \Bigg| \iint \alpha(r) \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right) \alpha(s) \, \mathrm{d}r \, \mathrm{d}s \\
& - \iint \alpha_m(r) \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right) \alpha_m(s) \, \mathrm{d}r \, \mathrm{d}s \Bigg| \\
& + \left| \iint \alpha_m(r) \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right) \alpha_m(s) \, \mathrm{d}r \, \mathrm{d}s \right| \\
\leq & 2 \|\alpha - \alpha_m\|_{\mathcal{L}^2} \|\alpha\|_{\mathcal{L}^2} \sqrt{ \iint \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right)^2 \mathrm{d}r \, \mathrm{d}s } \\
& + \sup_{1 \leq m \leq M} \left| \iint \alpha_m(r) \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right) \alpha_m(s) \, \mathrm{d}r \, \mathrm{d}s \right| \\
= & 2 \|\alpha - \alpha_m\|_{\mathcal{L}^2} \|\alpha\|_{\mathcal{L}^2} \left\| \widehat{\Sigma}_X - \Sigma_X \right\|_F \\
& + \sup_{1 \leq m \leq M} \left| \iint \alpha_m(r) \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right) \alpha_m(s) \, \mathrm{d}r \, \mathrm{d}s \right| \\
\leq & 2 C_X \delta \sqrt{\frac{\log(T)}{T}} + \sup_{1 \leq m \leq M} \left| \iint \alpha_m(r) \left(\frac{1}{T} \sum_{t=1}^{T} X_t(r) X_t(s) - \Sigma_X(r,s) \right) \alpha_m(s) \, \mathrm{d}r \, \mathrm{d}s \right| \\
= & 2 C_X \delta \sqrt{\frac{\log(T)}{T}} + \sup_{1 \leq m \leq M} \left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \alpha_m \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha_m, \alpha_m] \right|
\end{aligned}
$$

Therefore

$$
\mathbb{P}\left(\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \alpha \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha, \alpha] \right| \geq 2\gamma \Sigma_X[\alpha,\alpha] + 2 C_X \delta \sqrt{\frac{\log(T)}{T}} \text{ for all } \alpha \in \mathcal{F}_J \right) \leq \exp\left(D \log(2/\delta) - \gamma^2 T \right).
$$

So $\gamma = C \sqrt{\frac{D \log(T)}{T}}$ and $\delta = T^{-9/2}$ gives

$$
\mathbb{P}\left(\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \alpha \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha, \alpha] \right| \geq C_1 \left(\sqrt{\frac{D \log(T)}{T}} \Sigma_X[\alpha,\alpha] + \frac{\sqrt{\log(T)}}{T^5} \right) \text{ for all } \alpha \in \mathcal{F}_J \right) \leq T^{-6}.
$$

Let

$$
\mathcal{E} = \left\{ \alpha = \sum_{k=1}^{D} f_k L_{\mathbb{K}^{1/2}}(\Phi_k) \in \mathcal{H}(\mathbb{K}) : \Sigma_X[\alpha,\alpha] \leq \frac{1}{T^5}, \sum_{k=1}^{D} f_k^2 = \|\alpha\|_{\mathcal{H}(\mathbb{K})}^2 \leq 1 \right\}.
$$

The similar argument shows that

$$
\mathbb{P}\left(\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \alpha \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha, \alpha] \right| \geq C_1 \left(\sqrt{\frac{D \log(T)}{T}} \Sigma_X[\alpha,\alpha] + \frac{\sqrt{\log(T)}}{T^5} \right) \text{ for all } \alpha \in \mathcal{E} \right) \leq T^{-6}.
$$

Since $\Sigma_X[\alpha,\alpha] \le T^{-5}$ when $\alpha \in \mathcal{E}$, and $D \le T$,

$$\mathbb{P}\left(\left| \frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha,\alpha] \right| \ge C_1' \frac{\sqrt{\log(T)}}{T^5} \text{ for all } \alpha \in \mathcal{E} \right) \le T^{-6}.$$

Since

$$\{\alpha \in \mathcal{F} : \|\alpha\|_{\mathcal{H}(\mathbb{K})} \le 1\} = \bigcup_{2^{-J} \ge T^{-5}} \mathcal{F}_J \bigcup \mathcal{E},$$

by union bound,

$$\mathbb{P}\left(\left| \frac{1}{T}\sum_{t=1}^T \langle X_t, \alpha\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\alpha,\alpha] \right| \ge C_1' \left(\sqrt{\frac{D\log(T)}{T}} \Sigma_X[\alpha,\alpha] + \frac{\sqrt{\log(T)}}{T^5} \right) \text{ for all } \alpha \in \mathcal{F}, \|\alpha\|_{\mathcal{H}(\mathbb{K})} \le 1 \right) \tag{71}$$

$\le \log(T)T^{-6}$

Step 2. For any β such that $\|\beta\|_{\mathcal{H}(\mathbb{K})} \le 1$, it holds that $\|\mathcal{P}_{\mathcal{F}}(\beta)\|_{\mathcal{H}(\mathbb{K})} \le 1$. Therefore

$$\begin{aligned}
&\left| \frac{1}{T}\sum_{t=1}^T \langle X_t, \mathcal{P}_{\mathcal{F}}(\beta)\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] \right| \\
\le & C_1' \left(\sqrt{\frac{D\log(T)}{T}} \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] + \frac{\sqrt{\log(T)}}{T^5} \right) \\
\le & C_1' \left(\sqrt{\frac{D\log(T)}{T}} \Sigma_X[\beta,\beta] + \frac{\sqrt{\log(T)}}{T^5} \right)
\end{aligned}$$

where the first inequality follows from (71), and the last inequality holds because

$$\begin{aligned}
\Sigma_X[\beta,\beta] = & \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] + 2\Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] + \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \\
= & \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] + \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \\
\ge & \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)]
\end{aligned}$$

where by Theorem 25, $\Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] = 0$.

Step 3. Denote the eigen-expansion of linear map $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ as

$$L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}(\Phi_k) = \sum_{k=1}^\infty \xi_k \Phi_k.$$

Let

$$w_{t,k} = \frac{\langle L_{\mathbb{K}^{1/2}}(X_t), \Phi_k\rangle_{\mathcal{L}^2}}{\sqrt{\xi_k}}.$$

For any β, let $f_k = \langle L_{\mathbb{K}^{-1/2}}(\Phi_k), \beta\rangle_{\mathcal{L}^2}$. Note that if $\|\beta\|_{\mathcal{H}(\mathbb{K})}^2 \le 1$, then $\|\mathcal{P}_{\mathcal{F}^\perp}(\beta)\|_{\mathcal{H}(\mathbb{K})} \le 1$.

$$\langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta)\rangle_{\mathcal{L}^2} = \sum_{k=D+1}^\infty \langle L_{\mathbb{K}^{1/2}}(X_t), \Phi_k\rangle_{\mathcal{L}^2} \langle L_{\mathbb{K}^{-1/2}}(\beta), \Phi_k\rangle_{\mathcal{L}^2} = \sum_{k=D+1}^\infty \sqrt{\xi_k} w_{t,k} f_k.$$

Note that if $\|\beta\|_{\mathcal{H}(\mathbb{K})} \le 1$, then $\|\mathcal{P}_{\mathcal{F}^\perp}(\beta)\|_{\mathcal{H}(\mathbb{K})} \le 1$. Therefore

$$\sup_{\|\beta\|_{\mathcal{H}(\mathbb{K})}\le 1} \left| \frac{1}{T}\sum_{t=1}^{T} \langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta)\rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \right|$$
$$\le \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left| \sum_{k,l=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\xi_l} f_l \left(\frac{1}{T}\sum_{t=1}^{T} w_{i,k}w_{i,l} - \frac{1}{T}\sum_{t=1}^{T} E(w_{i,k}w_{i,l}) \right) \right|,$$

where the last inequality follows from Theorem 22.
Let $\eta = \sqrt{\frac{4\log(k)+4\log(l)+12\log(T)}{T}}$. Note that since $w_{t,k}w_{i,l}$ is SE(1),

$$\mathbb{P}\left(\left| \frac{1}{T}\sum_{i=1}^{n} w_{t,k}w_{i,l} - \frac{1}{T}\sum_{t=1}^{T} E(w_{i,k}w_{i,l}) \right| \le \eta + \eta^2 \text{ for all } k, l \ge 1 \right)$$
$$\le \sum_{k,l=1}^{\infty} P\left(\left| \frac{1}{T}\sum_{i=1}^{n} w_{t,k}w_{i,l} - \frac{1}{T}\sum_{t=1}^{T} E(w_{i,k}w_{i,l}) \right| \le \eta + \eta^2 \right)$$
$$\le \sum_{k,l=1}^{\infty} \frac{1}{T^6 k^2 l^2} \le \frac{1}{T^5}.$$

As a result

$$\sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left| \sum_{k,l=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\xi_l} f_l \left(\frac{1}{T}\sum_{t=1}^{T} w_{i,k}w_{i,l} - E(w_{i,k}w_{i,l}) \right) \right|$$
$$\le \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left| \sum_{k,l=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\xi_l} f_l \, (\eta + \eta^2) \right|$$
$$\le \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \frac{1}{\sqrt{T}} \left| \sum_{k,l=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\xi_l} f_l \left(\sqrt{4\log(k)} + \sqrt{4\log(l)} + \sqrt{12\log(T)} + \frac{4\log(k)+4\log(l)+12\log(T)}{\sqrt{T}} \right) \right| \tag{72}$$

Observe that for $D' \ge D$,

$$\sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left| \sum_{k,l=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\xi_l} f_l \sqrt{\log(k)} \right|$$
$$\le \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left(\sum_{l=D+1}^{\infty} \sqrt{\xi_l} f_l \right) \left(\sum_{k=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\log(k)} \right)$$
$$\le \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left(\sqrt{\sum_{l=D+1}^{\infty} l^{-2r}} \sqrt{\sum_{l=D+1}^{\infty} f_l^2} \right) \left(\sqrt{\sum_{k=D+1}^{\infty} k^{-2r}\log(k)} \sqrt{\sum_{l=D+1}^{\infty} f_k^2} \right)$$
$$\le \sqrt{D^{-2r+1}} \sqrt{\left(\frac{\log(p)}{D^{2r-1}} + \frac{1}{(D')^{r-1/2}} \right)},$$

where the last inequality follows from Theorem 30. In addition,

$$
\begin{aligned}
&\sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left| \sum_{k,l=D+1}^{\infty} \sqrt{\xi_k} f_k \sqrt{\xi_l} f_l \sqrt{\log(T)} \right| \\
\le &\sqrt{\log(T)} \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left(\sum_{l=D+1}^{\infty} \sqrt{\xi_l} f_l \right) \left(\sum_{k=D+1}^{\infty} \sqrt{\xi_k} f_k \right) \\
\le &\sqrt{\log(T)} \sup_{\sum_{k=D+1}^{\infty} f_k^2 \le 1} \left(\sqrt{\sum_{l=D+1}^{\infty} l^{-2r}} \sqrt{\sum_{l=D+1}^{\infty} f_l^2} \right) \left(\sqrt{\sum_{k=D+1}^{\infty} k^{-2r}} \sqrt{\sum_{l=D+1}^{\infty} f_k^2} \right) \\
\le &\sqrt{\log(T)} D^{-2r+1}.
\end{aligned}
$$

The rest terms in Equation (72) can be handled in a similar way. So

$$
\sup_{\|\beta\|_{\mathcal{H}(\mathbb{K})} \le 1} \left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta) \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \right| \tag{73}
$$

$$
\le C_2 \frac{1}{\sqrt{T}} \left(\sqrt{D^{-2r+1} \left(\frac{\log(p)}{D^{2r-1}} + \frac{1}{(D')^{r-1/2}} \right)} + \sqrt{\log(T)} D^{-2r+1} \right) \tag{74}
$$

$$
\le C_2 \left(\sqrt{\frac{D^{-4r+2} \log(T)}{T} + \frac{D^{-2r+1}}{T^{10}}} + \sqrt{\frac{\log(T)}{T}} D^{-2r+1} \right) \le C_2' \left(D^{-2r+1} \sqrt{\frac{\log(T)}{T}} + \sqrt{\frac{D^{-2r+1}}{T^{10}}} \right), \tag{75}
$$

where the second inequality follows by setting $D' = \max\{n^{\frac{9}{r-1/2}}, D\}$.

Step 4. So with high probability, it holds that for all β such that $\|\beta\|_{\mathcal{H}(\mathbb{K})} \le 1$,

$$
\begin{aligned}
\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \mathcal{P}_{\mathcal{F}}(\beta) \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] \right| &\le C_3 \left(\sqrt{\frac{D \log(T)}{T} \Sigma_X[\beta, \beta]} \right) + \frac{\sqrt{\log(T)}}{T^5} \\
\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta) \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \right| &\le C_3 \left(D^{-2r+1} \sqrt{\frac{\log(T)}{T}} + \sqrt{\frac{D^{-2r+1}}{T^{10}}} \right).
\end{aligned}
$$

Taking $D = \frac{cT}{\log(T)}$ for sufficiently small constant c gives,

$$
\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \mathcal{P}_{\mathcal{F}}(\beta) \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] \right| \le \frac{\tau}{2} \Sigma_X[\beta, \beta] + C_4 T^{-4} \tag{76}
$$

$$
\begin{aligned}
\left| \frac{1}{T} \sum_{t=1}^{T} \langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta) \rangle_{\mathcal{L}^2}^2 - \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \right| &\le C_4 \left(\left(\frac{\log(T)}{T} \right)^{2r-1/2} + \frac{1}{T^4} \right) \\
&\le C_4 \left(\left(\frac{\log(T)}{T} \right)^{\frac{2r}{2r+1}} + \frac{1}{T^4} \right),
\end{aligned} \tag{77}
$$

where $r > 1/2$ is used in the last inequality.

Step 5. Denote

$$\delta_T = \left(\frac{\log(T)}{T}\right)^{\frac{2r}{2r+1}}.$$

Note that if $k \neq l$,

$$\begin{aligned}
\mathbb{E}(w_{i,k}w_{i,l}) =& \mathbb{E}\left(\frac{\langle L_{\mathbb{K}^{1/2}}(X_t), \Phi_k\rangle_{\mathcal{L}^2}}{\sqrt{\xi_k}}\frac{\langle L_{\mathbb{K}^{1/2}}(X_t), \Phi_l\rangle_{\mathcal{L}^2}}{\sqrt{\xi_l}}\right) \\
=& \frac{1}{\sqrt{\xi_k\xi_l}} E\left(\langle X_t, L_{\mathbb{K}^{1/2}}(\Phi_k)\rangle_{\mathcal{L}^2}\langle X_t, L_{\mathbb{K}^{1/2}}(\Phi_l)\rangle_{\mathcal{L}^2}\right) \\
=& \frac{1}{\sqrt{\xi_k\xi_l}}\Sigma_X[L_{\mathbb{K}^{1/2}}(\Phi_k), L_{\mathbb{K}^{1/2}}(\Phi_l)] \\
=& \frac{1}{\sqrt{\xi_k\xi_l}}\langle L_{\mathbb{K}^{1/2}C\mathbb{K}^{1/2}}(\Phi_k), \Phi_l\rangle_{\mathcal{L}^2} = 0.
\end{aligned}$$

Therefore

$$\begin{aligned}
&\sup_{\|\beta\|_{\mathcal{H}(\mathbb{K})}\leq 1} \Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] \\
\leq& \sup_{\sum_{k=D+1}^\infty f_k^2 \leq 1}\left|\sum_{k,l=D+1}^\infty \sqrt{\xi_k}f_k\sqrt{\xi_l}f_l\frac{1}{T}\sum_{t=1}^T E(w_{i,k}w_{i,l})\right| \\
=& \sup_{\sum_{k=D+1}^\infty f_k^2 \leq 1}\left|\sum_{k=D+1}^\infty \sqrt{\xi_k}f_k\sqrt{\xi_k}f_k\frac{1}{T}\sum_{t=1}^T E(w_{i,k}w_{i,k})\right| \\
\leq& 2\sup_{\sum_{k=D+1}^\infty f_k^2 \leq 1}\left|\sum_{k=D+1}^\infty f_k^2 k^{-2r}\right| \\
\leq& 2\sup_{\sum_{k=D+1}^\infty f_k^2 \leq 1}\sqrt{\sum_{k=D+1}^\infty f_k^4}\sqrt{\sum_{k=D+1}^\infty k^{-4r}} \\
\leq& 2D^{-2r+1/2} \leq C_5\left(\frac{\log(T)}{T}\right)^{2r-1/2} \leq C_5\left(\frac{\log(T)}{T}\right)^{\frac{2r}{2r+1}} \leq C_5\log(T)\delta_T. \qquad (78)
\end{aligned}$$

Observe that

$$\Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] = 0.$$

So for any $\|\beta\|_{\mathcal{H}(\mathbb{K})} \leq 1$,

$$\begin{aligned}
&\left|\frac{1}{T}\sum_{t=1}^T\langle X_t, \mathcal{P}_{\mathcal{F}}(\beta)\rangle_{\mathcal{L}^2}\langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta)\rangle_{\mathcal{L}^2} - \Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)]\right| \\
\leq& \left|\sqrt{\frac{1}{T}\sum_{t=1}^T\langle X_t, \mathcal{P}_{\mathcal{F}}(\beta)\rangle_{\mathcal{L}^2}^2}\sqrt{\frac{1}{T}\sum_{t=1}^T\langle X_t, \mathcal{P}_{\mathcal{F}^\perp}(\beta)\rangle_{\mathcal{L}^2}^2}\right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \Big(\Sigma_X[\mathcal{P}_{\mathcal{F}}(\beta), \mathcal{P}_{\mathcal{F}}(\beta)] + \frac{\tau}{2}\Sigma_X[\beta,\beta] + C_4 T^{-4}\Big)^{1/2} \Big(\Sigma_X[\mathcal{P}_{\mathcal{F}^\perp}(\beta), \mathcal{P}_{\mathcal{F}^\perp}(\beta)] + C_4\Big(\delta_T + T^{-4}\Big)\Big)^{1/2} \\
&\leq \big(2\Sigma_X[\beta,\beta] + T^{-4}\big)^{1/2} \big(C_6\delta_T + T^{-4}\big)^{1/2} \\
&\leq \frac{\tau}{2}\Sigma_X[\beta,\beta] + C_7\delta_T, \qquad (79)
\end{aligned}
$$

where the second inequality follows from (76) and (77), and the third inequality follows from (78), and the last inequality holds if C_7 is a sufficiently large constant. The desired results follows from Equation (76), Equation (77) and Equation (79).

■

C.2 Additional lemmas for kernel alignment

Remark 20 *Note that following the same argument as that in the proof of Theorem 2 in Cai and Yuan (2012), the function* $L_{\mathbb{K}^{1/2}} : \mathcal{L}^2 \to \mathcal{H}(\mathbb{K})$ *admits a well-defined inverse function* $L_{\mathbb{K}^{-1/2}} : \mathcal{H}(\mathbb{K}) \to \mathcal{L}^2$ *such that*

$$L_{\mathbb{K}^{-1/2}}(\phi_k) = \mu_k^{-1/2}\phi_k,$$

where the eigen-expansion of $\mathbb{K}(r,s)$ *satisfies* $\mathbb{K}(r,s) = \sum_{k=1}^\infty \mu_k\phi_k(r)\phi_k(s)$.

Recall that the eigen-expansion of the linear map $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ takes the form

$$\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2} = \sum_{k=1}^\infty \xi_k\Phi_k.$$

Lemma 21 *For any* $f, g \in \mathcal{H}(\mathbb{K})$,

$$\langle f, g\rangle_{\mathcal{H}(\mathbb{K})} = \langle L_{\mathbb{K}^{-1/2}}(f), L_{\mathbb{K}^{-1/2}}(g)\rangle_{\mathcal{L}^2}.$$

In addition, for any $\mathcal{L}^2$ *basis* $\{\Phi_k\}_{k=1}^\infty$, *we have* $\{L_{\mathbb{K}^{1/2}}(\Phi_k)\}_{k=1}^\infty$ *is a collection of* $\mathcal{H}(\mathbb{K})$ *basis. (However,* $\{L_{\mathbb{K}^{1/2}}(\Phi_k)\}_{k=1}^\infty$ *is not necessary a collection of* $\mathcal{L}^2$ *basis.)*

Proof Let $f = \sum_{k=1}^\infty a_k\phi_k \in \mathcal{H}(\mathbb{K})$ and $g = \sum_{k=1}^\infty b_k\phi_k \in \mathcal{H}(\mathbb{K})$. Note that $L_{\mathbb{K}^{-1/2}}(f) = \sum_{k=1}^\infty \frac{a_k}{\sqrt{\mu_k}}\phi_k$. As a result

$$\|L_{\mathbb{K}^{-1/2}}(f)\|_{\mathcal{L}^2}^2 = \sum_{k=1}^\infty \frac{a_k^2}{\mu_k} = \|f\|_{\mathcal{H}(\mathbb{K})}.$$

So $L_{\mathbb{K}^{-1/2}}(f) \in \mathcal{L}^2$. Then

$$\langle f, g\rangle_{\mathcal{H}(\mathbb{K})} = \sum_{k=1}^\infty \frac{a_k b_k}{\mu_k} = \langle L_{\mathbb{K}^{-1/2}}(f), L_{\mathbb{K}^{-1/2}}(g)\rangle_{\mathcal{L}^2}.$$

In addition, from the above equality,

$$\langle L_{\mathbb{K}^{1/2}}(\Phi_i), L_{\mathbb{K}^{1/2}}(\Phi_j)\rangle_{\mathcal{H}(\mathbb{K})} = \langle L_{\mathbb{K}^{-1/2}}L_{\mathbb{K}^{1/2}}(\Phi_i), L_{\mathbb{K}^{-1/2}}L_{\mathbb{K}^{1/2}}(\Phi_j)\rangle_{\mathcal{L}^2} = \begin{cases} 1 \text{ if } i=j; \\ 0 \text{ if } i \neq j. \end{cases}$$

■

Lemma 22 *For any* $\beta \in \mathcal{H}(\mathbb{K})$*, it holds that*

$$\sum_{k=1}^{\infty}\langle L_{\mathbb{K}^{-1/2}}(\beta), \Phi_k\rangle_{\mathcal{L}^2}^2 = \|\beta\|_{\mathcal{H}(\mathbb{K})}^2.$$

If in addition $\beta \in \mathcal{F}^{\perp}$*, where*

$$\mathcal{F}^{\perp} := span\,\{L_{\mathbb{K}^{1/2}}(\Phi_k)\}_{k=D+1}^{\infty},$$

and D *is any positive integer, then*

$$\sum_{k=D+1}^{\infty}\langle L_{\mathbb{K}^{-1/2}}(\beta), \Phi_k\rangle_{\mathcal{L}^2}^2 = \|\beta\|_{\mathcal{H}(\mathbb{K})}^2$$

Proof Denote $\beta = \sum_{k=1}^{\infty} L_{\mathbb{K}^{1/2}}(\Phi_k)b_k$. Then $\|\beta\|_{\mathcal{H}(\mathbb{K})}^2 = \sum_{k=1}^{\infty} b_k^2$. So

$$\sum_{k=1}^{\infty}\langle L_{\mathbb{K}^{-1/2}}(\beta), \Phi_k\rangle_{\mathcal{L}^2}^2 = \sum_{k=1}^{\infty} b_k^2 = \|\beta\|_{\mathcal{H}(\mathbb{K})}^2.$$

For the second part, it suffices to observe that $\beta = \sum_{k=D+1}^{\infty} L_{\mathbb{K}^{1/2}}(\Phi_k)b_k$. ■

Suppose $W(r)$ is any Gaussian process with covariance operator

$$\mathbb{E}(W(r)W(s)) = \Sigma_W(r,s),$$

and the eigen-expansion of Σ_W satisfies

$$\Sigma_W(r,s) = \sum_{k=1}^{\infty}\sigma_k^2 h_k(r)h_k(s).$$

Then it holds that

$$W(r) = \sum_{k=1}^{\infty}\alpha_k h_k(r),$$

where $\{\alpha_k\}_{k=1}^{\infty}$ are independent Gaussian random variables with mean 0 and variance σ_k^2. Denote $f(r) = \sum_{k=1}^{\infty} b_k h_k(r) \in \mathcal{L}^2$. Then $\langle W, f\rangle_{\mathcal{L}^2}$ is Gaussian with mean 0 and variance

$$\sum_{k=1}^{\infty} b_k^2\sigma_k^2 = \mathbb{E}(\langle W, f\rangle_{\mathcal{L}^2}^2) = \iint_{[0,1]^2} f(r)\Sigma_W(r,s)f(s)\,\mathrm{d}r\,\mathrm{d}s = \langle L_{\Sigma_W}(f), f\rangle_{\mathcal{L}^2}.$$

Lemma 23 *Suppose R is symmetric. Let $W = L_R(X)$. Then*

$$\mathbb{E}(W(s)W(r)) = (R\Sigma_X R)(s,r).$$

Thus if X is a Gaussian process, then $L_R(X)$ is also a Gaussian process with covariance function $R\Sigma_X R$.

Proof For the first part, it suffices to observe that

$$\mathbb{E}(W(s)W(r)) = \iint_{[0,1]^2} R(s,u)\mathbb{E}(X(u)X(v))R(r,v)\,\mathrm{d}u\,\mathrm{d}v = \iint_{[0,1]^2} R(s,u)\Sigma_X(u,v)R(r,v)\,\mathrm{d}u\,\mathrm{d}v.$$

For the second part, it suffices to observe that $\langle L_R(X), f\rangle_{\mathcal{L}^2} = \langle X, L_R(f)\rangle_{\mathcal{L}^2}$ is Gaussian with standard deviation

$$\sqrt{\mathbb{E}(\langle X, L_R(f)\rangle^2_{\mathcal{L}^2})} = \sqrt{\mathbb{E}(\langle L_R(X), f\rangle^2_{\mathcal{L}^2})} = \sqrt{\iint_{[0,1]^2} f(s)(R\Sigma_X R)(s,r)f(r)\,\mathrm{d}s\,\mathrm{d}r}.$$

■

Lemma 24 *Let X be a centered Gaussian process. Suppose eigenvalues of $\{\xi_k\}_{k=1}^{\infty}$ of the linear operator $L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}$ satisfy*

$$\xi_k \asymp k^{-2r}$$

for some $r > 1/2$. Let $w_k = \frac{\langle L_{\mathbb{K}^{1/2}}(X), \Phi_k\rangle_{\mathcal{L}^2}}{\sqrt{\xi_k}}$. Then w_k is Gaussian with variance 1.

Proof Since $\langle L_{\mathbb{K}^{1/2}}(X), \Phi_k\rangle_{\mathcal{L}^2}$ is Gaussian with standard deviation

$$\sqrt{\iint_{[0,1]^2} \Phi_k(s)L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}(s,r)\Phi_k(r)\,\mathrm{d}s\,\mathrm{d}r} = \sqrt{\xi_k}.$$

So w_k is $\mathcal{N}(0,1)$. ■

Lemma 25 *Under the same conditions as in Theorem 24, it holds that*

$$\Sigma_X[L_{\mathbb{K}^{1/2}}(\Phi_k), L_{\mathbb{K}^{1/2}}(\Phi_l)] = \begin{cases} \xi_k, & \textit{if } k = l; \\ 0, & \textit{if } k \neq l; \end{cases} \quad k, l = 1, 2, 3, \ldots$$

Proof It suffices to observe that

$$\Sigma_X[L_{\mathbb{K}^{1/2}}(\Phi_k), L_{\mathbb{K}^{1/2}}(\Phi_l)] = \langle L_{\Sigma_X}(L_{\mathbb{K}^{1/2}}\Phi_k), L_{\mathbb{K}^{1/2}}\Phi_l\rangle_{\mathcal{L}^2} = \langle L_{\mathbb{K}^{1/2}\Sigma_X\mathbb{K}^{1/2}}\Phi_k, \Phi_l\rangle_{\mathcal{L}^2} = \langle \xi_k\Phi_k, \Phi_l\rangle_{\mathcal{L}^2}.$$

■

Appendix D. Proof of the lower bound result

Proof [Proof of Theorem 7] Since the proof is about lower bounds, it suffices to assume that $1 \leq \mathfrak{s} < p/3$ and $T \geq \mathfrak{s}^2 \log(2p)/24$.

Step 1. Consider the following two special cases of model (1):

$$Y(r) = \int_{[0,1]} A^*(r,s)X(s)\,\mathrm{d}s + \epsilon(r), \quad r \in [0,1] \tag{80}$$

$$Y(r) = \sum_{j=1}^{p} Z_j \beta_j^*(r) + \epsilon(r), \quad r \in [0,1]. \tag{81}$$

Denote

$$\mathcal{E}^*(\widehat{A}) = \int_{[0,1]} \left\{ \iint_{[0,1]\times[0,1]} \Delta_A(r,s_1)\Sigma_X(s_1,s_2)\Delta_A(r,s_2)\,\mathrm{d}s_1\,\mathrm{d}s_2 \right\} \mathrm{d}r$$

and

$$\mathcal{E}^*(\widehat{\beta}) = \int_{[0,1]} \Delta_\beta^\top(r)\Sigma_Z\Delta_\beta(r)\,\mathrm{d}r,$$

as the excess risks of model (80) and model (81) respectively. Since both model (80) and model (81) are special cases of model (1), standard minimax analysis shows that

$$\begin{aligned}\inf_{\widehat{A},\widehat{\beta}} \sup_{\substack{A^*\in\mathcal{C}_A,\\ \beta^*\in\mathcal{C}_\beta}} \mathbb{E}\{\mathcal{E}^*(\widehat{A},\widehat{\beta})\} &\geq \max\left\{ \inf_{\widehat{A},\widehat{\beta}} \sup_{\substack{A^*\in\mathcal{C}_A,\\ \beta^*=0}} \mathbb{E}\{\mathcal{E}^*(\widehat{A})\},\ \inf_{\widehat{A},\widehat{\beta}} \sup_{\substack{A^*=0\\ \beta^*\in\mathcal{C}_\beta}} \mathbb{E}\{\mathcal{E}^*(\widehat{\beta})\} \right\} \\ &= \max\left\{ \inf_{\widehat{A}} \sup_{\substack{A^*\in\mathcal{C}_A,\\ \beta^*=0}} \mathbb{E}\{\mathcal{E}^*(\widehat{A})\},\ \inf_{\widehat{\beta}} \sup_{\substack{A^*=0,\\ \beta^*\in\mathcal{C}_\beta}} \mathbb{E}\{\mathcal{E}^*(\widehat{\beta})\} \right\}.\end{aligned}$$

By the above arguments, the task of finding a lower bound on the excess risk $\mathcal{E}^*(\widehat{A},\widehat{\beta})$ can be separated into two tasks of finding lower bounds on

$$\inf_{\widehat{A}} \sup_{\substack{A^*\in\mathcal{C}_A,\\ \beta^*=0}} \mathbb{E}\{\mathcal{E}^*(\widehat{A})\} \quad \text{and} \quad \inf_{\widehat{\beta}} \sup_{\substack{A^*=0,\\ \beta^*\in\mathcal{C}_\beta}} \mathbb{E}\{\mathcal{E}^*(\widehat{\beta})\}. \tag{82}$$

Theorem 2 in Sun et al. (2018) shows that under the same eigen-decay assumption in condition (9) of Theorem 4, it holds that

$$\inf_{\widehat{A}} \sup_{\substack{A^*\in\mathcal{C}_A,\\ \beta^*=0}} \mathbb{E}\{\mathcal{E}^*(\widehat{A})\} \geq cT^{-\frac{2r}{2r+1}}, \tag{83}$$

where $c > 0$ is a sufficiently small constant. Therefore, to provide a lower bound on the excess risk $\mathcal{E}^*(\widehat{A},\widehat{\beta})$, the only remaining task is to provide a lower bound of $\inf_{\widehat{\beta}} \sup_{\substack{A^*=0,\\ \beta^*\in\mathcal{C}_\beta}} \mathbb{E}\{\mathcal{E}^*(\widehat{\beta})\}$ based on model (81).

Step 2. For $0 < \delta < \sqrt{1/(2\mathfrak{s})}$ to be specified later, let $\widetilde{\Theta}(\delta, \mathfrak{s})$ denote

$$\mathcal{B}(\mathfrak{s}) = \left\{ \{\beta_j\}_{j=1}^p \subset \mathcal{H}_1;\ \sum_{j=1}^p \mathbb{1}\{\beta_j \neq 0\} \leq \mathfrak{s},\ \max_{j=1,\ldots,p} \|\beta_j\|_{\mathcal{H}_1} \leq \delta\sqrt{2/\mathfrak{s}} \right\}, \quad \mathfrak{s} \in \{1, \ldots, p\}.$$

Since $\delta\sqrt{2\mathfrak{s}} \leq 1$, $\mathcal{B}(\mathfrak{s}) \subset \mathcal{C}_\beta$ with $C_\beta = 1$. Let $\phi \in \mathcal{H}(\mathbb{K}_\beta)$ be the leading eigenfunction of $\mathcal{H}(\mathbb{K}_\beta)$. Since $\mathcal{H}(\mathbb{K}_\beta)$ is generated by a bounded kernel, we have that $\|\phi\|_\infty \leq 1$ and $\|\phi\|_{\mathcal{L}^2} = 1$. For any $\theta \in \mathbb{R}^p$ such that $\|\theta\|_\infty \leq 1$, denote $\beta^\theta = (\beta_j^\theta,\, j = 1, \ldots, p) \subset \mathcal{H}(\mathbb{K}_\beta)$ be such that

$$\beta_j^\theta(\cdot) = \theta_j \phi(\cdot), \quad j = 1, \ldots, p.$$

Provided that $\|\theta\|_0 \leq \mathfrak{s}$, we have that $\beta^\theta \in \mathcal{B}(\mathfrak{s}) \subset \mathcal{C}_\beta$. For $t \in \{1, \ldots, T\}$, let $\epsilon_t(s) = \varepsilon_t \phi(s)$, where $\{\varepsilon_t\}_{t=1}^T| \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We thus have ε_t is a Gaussian process with bounded second moment.

Denote P_θ as the joint distribution of $\{Y_t, Z_t\}_{t=1}^T$ with $\beta^* = \beta^\theta$. Since P_θ is supported on the subspace spanned by ϕ, for any θ, θ', the Kullback–Leibler divergence between P_θ and $P_{\theta'}$ is

$$\mathrm{KL}(P_\theta | P_{\theta'}) = \mathbb{E}_{P_\theta} \left\{ \log\left(\frac{\mathrm{d}P_\theta}{\mathrm{d}P_{\theta'}} \right) \right\} = \mathbb{E}_{Q_\theta} \left\{ \log\left(\frac{\mathrm{d}Q_\theta}{\mathrm{d}Q_{\theta'}} \right) \right\},$$

where Q_θ is the joint distribution of $\{y_t, Z_t\}_{t=1}^T$ with

$$y_t = \langle Z_t, \theta \rangle_p + \varepsilon_t, \quad t \in \{1, \ldots, T\}.$$

Since $\{Z_t\}_{t=1}^T$ and $\{\varepsilon_t\}_{t=1}^T$ are independent, we have that

$$\log\left(\frac{\mathrm{d}Q_\theta}{\mathrm{d}Q_{\theta'}} \right) = \sum_{t=1}^T \left(y_t \langle Z_t, \theta \rangle_p - \frac{1}{2}\langle Z_t, \theta \rangle_p^2 - y_t \langle Z_t, \theta' \rangle_p + \frac{1}{2}\langle Z_t, \theta' \rangle_p^2 \right).$$

Conditioning on $\{Z_t\}_{t=1}^T$, we have that $y_t \sim \mathcal{N}(\langle Z_t, \theta \rangle_p, 1)$. Therefore,

$$\begin{aligned}\mathrm{KL}(P_\theta | P_{\theta'}) =& \mathbb{E}_{Q_\theta} \left\{ \log\left(\frac{\mathrm{d}Q_\theta}{\mathrm{d}Q_{\theta'}} \right) \right\} = \mathbb{E}_{\{Z_t\}_{t=1}^T} \left\{ \mathbb{E}_{Q_\theta | \{Z_t\}_{t=1}^T} \log\left(\frac{\mathrm{d}Q_\theta}{\mathrm{d}Q_{\theta'}} \right) \right\} \\ =& \mathbb{E}_{\{Z_t\}_{t=1}^T} \left(\sum_{t=1}^T \langle Z_t, \theta - \theta' \rangle_p^2 \right) = T(\theta - \theta')^\top \Sigma_Z (\theta - \theta') = T\|\theta - \theta'\|_2^2.\end{aligned}$$

For $0 < \delta < \sqrt{1/(2\mathfrak{s})}$ to be specified later, let $\widetilde{\Theta}(\delta, \mathfrak{s})$ be defined as in Theorem 26. Then for any $\theta \in \widetilde{\Theta}(\delta, \mathfrak{s})$, we have that

$$\|\theta\|_1 \leq \delta\sqrt{2\mathfrak{s}} \leq 1$$

and thus $\beta(\theta) \in C_\beta$. For $\theta \neq \theta'$ in $\widetilde{\Theta}(\delta, \mathfrak{s})$, it holds that

$$\sum_{j=1}^p \|\beta_j(\theta) - \beta_j(\theta')\|_{\mathcal{L}^2}^2 = \sum_{j=1}^p (\theta_j - \theta_j')^2 \|\phi\|_{\mathcal{L}^2}^2 = \|\theta - \theta'\|_2^2 \geq \delta^2$$

and that

$$\mathrm{KL}(P_\theta | P_{\theta'}) = T\|\theta - \theta'\|_2^2 \leq 8T\delta^2.$$

Using Fano's lemma (e.g. Yu, 1997), we have that

$$\mathbb{P}\left(\inf_{\widehat{\beta}} \sup_{\beta^* \in \mathcal{B}(s)} \sum_{j=1}^p \|\widehat{\beta}_j - \beta_j^*\|_{\mathcal{L}^2}^2 \geq \delta^2/4 \right) \geq 1 - \frac{\frac{2}{M(M-1)} \sum_{\substack{\theta,\theta' \in \widetilde{\Theta}(\delta,s) \\ \theta \neq \theta'}} \mathrm{KL}(P_\theta | P_{\theta'}) + \log(2)}{\log(M)},$$

where

$$M = |\widetilde{\Theta}(\delta, \mathfrak{s})| \geq \exp\left\{ \frac{s}{2} \log\left(\frac{p - \mathfrak{s}}{\mathfrak{s}/2} \right) \right\}.$$

Note that

$$\frac{\frac{2}{M(M-1)} \sum_{\substack{\theta,\theta' \in \widetilde{\Theta}(\delta,\mathfrak{s}) \\ \theta \neq \theta'}} \mathrm{KL}(P_\theta | P_{\theta'}) + \log(2)}{\log(M)} \leq \frac{8T\delta^2 + \log(2)}{\frac{\mathfrak{s}}{2} \log(\frac{p-\mathfrak{s}}{\mathfrak{s}/2})}.$$

Provided that $\delta^2 = \frac{\mathfrak{s}}{48T} \log(\frac{p-\mathfrak{s}}{\mathfrak{s}/2})$, we have that

$$\mathbb{P}\left\{ \inf_{\widehat{\beta}} \sup_{\beta^* \in \mathcal{B}(\mathfrak{s})} \sum_{j=1}^p \|\widehat{\beta}_j - \beta_j^*\|_{\mathcal{L}^2}^2 \geq \frac{\mathfrak{s}}{200T} \log\left(\frac{p - \mathfrak{s}}{\mathfrak{s}/2} \right) \right\} \geq 1 - 1/2.$$

Since $T \geq \frac{\mathfrak{s}^2 \log(2p)}{24}$, it holds that

$$2\mathfrak{s}\delta^2 = \frac{\mathfrak{s}^2}{24T} \log\left(\frac{p - \mathfrak{s}}{\mathfrak{s}/2} \right) \leq \frac{\mathfrak{s}^2 \log(2p)}{24T} \leq 1,$$

hence $\delta < \sqrt{1/(2\mathfrak{s})}$ as desired. Since Σ_Z is positive definite , it holds that $v^\top \Sigma_Z v \geq \kappa \|v\|^2$ for any $v \in \mathbb{R}^p$.
Finally, we have that

$$\mathbb{P}\left\{ \inf_{\widehat{\beta}} \sup_{\beta^* \in \mathcal{B}(\mathfrak{s})} \sum_{j=1}^p \int_{[0,1]} \Delta_\beta^\top(r) \Sigma_Z \Delta_\beta(r) \, \mathrm{d}r \geq \frac{\mathfrak{s}}{200T} \log\left(\frac{p - \mathfrak{s}}{\mathfrak{s}/2} \right) \right\} \geq 1/2,$$

which directly implies the desired result. ■

Lemma 26 *For every $\delta > 0$ and $\mathfrak{s} < p/3$, denote*

$$\Theta(\delta, s) = \left\{ \theta \in \{ -\delta\sqrt{2/\mathfrak{s}}, 0, \delta\sqrt{2/\mathfrak{s}} \}^p : \|\theta\|_0 = \mathfrak{s} \right\}.$$

There exists a set $\widetilde{\Theta}(\delta, \mathfrak{s}) \subset \Theta(\delta, \mathfrak{s})$, *such that*

$$|\widetilde{\Theta}(\delta, \mathfrak{s})| \geq \exp\left\{\frac{\mathfrak{s}}{2}\log\left(\frac{p-\mathfrak{s}}{\mathfrak{s}/2}\right)\right\}.$$

This implies that for any $\theta, \theta' \in \widetilde{\Theta}(\delta, \mathfrak{s})$, $\theta \neq \theta'$, *it holds that*

$$\|\theta - \theta'\|_2^2 \geq \delta^2 \quad \text{and} \quad \|\theta - \theta'\|_2^2 \leq 8\delta^2.$$

Proof Theorem 26 is Lemma 4 in Raskutti et al. (2011). ■

Appendix E. Additional technical results

Theorem 27 *Suppose that* $Z_i \sim \mathcal{N}(0, \Sigma_Z)$. *Then there exist universal constants* $c, C > 0$ *such that*

$$\mathbb{P}\left(\frac{1}{T}\sum_{t=1}^{T}(Z_t^\top v)^2 \geq \frac{2}{3}v^\top \Sigma_Z v - C\frac{\log(p)}{T}\|v\|_1^2 \quad \text{for all } v \in \mathbb{R}^p\right) \leq \exp(-cT).$$

Proof This is the well known Restricted Eigenvalue condition. The proof can be found in Theorem 1 of Raskutti et al. (2010). See Loh and Wainwright (2012) for better constants. ■

Lemma 28 *Suppose* $\{w_i\}_{i=1}^n$ *are i.i.d. centered sub-Gaussian random variables with variance 1 (denoted as SG(1)).Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} w_i\right| \geq \gamma\right) \leq 2\exp(-2\gamma^2 n).$$

Lemma 29 *Suppose* $\{z_i\}_{i=1}^n$ *are i.i.d. centered sub-Exponential random variables with parameter* 1. *Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \geq \gamma\right) \leq 2\exp(-2n\min\{\gamma^2, \gamma\}).$$

See Vershynin (2018) for the proofs of these two lemmas.

Lemma 30 *Suppose* $r > 1/2$. *For any* $p \geq D$, *it holds that*

$$\sum_{k=D+1}^{\infty}\frac{\log(k)}{k^{2r}} \leq C_r\left(\frac{\log(p)}{D^{2r-1}} + \frac{1}{p^{r-1/2}}\right).$$

Proof Let $\nu \in \mathbb{R}$ $\nu = r - 1/2 > 0$.

$$\sum_{k=D+1}^{\infty}\frac{\log(k)}{k^{2r}} \leq \sum_{k=D+1}^{p}\frac{\log(k)}{k^{2r}} + \sum_{k=p+1}^{\infty}\frac{\log(k)}{k^{2r}}$$

$$
\begin{aligned}
&\leq \frac{\log(p)}{D^{2r-1}} + C_r \sum_{k=p+1}^{\infty} \frac{k^{\nu}}{k^{2r}} \\
&\leq \frac{\log(p)}{D^{2r-1}} + C_r \frac{1}{p^{2r-1-\nu}},
\end{aligned}
$$

where the second inequality holds because there exists a constant C_r such that $\log(k) \leq C_\nu k^\nu$ for all $k \geq 1$. ■

Theorem 31 (Hanson–Wright) *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector where the components X_i are independent SG(K). Let A be any $n \times n$ matrix. Then for every $t \geq 0$,*

$$
\mathbb{P}\left(\left|X^\top AX - E(X^\top AX)\right| \geq t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_{op}}\right)\right].
$$

Proof The proof of this theorem can be found in Vershynin (2018). ■

Lemma 32 *Suppose X is a centered Gaussian process with covaraince function Σ_X. Suppose that*

$$
\mathbb{E}(\|X\|_{\mathcal{L}^2}^2) = C_X < \infty.
$$

Then there exist absolute constants $c, C > 0$ only depending on C_X such that

$$
\mathbb{P}\bigg(\|X\|_{\mathcal{L}^2}^2 \geq C\eta\bigg) \leq \exp(-c\min\{\eta, \eta^2\}).
$$

Proof Let

$$
\Sigma_X(r,s) = \sum_{k=1}^{\infty} \sigma_k^2 \psi_k^x(r)\psi_k^x(s)
$$

be the eigen-expansion of Σ_X. Let $z_k = \langle X, \psi_k^x \rangle$. Observe that $\{z_k\}_{k=1}^{\infty}$ is a collection of independent Gaussian random variables such that $z_k \sim N(0, \sigma_k^2)$. Since $E(\|X\|_{\mathcal{L}^2}^2) = C_X$, we have that

$$
\mathbb{E}(\|X\|_{\mathcal{L}^2}^2) = \sum_{k=1}^{\infty} \mathbb{E}(z_k^2) = \sum_{k=1}^{\infty} \sigma_k^2 = C_X.
$$

Observe that $\|X\|_{\mathcal{L}^2}^2 = \sum_{k=1}^{\infty} z_k^2$ is a sub-Exponential random variable with parameter $\sum_{k=1}^{\infty} \sigma_k^4 \leq C_X^2$. So by sub-Exponential tail bound, it holds that for any $\eta > 0$,

$$
\mathbb{P}\bigg(\|X\|_{\mathcal{L}^2}^2 \geq C'\{\sum_{k=1}^{\infty}\sigma_k^4\}\eta\bigg) \leq \exp(-c\min\{\eta, \eta^2\}),
$$

where C' and c are absolute constants. ■

Lemma 33 *Suppose X is a centered Gaussian process with*

$$\mathbb{E}(\|X\|^2_{W^{\alpha,2}}) = C_X < \infty.$$

Then there exist absolute constants $c, C > 0$ depending only on C_X such that

$$\mathbb{P}\bigg(\|X\|^2_{W^{\alpha,2}} \geq C\eta \bigg) \leq \exp(-c \min\{\eta, \eta^2\}).$$

Proof Let $\mathbb{K}_\alpha$ be the generating kernel of $W^{\alpha,2}$ with eigen-expansion

$$\mathbb{K}_\alpha(r,s) = \sum_{k=1}^\infty \omega_k \psi_k^\alpha(r)\psi_k^\alpha(s).$$

Note that we have $\omega_k \asymp k^{-2\alpha}$ and $\|\psi_k\|^2_{W^{\alpha,2}} = \omega_k^{-1} \asymp k^{2\alpha}$ due to the property of Sobolev space. Let $b_k = \langle X, \psi_k^\alpha \rangle_{\mathcal{L}^2}$. So $\{b_k\}_{k=1}^\infty$ is a collection of centered correlated Gaussian random variables and $X = \sum_{k=1}^\infty b_k \psi_k^\alpha$.

Step 1. Since $X \in W^{\alpha,2}$ almost surely, by Theorem 20, $Y := L_{\mathbb{K}_\alpha^{-1/2}}(X) \in \mathcal{L}^2$ is well defined. So $X = L_{\mathbb{K}_\alpha^{1/2}}(Y)$ and that

$$\mathbb{E}(\|L_{\mathbb{K}_\alpha^{1/2}}(Y)\|^2_{W^{\alpha,2}}) = C_X.$$

From Theorem 21, $\|L_{\mathbb{K}_\alpha^{1/2}}(Y)\|^2_{W^{\alpha,2}} = \|Y\|^2_{\mathcal{L}^2}$. So $E(\|Y\|^2_{\mathcal{L}^2}) = C_X$. For any generic function $f \in \mathcal{L}^2$ with $\|f\|_{\mathcal{L}^2} < \infty$, then

$$\langle Y, f \rangle_{\mathcal{L}^2} = \bigg\langle L_{\mathbb{K}_\alpha^{-1/2}}\Big(\sum_{k=1}^\infty b_k \psi_k^\alpha\Big), f \bigg\rangle_{\mathcal{L}^2} = \sum_{k=1}^\infty b_k \bigg\langle L_{\mathbb{K}_\alpha^{-1/2}}(\psi_k^\alpha), f \bigg\rangle_{\mathcal{L}^2},$$

where $L_{\mathbb{K}_\alpha^{-1/2}}(\psi_k^\alpha) \in \mathcal{L}^2$ because $\psi_k^\alpha \in W^{\alpha,2}$. Therefore $\langle Y, f \rangle_{\mathcal{L}^2}$ is a centered Gaussian random variable with

$$Var(\langle Y, f \rangle_{\mathcal{L}^2}) = \mathbb{E}(\langle Y, f \rangle^2_{\mathcal{L}^2}) \leq \mathbb{E}\{\|Y\|^2_{\mathcal{L}^2} \|f\|^2_{\mathcal{L}^2}\} = C_X \|f\|^2_{\mathcal{L}^2} < \infty.$$

Step 2. Let Σ_Y be the covariance function of Y. Since $E(\|Y\|^2_{\mathcal{L}^2}) < \infty$, by the Hilbert–Schmidt theorem, the eigen-expansion of Σ_Y can be written as

$$\Sigma_Y(r,s) = \sum_{k=1}^\infty \delta_k^2 \psi_k^y(r)\psi_k^y(s).$$

Then by **Step 1**, $\langle Y, \psi_k^y \rangle_{\mathcal{L}^2}$ is Gaussian and that

$$\mathbb{E}\bigg\{ \langle Y, \psi_k^y \rangle_{\mathcal{L}^2} \langle Y, \psi_l^y \rangle_{\mathcal{L}^2} \bigg\} = \Sigma_Y[\psi_k^y, \psi_l^y] = 0$$

whenever $l \neq k$. So $\{\langle Y, \psi_k^y \rangle_{\mathcal{L}^2}\}_{k=1}^\infty$ is a collection of independent Gaussian random variables. By Theorem 32,

$$\mathbb{P}\bigg(\|Y\|_{\mathcal{L}^2}^2 \geq C\eta \bigg) \leq \exp(-c \min\{\eta, \eta^2\}).$$

The desired result follows by observing that

$$\|Y\|_{\mathcal{L}^2}^2 = \|L_{\mathbb{K}_\alpha^{-1/2}}(X)\|_{\mathcal{L}^2}^2 = \|X\|_{W^{\alpha,2}}^2.$$

■

Lemma 34 *Suppose that $\{X_t\}_{t=1}^T$ is a collection of independent centered Gaussian process with covariance operator Σ_X. Let*

$$\Sigma_X(r,s) = \sum_{k=1}^\infty \sigma_k^2 \psi_k^x(r)\psi_k^x(s)$$

be the eigen expansion of Σ_X. Suppose $\mathbb{E}(\|X_t\|_{\mathcal{L}^2}^2) = \sum_{k=1}^\infty \sigma_k^2 < \infty$ and that $\{\sigma_k\}_{k=1}^\infty$ decay to 0 at polynomial rate. Let $\{z_t\}_{t=1}^T \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. If in addition, $T \geq \log(p)$, then with probability at least $1 - 10p^{-4}$

$$\bigg\| \frac{1}{T}\sum_{t=1}^T X_t z_t - \mathbb{E}(X_1 z_1) \bigg\|_{\mathcal{L}^2} \leq C\sqrt{\frac{\log(p)}{T}}.$$

Proof Let $w_{t,k} = \langle X_t, \psi_k^x \rangle_{\mathcal{L}^2}$. Then $\{w_{t,k}\}_{1\leq t \leq T, 1 \leq k < \infty}$ is a collection of centered independent random variables with $Var(w_{t,k}) = \sigma_k^2$. Note that

$$\bigg\| \frac{1}{T}\sum_{t=1}^T X_t z_t - \mathbb{E}(X_1 z_1) \bigg\|_{\mathcal{L}^2}^2 = \bigg\| \frac{1}{T}\sum_{t=1}^T \sum_{k=1}^\infty \big(w_{t,k} z_t - \mathbb{E}(w_{1,k} z_1)\big)\psi_k^x \bigg\|_{\mathcal{L}^2}^2 = \sum_{k=1}^\infty \bigg(\frac{1}{T}\sum_{t=1}^T \Big\{ w_{t,k} z_t - \mathbb{E}(w_{1,k} z_1) \Big\} \bigg)^2, \tag{84}$$

where the last inequality follows from the fact that $\{\psi_k^x\}$ is a collection of basis functions in $\mathcal{L}^2$.

Note that $w_{t,k} z_t$ is SE with parameter σ_k^2. So

$$\mathbb{P}\bigg(\bigg| \frac{1}{T}\sum_{t=1}^T w_{t,k} z_t - \mathbb{E}(w_{1,k} z_1) \bigg| \geq \eta\sigma_k \bigg) \leq \exp(-n\{\eta, \eta^2\}).$$

Therefore

$$\mathbb{P}\bigg(\bigg| \frac{1}{T}\sum_{t=1}^T w_{t,k} z_t - \mathbb{E}(w_{1,k} z_1) \bigg| \geq C\sigma_k \bigg\{ \sqrt{\frac{\log(p) + \log(k)}{T}} + \frac{\log(p) + \log(k)}{T} \bigg\} \bigg) \leq \frac{1}{k^3 p^4}.$$

Since $\sum_{k=1}^\infty k^{-3} < 3$, with probability at least $3p^{-4}$,

$$\bigg| \frac{1}{T}\sum_{t=1}^T w_{t,k} z_t - \mathbb{E}(w_{1,k} z_1) \bigg| \leq C\sigma_k \bigg\{ \sqrt{\frac{\log(p) + \log(k)}{T}} + \frac{\log(p) + \log(k)}{T} \bigg\} \quad \text{for all } 1 \leq k < \infty.$$

Under this good event, Equation (84) implies that

$$\left\|\frac{1}{T}\sum_{t=1}^{T} X_t z_t - \mathbb{E}(X_1 z_1)\right\|_{\mathcal{L}^2}^2 \leq 2C\sigma_k^2 \left\{\frac{\log(p)+\log(k)}{T} + \left(\frac{\log(p)+\log(k)}{T}\right)^2\right\}. \tag{85}$$

Since $\sum_{k=1}^{\infty} \sigma_k^2 < \infty$, $\sigma_k \asymp k^{-1-\nu}$ for some $\nu > 0$. Therefore $\sum_{k=1}^{\infty} \sigma_k^2 \log^2(k) < \infty$. Since in addition, $\frac{\log(p)}{T} < 1$, Equation (85) implies that

$$\left\|\frac{1}{T}\sum_{t=1}^{T} X_t z_t - \mathbb{E}(X_1 z_1)\right\|_{\mathcal{L}^2}^2 \leq C' \frac{\log(p)}{T}.$$

■

Theorem 35 *Suppose $A : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$ is a compact linear operator. There exist $\{\psi_k\}_{k=1}^{\infty}, \{\omega_k\}_{k=1}^{\infty} \subset \mathcal{H}(\mathbb{K})$, two orthogonal basis in $\mathcal{L}^2$, such that the associated bivariate function A can be written as*

$$A(r,s) = \sum_{k=1}^{\infty} a_k \psi_k(s)\omega_k(r), \quad r, s \in [0,1],$$

where $\{a_k\}_{k=1}^{\infty} \subset \mathbb{R}$.

Proof This is the well known spectral theory for compact operators on Hilbert space. See Chapter 5 of Brezis (2010) for a detailed proof. ■

Lemma 36 *Let $A : \mathcal{H}(\mathbb{K}) \to \mathcal{H}(\mathbb{K})$ be any Hilbert-Schmidt operator. We have that*

$$\max\left\{\sup_{r\in[0,1]} \|A(r,\cdot)\|_{\mathcal{H}(\mathbb{K})}, \sup_{s\in[0,1]} \|A(\cdot,s)\|_{\mathcal{H}(\mathbb{K})}, \|A(\cdot,\cdot)\|_{\infty}\right\} \leq C_{\mathbb{K}} \|A\|_{\mathrm{F}(\mathbb{K})},$$

where $C_{\mathbb{K}}$ is some constant only depending on $\mathbb{K}$.

Proof Since $\mathcal{H}(\mathbb{K})$ is generated by a bounded kernel $\mathbb{K}$, we have that for any $f \in \mathcal{H}(\mathbb{K})$ and any $r \in [0,1]$,

$$f(r) = \langle f, \mathbb{K}(r,\cdot)\rangle_{\mathcal{H}(\mathbb{K})} \leq \|f\|_{\mathcal{H}(\mathbb{K})} \|K(r,\cdot)\|_{\mathcal{H}(\mathbb{K})} = \|f\|_{\mathcal{H}(\mathbb{K})} \sqrt{\langle \mathbb{K}(\cdot,r), \mathbb{K}(\cdot,r)\rangle_{\mathcal{H}(\mathbb{K})}} \leq \|f\|_{\mathcal{H}(\mathbb{K})} \sqrt{\mathbb{K}(r,r)}.$$

Since $\mathbb{K}$ is a bounded kernel, letting $C_{\mathbb{K}} := \sup_r \mathbb{K}(r,r)$, we have that

$$\|f\|_{\infty} \leq \|f\|_{\mathcal{H}(\mathbb{K})} C_{\mathbb{K}}.$$

Without loss of generality, it suffices to assume that $C_{\mathbb{K}} = 1$. By Theorem 35, there exist $\{\psi_k\}_{k=1}^{\infty}, \{\omega_k\}_{k=1}^{\infty} \subset \mathcal{H}(\mathbb{K})$, two orthogonal basis in $\mathcal{L}^2$, and $\{a_k\}_{k=1}^{\infty} \subset \mathbb{R}$, such that

$$A(r,s) = \sum_{k=1}^{\infty} a_k \psi_k(r)\omega_k(s).$$

Since $\|\psi_k\|_{\mathcal{H}(\mathbb{K})} = \|\omega_k\|_{\mathcal{H}(\mathbb{K})} = 1$, $k \in \mathbb{N}_*$, $\|A\|^2_{\mathrm{F}(\mathbb{K})} = \sum_{k=1}^{\infty} a_k^2$.
Observe that for any $f \in \mathcal{H}(\mathbb{K})$ such that $\|f\|_{\mathcal{H}} = 1$, it holds that $f = \sum_{k=1}^{\infty} b_k \omega_k$, where $\sum_{k=1}^{\infty} b_k^2 = 1$. Therefore, for any $r \in [0,1]$, we have that

$$\begin{aligned}
\|A(r,\cdot)\|_{\mathcal{H}(\mathbb{K})} &= \sup_{\|f\|_{\mathcal{H}(\mathbb{K})}=1} \Big\langle \sum_{k=1}^{\infty} a_k \psi_k(r)\omega_k(\cdot), f(\cdot) \Big\rangle_{\mathcal{H}(\mathbb{K})} = \sup_{\sum_{k=1}^{\infty} b_k^2=1} \sum_{k=1}^{\infty} a_k \psi_k(r) b_k \\
&\leq \sup_{\sum_{k=1}^{\infty} b_k^2=1} \sum_{k=1}^{\infty} |a_k||b_k| \leq \sup_{\sum_{k=1}^{\infty} b_k^2=1} \sqrt{\sum_{k=1}^{\infty} a_k^2}\sqrt{\sum_{k=1}^{\infty} b_k^2} = \|A\|_{\mathrm{F}(\mathbb{K})},
\end{aligned}$$

where the first inequality is due to that $\|\psi_k\|_{\infty} \leq \|\psi_k\|_{\mathcal{H}(\mathbb{K})} = 1$, $k \in \mathbb{N}_*$. Similar arguments show that $\sup_{s\in[0,1]} \|A(\cdot,s)\|_{\mathcal{H}(\mathbb{K})} \leq \|A\|_{\mathrm{F}(\mathbb{K})}$.

Moreover, observe that for any fixed $r, s \in [0,1]$, it holds that

$$A(r,s) = \langle A(r,\cdot), \mathbb{K}(s,\cdot)\rangle_{\mathcal{H}(\mathbb{K})}.$$

Therefore it holds that

$$\sup_{s\in[0,1]} |A(r,s)| \leq \sup_{s\in[0,1]} |\langle A(r,\cdot), \mathbb{K}(s,\cdot)\rangle_{\mathcal{H}(\mathbb{K})}| \leq \|A(r,\cdot)\|_{\mathcal{H}(\mathbb{K})} \sup_{s\in[0,1]} \|\mathbb{K}(s,\cdot)\|_{\mathcal{H}(\mathbb{K})} \leq \|A(r,\cdot)\|_{\mathcal{H}(\mathbb{K})},$$

where $\|\mathbb{K}(s,\cdot)\|^2_{\mathcal{H}(\mathbb{K})} = \mathbb{K}(s,s) \leq 1$, $s \in [0,1]$, is used in the last inequality. Finally, we have that

$$\|A(\cdot,\cdot)\|_{\infty} \leq \sup_{r\in[0,1]} \|A(r,\cdot)\|_{\mathcal{H}(\mathbb{K})} \leq \|A\|_{\mathrm{F}(\mathbb{K})},$$

which concludes the proof. ■

Lemma 37 *It holds that*

$$\mathcal{E}^*(\widehat{A},\widehat{\beta}) = \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(\int_{[0,1]} \Delta_A(r,s)X^*(s)\,\mathrm{d}s + \langle Z^*, \Delta_\beta(r)\rangle_p\right)^2 \mathrm{d}r\right\}.$$

Proof We have that

$$\begin{aligned}
\mathcal{E}^*(\widehat{A},\widehat{\beta}) &= \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(Y^*(r) - \int_{[0,1]} \widehat{A}(r,s)X^*(s)\,\mathrm{d}s - \langle Z^*, \widehat{\beta}(r)\rangle_p\right)^2 \mathrm{d}r\right\} \\
&\quad - \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(Y^*(r) - \int_{[0,1]} A^*(r,s)X^*(s)\,\mathrm{d}s - \langle Z^*, \beta^*(r)\rangle_p\right)^2 \mathrm{d}r\right\} \\
=&\mathbb{E}_{X^*,Z^*,Y^*}\bigg\{\int_{[0,1]}\bigg(\int_{[0,1]} A^*(r,s)X^*(s)\,\mathrm{d}s + \langle Z^*, \beta^*(r)\rangle_p + \epsilon_t^*(r) \\
&\quad - \int_{[0,1]} \widehat{A}(r,s)X^*(s)\,\mathrm{d}s - \langle Z^*, \widehat{\beta}(r)\rangle_p\bigg)^2 \mathrm{d}r\bigg\} - \mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]} \{\epsilon_t^*(r)\}^2\,\mathrm{d}r\right\}
\end{aligned}$$

$$
\begin{aligned}
=&\mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s+\langle Z^*,\Delta_\beta(r)\rangle_p\right)^2\,\mathrm{d}r\right\}\\
&+2\mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s+\langle Z^*,\Delta_\beta(r)\rangle_p\right)\epsilon_t^*(r)\,\mathrm{d}r\right\}\\
=&\mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s+\langle Z^*,\Delta_\beta(r)\rangle_p\right)^2\,\mathrm{d}r\right\}\\
&+2\int_{[0,1]}\mathbb{E}_{X^*,Z^*,Y^*}\left(\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s+\langle Z^*,\Delta_\beta(r)\rangle_p\right)\mathbb{E}_{X^*,Z^*,Y^*}\left\{\epsilon_t^*(r)\right\}\,\mathrm{d}r\\
=&\mathbb{E}_{X^*,Z^*,Y^*}\left\{\int_{[0,1]}\left(\int_{[0,1]}\Delta_A(r,s)X^*(s)\,\mathrm{d}s+\langle Z^*,\Delta_\beta(r)\rangle_p\right)^2\,\mathrm{d}r\right\},
\end{aligned}
$$

where the second identity is due to the definition of $Y^*(r)$, the fourth identity is due to the independence between $\epsilon_t(\cdot)$ and the rest, and the final identity is due to the mean-zero assumption of $\epsilon_t^*(\cdot)$. ■

E.1 Properties of Soboblev Spaces

Definition 38 *Let $f:[0,1]\to\mathbb{R}$ be any real function. Then f is said to be Hölder continuous of order $\kappa\in(0,1]$ if*

$$
\sup_{x,x'\in[0,1]}\frac{|f(x)-f(x')|}{|x-x'|^\kappa}<\infty.
$$

Theorem 39 *Suppose that $f:[0,1]\to\mathbb{R}$ is such that $f\in W^{\alpha,2}$ with $\alpha>1/2$. Then f is Hölder continuous with $\kappa=\alpha-1/2$ and there exists an absolute constant C such that*

$$
\sup_{x,x'\in[0,1]}\frac{|f(x)-f(x')|}{|x-x'|^{\alpha-1/2}}<C\|f\|_{W^{\alpha,2}}.
$$

Proof This is the well-known Morrey's inequality. See Evans (2010) for a proof. ■

In the following lemma, we show that the integral of a Hölder smooth function f can be approximated by the discrete sum of f evaluated at the sample points $\{s_i\}_{i=1}^n$.

Lemma 40 *Suppose that $\{s_i\}_{i=1}^n$ is a collection of sample points satisfying Assumption 3***a***. Suppose in addition that $f:[0,1]\to\mathbb{R}$ is Hölder continuous with*

$$
\sup_{x,x'\in[0,1]}\frac{|f(x)-f(x')|}{|x-x'|^{\alpha-1/2}}<C_f.
$$

Let $w_s(i) = (s_i - s_{i-1})n$*. Then there exists an absolute constant* C *such that*

$$\left| \int_{[0,1]} f(s)ds - \frac{1}{n}\sum_{i=1}^n w_s(i) f(s_i) \right| \le C C_f n^{-\alpha+1/2.}$$

Consequently if $f \in W^{\alpha,2}$*, then*

$$\left| \int_{[0,1]} f(s)ds - \frac{1}{n}\sum_{i=1}^n w_s(i) f(s_i) \right| \le C' \|f\|_{W^{\alpha,2}} n^{-\alpha+1/2.}$$

Proof Let $s_0 = 0$. Observe that

$$\begin{aligned}
\left| \int_{[0,1]} f(s)ds - \frac{1}{n}\sum_{i=1}^n w_s(i) f(s_i) \right| =& \left| \sum_{i=1}^n \int_{s_{i-1}}^{s_i} f(s)\,\mathrm{d}s - \frac{1}{n}\sum_{i=1}^n w_s(i) f(s_i) \right| \\
=& \left| \sum_{i=1}^n \int_{s_{i-1}}^{s_i} f(s)\,\mathrm{d}s - \sum_{i=1}^n \int_{s_{i-1}}^{s_i} f(s_i)\,\mathrm{d}s \right| \\
\le& \sum_{i=1}^n \int_{s_{i-1}}^{s_i} \left| f(s) - f(s_i) \right| \mathrm{d}s \\
\le& \sum_{i=1}^n (s_i - s_{i-1}) C_f {C_d}^{\alpha-1/2} n^{-\alpha+1/2} \\
=& C_f C_d^{\alpha-1/2} n^{-\alpha+1/2}.
\end{aligned}$$

■

Lemma 41 *Suppose that* $f, g : [0,1] \to \mathbb{R}$ *are such that* $f, g \in W^{\alpha,2}$ *for some* $\alpha > 1/2$*. Then there exist absolute constants* C_1 *and* C_2 *such that*

$$\|fg\|_{W^{\alpha,2}} \le C_1 \|f\|_{W^{\alpha,2}} \|g\|_{W^{\alpha,2}} \quad \text{and} \quad \|f\|_\infty \le C_2 \|f\|_{W^{\alpha,2}}. \tag{86}$$

Proof These are well know Sobolev inequalities. See Evans (2010) for proofs. ■

Appendix F. Technical details of the optimization

F.1 Detailed formulation of the convex optimization

With the notation defined as in Section 4.1, in the following, we provide the details for rewriting the penalized optimization problem in (13) into (14), which is a convex function of R and $B = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_p]$. Specifically, we examine the three components of (13) one by one.

- The squared loss can be rewritten as

$$\sum_{t=1}^T \sum_{j=1}^{n_2} w_r(j) \left(Y_t(r_j) - \frac{1}{n_1}\sum_{i=1}^{n_1} w_s(i) A(r_j, s_i) X_t(s_i) - \langle \beta(r_j), Z_t \rangle_p \right)^2$$

$$=\sum_{t=1}^{T}\left\|W_R\left(Y_t-\frac{1}{n_1}K_1^\top RK_2W_SX_t-K_1^\top BZ_t\right)\right\|_2^2=\left\|W_R\left(Y-\frac{1}{n_1}K_1RK_2W_SX-K_1BZ\right)\right\|_{\mathrm{F}}^2,$$

where $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ are the ℓ_2-norm of a vector and the Frobenius norm of a matrix.

- As for the Frobenius norm penalty $\|A\|^2_{\mathrm{F}(\mathbb{K})}$, first, it is easy to see $A^\top(r,s)=k_1(s)^\top Rk_2(r)$, where $A^\top$ is the adjoint operator of A. Let $u(s)=k_2(s)^\top c$, for any $c\in\mathbb{R}^{n_1}$. We have that

$$\begin{aligned}&A^\top A[u](s)=\langle A^\top(s,r),A[u](r)\rangle_{\mathcal{H}}=\langle k_1(r)^\top Rk_2(s),A[u](r)\rangle_{\mathcal{H}}\\=&\langle k_2(s)^\top R^\top k_1(r),\langle A(r,s),u(s)\rangle_{\mathcal{H}}\rangle_{\mathcal{H}}=k_2(s)^\top R^\top\langle k_1(r),k_1(r)^\top\rangle_{\mathcal{H}}R\langle k_2(s),u(s)\rangle_{\mathcal{H}}\\=&k_2(s)^\top R^\top K_1RK_2c.\end{aligned}$$

  Thus, the eigenvalues of $A^\top A$ are the same as those of $R^\top K_1RK_2$ and $\|A\|^2_{\mathrm{F}(\mathbb{K})}=\mathrm{tr}(R^\top K_1RK_2)$.

- As for the $\mathcal{H}(\mathbb{K})$-norm penalty $\|\beta_l\|_{\mathcal{H}(\mathbb{K})}$ and the group Lasso-type penalty $\|\beta_l\|_{n_2}$, we have $\|\beta_l\|_{\mathcal{H}(\mathbb{K})}=\sqrt{\langle\beta_l(r),\beta_l(r)\rangle_{\mathcal{H}(\mathbb{K})}}=\sqrt{\mathbf{b}_l^\top K_1\mathbf{b}_l}$ and

$$\begin{aligned}\|\beta_l\|_{n_2}=&\sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2}w_r(j)\beta_l^2(r_j)}=\sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2}w_r(j)\mathbf{b}_l^\top k_1(r_j)k_1(r_j)^\top\mathbf{b}_l}\\=&\sqrt{\mathbf{b}_l^\top\frac{1}{n_2}\sum_{j=1}^{n_2}w_r(j)k_1(r_j)k_1(r_j)^\top\mathbf{b}_l}=\sqrt{\frac{1}{n_2}\mathbf{b}_l^\top K_1W_R^2K_1\mathbf{b}_l},\ \text{for } l=1,\ldots,p.\end{aligned}$$

Combining all three components together, it is easy to see that the optimization problem in (13) can be written as (14).

F.2 Detailed optimization of the structured ridge regression (16)

By simple linear algebra, the first order condition forms the linear system

$$(S_1^\top S_1+\lambda_1I_{n_1n_2})\mathbf{e}=S_1^\top\widetilde{\mathbf{y}}.$$

Denote

$$S(\lambda_1)=S_1^\top S_1+\lambda_1I_{n_1n_2}.$$

Note that $S(\lambda_1)$ is a positive definite matrix and the optimization has a unique solution. However, when n_1n_2 is large, the inverse of the matrix $S(\lambda_1)$ becomes computationally expensive. We thus further exploit the structure of $S(\lambda_1)$.

Denote $S_4=\frac{1}{n_1}K_2^{1/2}W_SX\in\mathbb{R}^{n_1\times T}$, we have $S_1^\top S_1=(S_4S_4^\top)\otimes(K_1^{1/2}W_R^2K_1^{1/2})$, and thus

$$S(\lambda_1)=(S_4S_4^\top)\otimes(K_1^{1/2}W_R^2K_1^{1/2})+\lambda_1I_{n_1n_2}.$$

Denote the singular value decomposition (SVD) of $S_4S_4^\top = \frac{1}{n_1^2}K_2^{1/2}W_SXX^\top W_SK_2^{1/2}$ as $S_4S_4^\top = U_1D_1U_1^\top$ and the SVD of $K_1^{1/2}W_R^2K_1^{1/2}$ as $K_1^{1/2}W_R^2K_1^{1/2} = U_2D_2U_2^\top$, we have

$$S(\lambda_1) = (S_4S_4^\top) \otimes (K_1^{1/2}W_R^2K_1^{1/2}) + \lambda_1 I_{n_1n_2} = (U_1 \otimes U_2)(D_1 \otimes D_2 + \lambda_1 I_{n_1n_2})(U_1^\top \otimes U_2^\top).$$

Thus, we have

$$\begin{aligned}
S^{-1}(\lambda_1) &= (U_1 \otimes U_2)(D_1 \otimes D_2 + \lambda_1 I_{n_1n_2})^{-1}(U_1^\top \otimes U_2^\top) \\
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{\lambda_1 + D_{1i}D_{2j}}(U_{1i} \otimes U_{2j})(U_{1i}^\top \otimes U_{2j}^\top) \\
&= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{\lambda_1 + D_{1i}D_{2j}}(U_{1i}U_{1i}^\top) \otimes (U_{2j}U_{2j}^\top).
\end{aligned}$$

Thus, we have

$$\mathbf{e} = S^{-1}(\lambda_1)S_1^\top \widetilde{\mathbf{y}} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{\lambda_1 + D_{1i}D_{2j}}(U_{1i}U_{1i}^\top S_4) \otimes (U_{2j}U_{2j}^\top K_1^{1/2}W_R)\widetilde{\mathbf{y}},$$

which implies that

$$\begin{aligned}
E =& \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{\lambda_1 + D_{1i}D_{2j}}(U_{2j}U_{2j}^\top K_1^{1/2}W_R)\widetilde{Y}(S_4^\top U_{1i}U_{1i}^\top) \\
=& \frac{1}{n_1}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{\lambda_1 + D_{1i}D_{2j}}(U_{2j}U_{2j}^\top)(K_1^{1/2}W_R\widetilde{Y}X^\top W_SK_2^{1/2})(U_{1i}U_{1i}^\top),
\end{aligned}$$

and we update $R = K_1^{-1/2}EK_2^{-1/2}$.

Appendix G. Additional numerical analysis

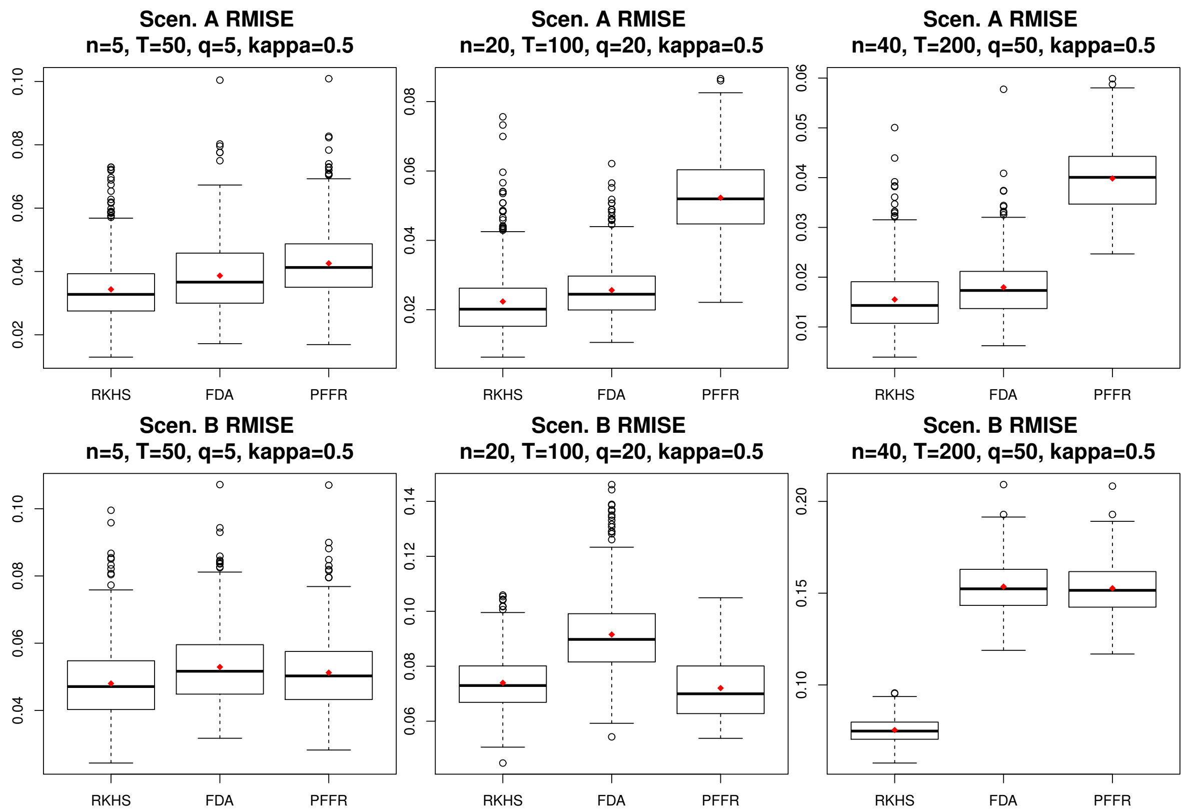

Figure 7: Boxplot of RMISE of RKHS, FDA and PFFR across 500 experiments under function-on-function regression with $\kappa = 0.5$. Red points denote the average RMISE.

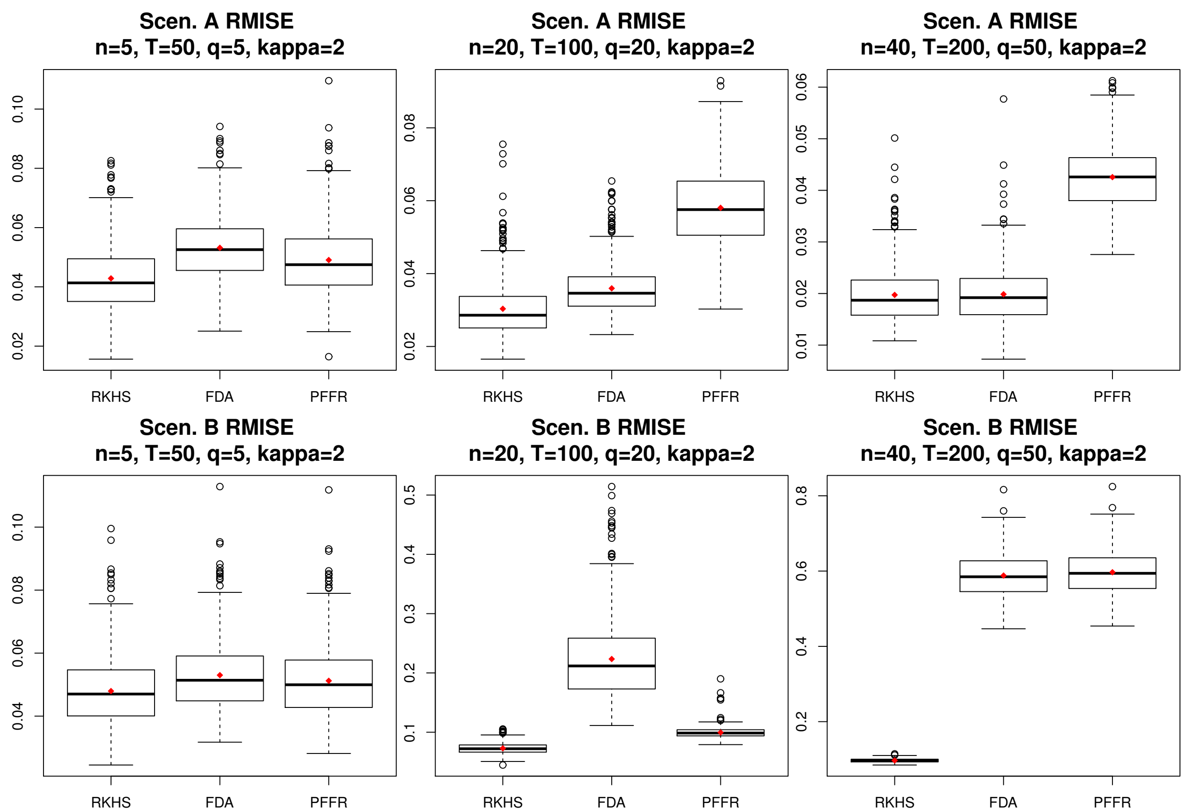

Figure 8: Boxplot of RMISE of RKHS, FDA and PFFR across 500 experiments under function-on-function regression with $\kappa = 2$. Red points denote the average RMISE.

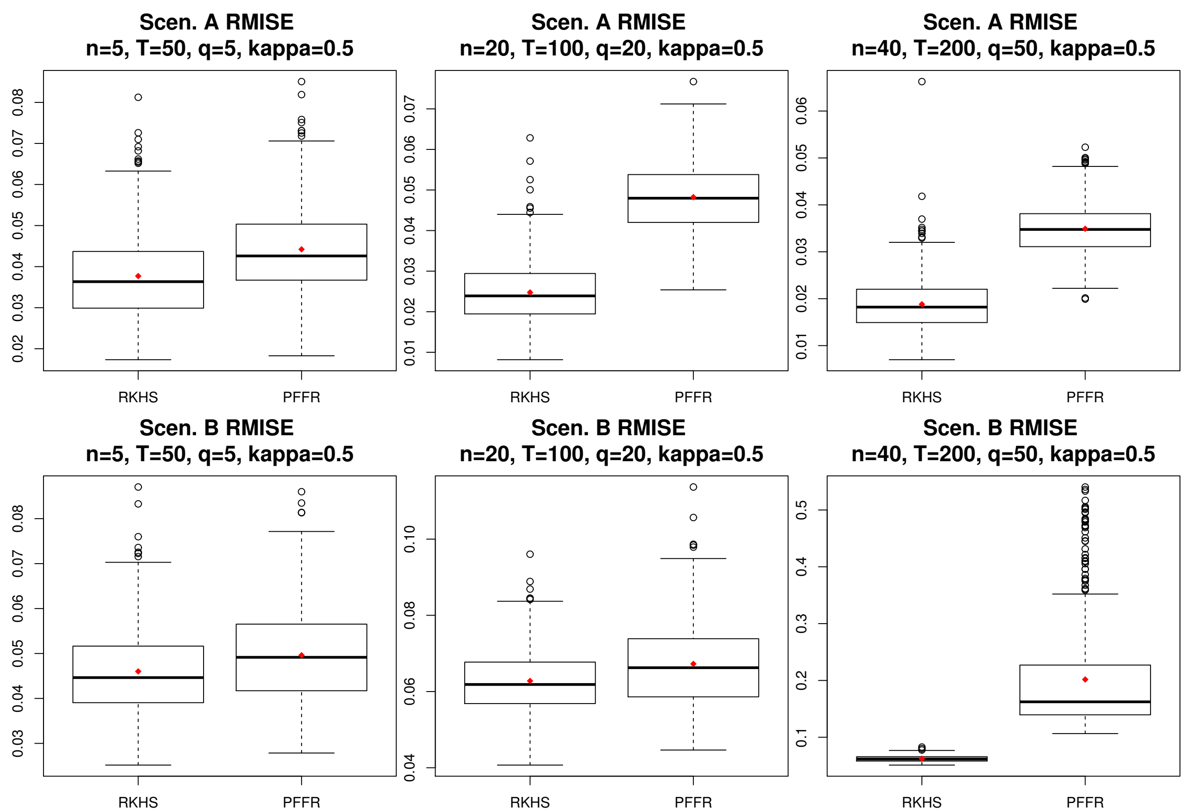

Figure 9: Boxplot of RMISE of RKHS and PFFR across 500 experiments under mixed functional regression with $\kappa = 0.5$. Red points denote the average RMISE.

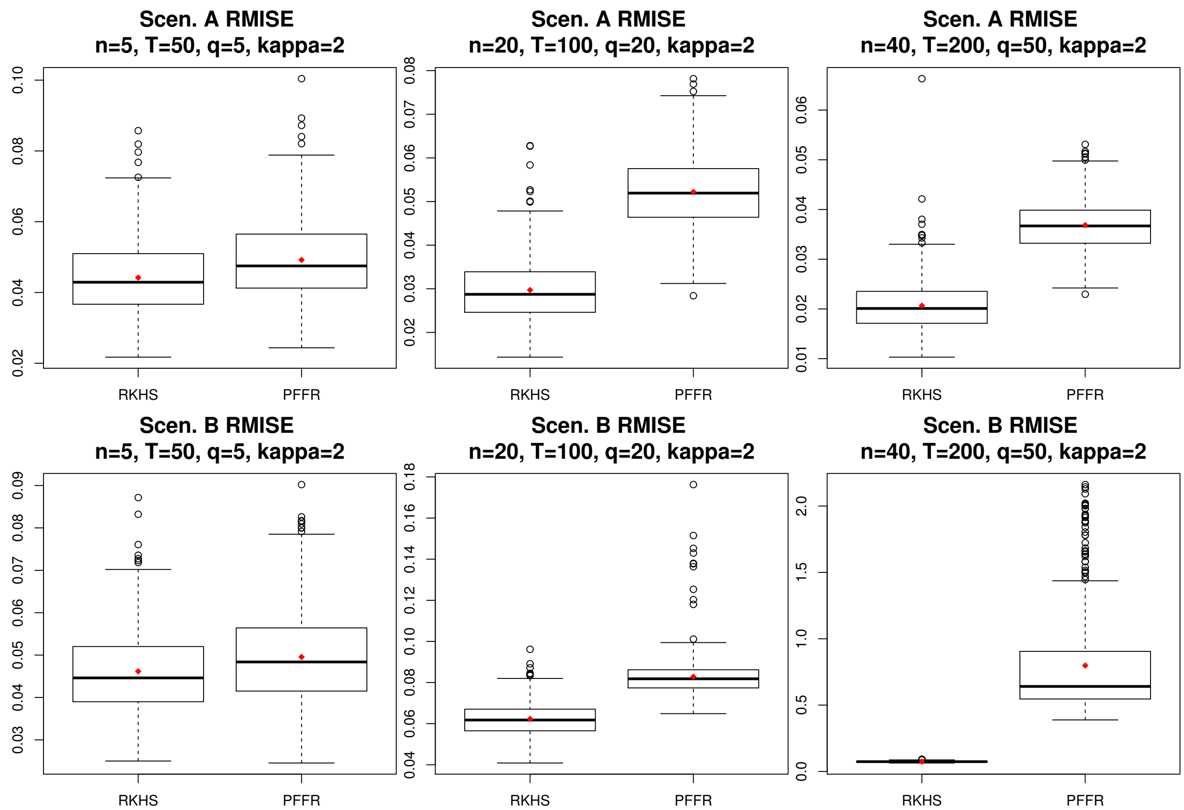

Figure 10: Boxplot of RMISE of RKHS and PFFR across 500 experiments under mixed functional regression with $\kappa = 2$. Red points denote the average RMISE.

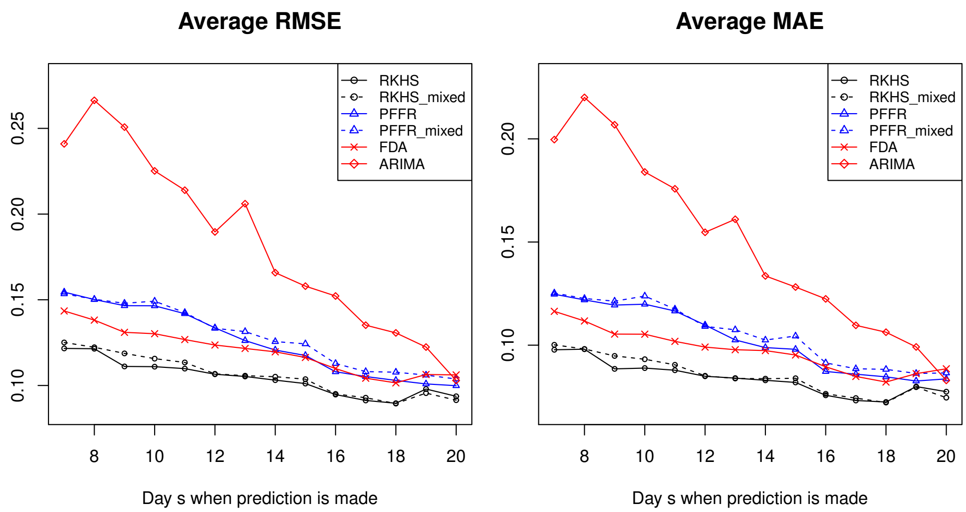

Figure 11: Average RMSE and MAE achieved by different functional regression methods. For campaign t at prediction time s, ARIMA is trained on $\{N_t(r_i), r_i \in [0, s]\}$ for the prediction of $\{N_t(r_i), r_i \in (s, 30]\}$. Note that ARIMA can be used as $\{N_t(r_i)\}_{i=1}^{60}$ can be seen as an evenly spaced univariate time series.

Appendix H. Robustness check for real data analysis

In this section, we further provide a robustness check for the real data analysis presented in Section 5.4. Specifically, we repeat the two-fold CV procedure in Section 5.4 independently for 100 times for RKHS, FDA and PFFR[3]. In addition, we also implement FDA with 50 basis functions (FDA50) and PFFR with 30 basis functions (PFFR30). We further compare with the popular functional PCA based regression method (FPCA) proposed in Yao et al. (2005).

Thus, for every method (RKHS, FDA, PFFR, FPCA), on each prediction day $s \in \{7, 8, 9, \cdots, 20\}$, we have 100 RMSE_s and 100 MAE_s obtained from the 100 times two-fold CV. We refer to Section 5.4 for more detailed definition of RMSE and MAE.

Figure 12 visualizes the average RMSE_s and MAE_s over the 100 times CV achieved by different functional regression methods across $s \in \{7, 8, 9, \cdots, 20\}$. As can be seen clearly, the pattern exhibited in Figure 12 is the same as the one in Figure 5 in Section 5.4 and RKHS is the winner by a notable margin. Note that FDA20 and FDA50 give essentially the same performance, while PFFR30 indeed gives worse performance than PFFR20. This suggests that 20 basis dimension is sufficient for the current real data application.

Table 3 further gives the number of two-fold CV (out of 100 times) where RKHS achieves the smallest RMSE_s and MAE_s for each $s \in \{7, 8, 9, \cdots, 20\}$. As can be seen, RKHS is almost always the winner for every $s \in \{7, 8, 9, \cdots, 20\}$ except for $s = 19$, where RKHS still provides the best performance more than two thirds of the time.

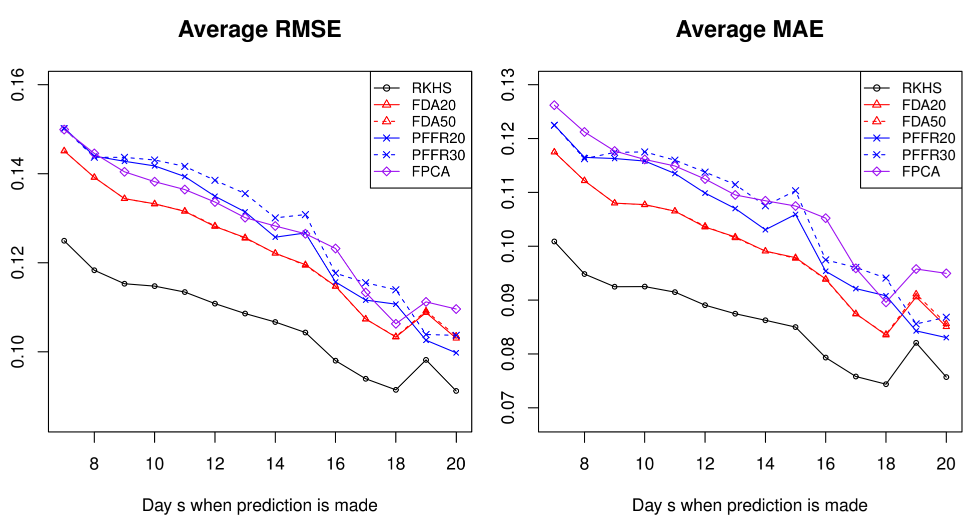

Figure 12: RMSE_s and MAE_s achieved by different functional regression methods averaged over 100 two-fold cross-validations for each $s \in \{7, 8, 9, \cdots, 20\}$.

3. For simplicity, we do not include $\text{RKHS}_{\text{mixed}}$ and $\text{PFFR}_{\text{mixed}}$ in the comparison, as the result in Section 5.4 (see Figure 5) clearly indicates that the three scalar predictors do not seem to improve the prediction performance.

s	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RMSE_s	100	100	100	100	100	100	100	100	100	100	100	100	77	97
MAE_s	100	100	100	100	100	100	100	100	100	100	100	100	66	96

Table 3: Number of two-fold cross-validations (out of 100 times) where RKHS achieves the smallest RMSE_s and MAE_s for each $s \in \{7, 8, 9, \cdots, 20\}$.

Appendix I. Numerical results for observations with measurement error

In this section, we conduct simulation experiments to examine the performance of different methods, including FDA, PFFR and the proposed RKHS, under the setting where the observed functional responses are additionally corrupted with measurement error.

The simulation analysis is done under the function-on-function regression setting as in Section 5.2[4]. Specifically, for the ease of comparison, we use exactly the same simulation setting that generates Table 1 in Section 5.2. The only difference is that we assume we do not observe the functional response $Y_t(r_j)$ directly but instead a noisy version of it, denoted by y_{tj}, corrupted with measurement error such that

$$y_{tj} = Y_t(r_j) + \mathfrak{E}_{tj},$$

where $\mathfrak{E}_{tj}$ are independent Gaussian random variable with mean 0 and variance $\sigma^2_{\mathfrak{E}}$ for $t = 1, \ldots, T$ and $j = 1, \ldots, n_2$. Note that for $\sigma^2_{\mathfrak{E}} = 0$, we get back to the identical setting as in Section 5.2. It is easy to derive that the overall variability of the functional noise (i.e. $\mathbb{E} \int_{[0,1]} \epsilon_t(r)^2 \, dr$) under the simulation setting in Section 5.2 is around $0.4^2/12 \approx 0.115^2$, which can be calculated easily using the fact that $\epsilon_t(r) = \sum_{i=1}^q e_{ti} u_i(r)$. We refer to Sections 5.1 and 5.2 for more details.

Thus, in the following, we set the variance of the measurement error as $\sigma^2_{\mathfrak{E}} = 0.12^2$ such that the variability of the functional noise and the measurement error are comparable. For each setting, we conduct 500 experiments. We again use RMISE to evaluate the excess risk of each estimator. Due to limited space, we only report the result for the case where the spectral norm $\kappa = 1$. The results for $\kappa = 0.5$ and $\kappa = 2$ are similar and omitted.

For each method, Table 4 reports its average nRMISE across 500 experiments under all simulation settings. Note that Table 4 is directly comparable with Table 1 in Section 5.2 (for $\kappa = 1$), which reports the performance of each method for $\sigma^2_{\mathfrak{E}} = 0$. Thus, for ease of comparison, we copy the result reported in Table 1 (for $\kappa = 1$) to Table 4 under the name $\sigma^2_{\mathfrak{E}} = 0$.

Examining Table 4, it can be seen clearly that, compared to the case of no measurement error (i.e. $\sigma^2_{\mathfrak{E}} = 0$), the performance of all three estimators (RKHS, FDA and PFFR) only worsen slightly when the variability of the measurement error and functional noise are comparable (i.e. $\sigma^2_{\mathfrak{E}} = 0.12^2$), suggesting that functional noise is more difficult to handle than the independent measurement error. This can also be viewed as numerical support for the result in Theorem 8, i.e. the convergence rate of RKHS is not affected when the functional response is corrupted by independent mean zero measurement errors with a bounded variance.

4. Note that we use the function-on-function regression setting so that we can include FDA in the analysis as FDA cannot handle the general setting of functional regression with mixed covariates.

	Scenario A: $n = 5, T = 50, q = 5$					Scenario B: $n = 5, T = 50, q = 5$				
$\sigma_{\mathfrak{E}}^2$	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)
0	9.14	10.31	10.98	12.81	68	10.42	11.50	11.12	6.63	71
0.12^2	10.14	11.40	12.20	12.45	64	12.61	13.87	13.23	4.96	64
	Scenario A: $n = 20, T = 100, q = 20$					Scenario B: $n = 20, T = 100, q = 20$				
$\sigma_{\mathfrak{E}}^2$	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)
0	5.89	6.75	12.84	14.66	81	18.39	32.41	22.32	21.37	92
0.12^2	5.87	6.86	12.44	16.93	84	22.09	34.67	25.22	14.17	83
	Scenario A: $n = 40, T = 200, q = 50$					Scenario B: $n = 40, T = 200, q = 50$				
$\sigma_{\mathfrak{E}}^2$	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)	RKHS	FDA	PFFR	R_{avg}(%)	R_w (%)
0	3.98	4.41	9.68	10.96	75	20.75	76.74	77.40	269.73	100
0.12^2	4.02	4.47	9.75	11.32	77	23.71	76.95	77.52	224.49	100

Table 4: Numerical performance of RKHS, FDA and PFFR under function-on-function regression with no measurement error ($\sigma_{\mathfrak{E}}^2 = 0$) and with measurement error ($\sigma_{\mathfrak{E}}^2 = 0.12^2$). The reported nRMISE_{avg} is multiplied by 100 in scale. R_{avg} reflects the percent improvement of RKHS over the best performing competitor, and R_w reflects the percentage of experiments in which RKHS achieves the lowest RMISE.

Appendix J. FDA and PFFR with larger number of basis functions

As discussed in Section 5.2, the performance of the penalized basis function approaches, such as FDA and PFFR, may be sensitive to the hyper-parameter N_b, which is the number of basis dimension used in the estimation.

In this section, we further examine the performance of FDA and PFFR with a larger N_b. Specifically, we use the identical simulation setting as that in Section 5.2. We keep the implementation of RKHS the same and the only difference is that we implement FDA with $N_b = 50$ and PFFR with $N_b = 30$ instead of $N_b = 20$. (For PFFR we can only do $N_b = 30$ as the computational cost of PFFR with $N_b > 30$ is practically forbidden for large-scale comparison. See later for more details.)

For each method, Table 5 reports the average nRMISE (nRMISE_{avg}) across 500 experiments under all simulation settings. Table 5 is directly comparable with Table 1 in Section 5.2. Cross-examining Table 5 and Table 1, we have the following observations. For Scenario A, where the bivariate function $A^*(r, s)$ is a simple exponential function, FDA50 and PFFR30 give essentially the same performance as $N_b = 20$. The same applies to Scenario B with $q = 5$ and with $q = 20$ (for low SNR parameter $\kappa = 0.5$). This indicates that $N_b = 20$ is sufficient for simulation settings with low model complexity. On the other hand, for Scenario B with $q = 20$ (for high SNR parameter $\kappa = 1, 2$) and $q = 50$, compared to $N_b = 20$, FDA50 and PFFR30 give much improved performance due to lower approximation bias, though still having a notable performance gap compared to RKHS. We further give the boxplot of RMISE for each method in Figure 13 under $\kappa = 1$. The general pattern is the same as that exhibited in Figure 2 for $N_b = 20$, where RKHS provides the best overall performance.

Computational Time: Note that a larger N_b can significantly increase the computational cost of the penalized basis function approaches. For example, for FDA with the same

	Scenario A: $n = 5, T = 50, q = 5$					Scenario B: $n = 5, T = 50, q = 5$				
κ	RKHS	FDA50	PFFR30	R_{avg}(%)	R_w (%)	RKHS	FDA50	PFFR30	R_{avg}(%)	R_w (%)
0.5	15.98	17.98	19.77	12.41	64	20.86	22.97	22.30	6.88	64
1	9.14	10.32	10.97	12.74	68	10.42	11.51	11.12	6.66	71
2	4.99	6.17	5.69	14.22	72	5.21	5.75	5.55	6.69	74
	Scenario A: $n = 20, T = 100, q = 20$					Scenario B: $n = 20, T = 100, q = 20$				
κ	RKHS	FDA50	PFFR30	R_{avg}(%)	R_w (%)	RKHS	FDA50	PFFR30	R_{avg}(%)	R_w (%)
0.5	10.61	11.94	25.93	12.55	77	37.41	40.65	40.53	8.34	58
1	5.89	6.77	13.32	14.96	81	18.39	20.63	19.75	7.43	73
2	3.60	3.89	6.98	7.99	72	9.19	11.75	9.98	8.52	87
	Scenario A: $n = 40, T = 200, q = 50$					Scenario B: $n = 40, T = 200, q = 50$				
κ	RKHS	FDA50	PFFR30	R_{avg}(%)	R_w (%)	RKHS	FDA50	PFFR30	R_{avg}(%)	R_w (%)
0.5	7.37	8.54	21.00	15.86	82	39.22	48.27	72.42	23.40	94
1	3.98	4.40	10.70	10.75	75	20.75	35.82	70.61	76.05	100
2	2.34	2.34	5.57	0.11	47	12.59	29.77	70.20	151.92	100

Table 5: Numerical performance of RKHS, FDA and PFFR under function-on-function regression. The reported nRMISE_{avg} is multiplied by 100 in scale. R_{avg} reflects the percent improvement of RKHS over the best performing competitor, and R_w reflects the percentage of experiments in which RKHS achieves the lowest RMISE.

number of basis functions N_b for the first and second argument of the bivariate coefficient function $A(r, s)$, the computation of FDA involves inversion of an $N_b^2 \times N_b^2$-dimension matrix (see equations (16.14) and (16.15) in Ramsay and Silverman (2005) for more details), which may not scale well with the number of basis N_b. PFFR uses restricted MLE (REML) for its model estimation by recasting the functional regression model as a penalized additive model with mixed effects (see Ivanescu et al. (2015) for more details), which seems to be slow in terms of computation and also does not scale well with N_b.

For illustration, Figure 14 gives the boxplot of computational time (in log-scale) for RKHS, FDA20, FDA50 with a fixed tuning parameter ($\lambda = 10^{-15}$) and for PFFR20 and PFFR30 across 500 experiments. To conserve space, we present the case for Scenario B with $\kappa = 1$. Results under other settings are similar and thus omitted. Note that for FDA and RKHS, the computational time does not depend on λ as we have closed-form solutions. For PFFR, it uses REML to automatically select the roughness penalty and does not require cross-validation. Thus, for fair comparison, if RKHS and FDA requires a 5-fold cross-validation to search for the best tuning parameters among 50 candidates, we need to add $\log(250) \approx 4.6$ to the log time for FDA and RKHS on Figure 14.

As can be seen from Figure 14, FDA50 incurs much higher computational cost compared to FDA20. Similarly, PFFR30 has noticeably higher computational cost than PFFR20. For the current simulation settings, RKHS has the lowest computational cost.

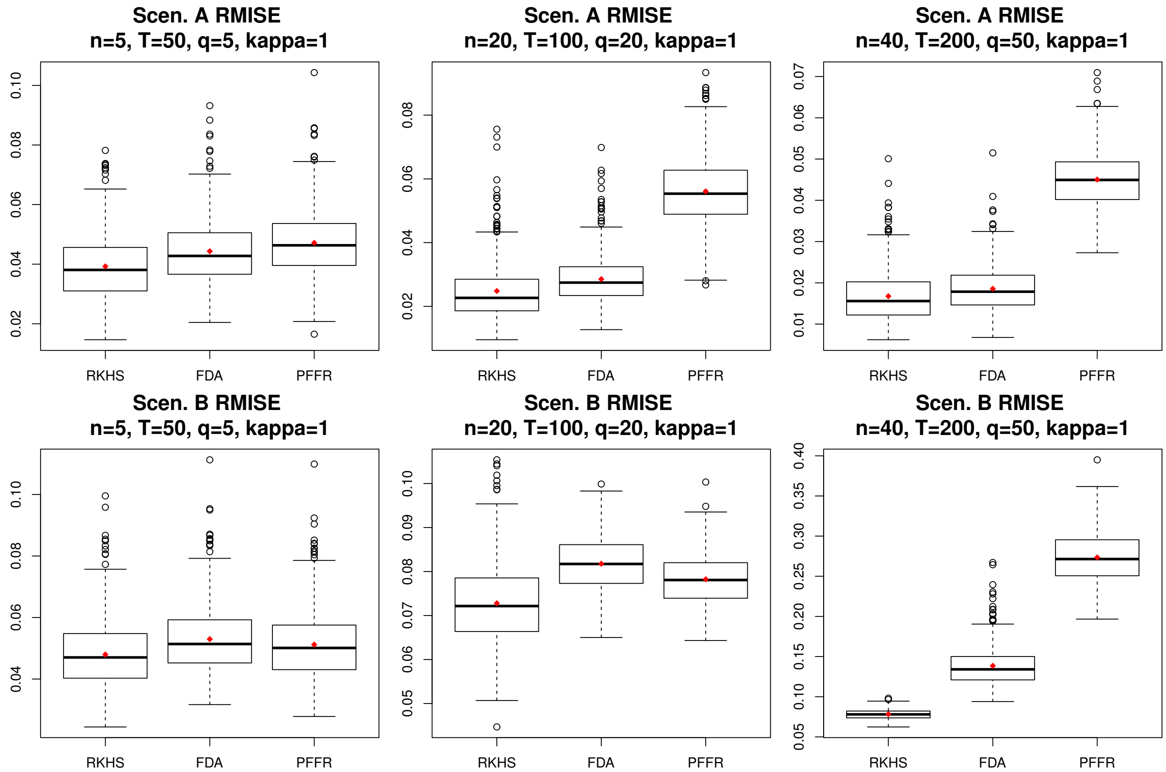

Figure 13: Boxplots of RMISE of RKHS, FDA50 and PFFR30 across 500 experiments under function-on-function regression with $\kappa = 1$. Red points denote the average RMISE.

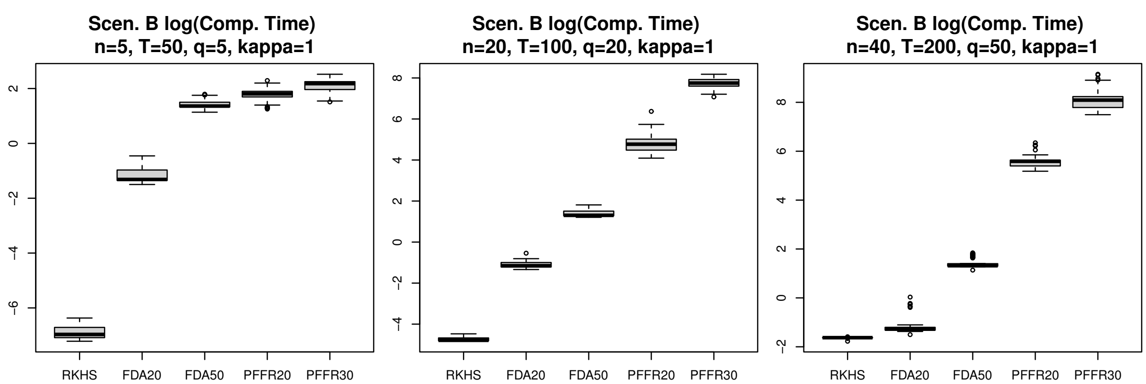

Figure 14: Boxplots of computational time (in log-scale) for RKHS, FDA20, FDA50 with a fixed tuning parameter ($\lambda = 10^{-15}$) and for PFFR20 and PFFR30 across 500 experiments.

Appendix K. Additional numerical comparison with FPCA

In this section, we further provide numerical comparison with the popular functional PCA based regression method (FPCA) proposed in Yao et al. (2005), which estimates the functional linear model based on Karhunen–Loéve expansion. FPCA is implemented via the

Matlab package PACE (FPCreg function). We keep all tuning parameters as default values suggested in the PACE package[5].

Function-on-function regression: We first consider the function-on-function regression setting as in Section 5.2 of the main text. Specifically, we use the identical simulation setting that generates Figure 2 in Section 5.2. The only difference is that besides RKHS, FDA and PFFR, we further implement FPCA in the simulation.

For each setting, we conduct 500 experiments. We again use RMISE to evaluate the excess risk of each estimator. Due to limited space, we only report the result for the case where the spectral norm $\kappa = 1$. The results for $\kappa = 0.5$ and $\kappa = 2$ are similar and omitted. For each method (RKHS, FDA, PFFR and FPCA), Figure 15 visualizes the boxplot of its RMISE across 500 experiments under each simulation setting. Note that the only different between Figure 2 and Figure 15 is that Figure 15 further includes the boxplot of RMISE for FPCA.

As can be seen, for Scenario A, where the coefficient function $A(s,t)$ is the simple exponential function, FPCA provides reasonable performance. However, its performance is not satisfactory when the coefficient function $A(s,t)$ becomes more complex as in Scenario B. This phenomenon is well-known in the literature as the functional space spanned by the eigenfunctions associated with the leading principle components may not be able to well approximate the coefficient function $A(s,t)$, see Cai and Yuan (2012) and Sun et al. (2018) for similar observations.

Functional regression with mixed predictors: We then consider the functional regression with mixed covariates setting as in Section 5.3 of the main text. Specifically, we use the identical simulation setting that generates Figure 3 in Section 5.3. The only difference is that besides RKHS and PFFR, we further implement FPCA in the simulation.

Note that the original FPCreg function in the Matlab package PACE only supports function-on-function regression and does not support the general setting of functional regression with mixed covariates. Thus, we modify the original code in FPCreg to implement FPCA for functional regression with mixed covariates. The modification is straightforward, where we further include the observed scalar predictors (Z_{t1}, Z_{t2}, Z_{t3}) together with the estimated FPCs of the functional predictor X_t to predict the FPCs of the functional response Y_t.

For each setting, we conduct 500 experiments. We again use RMISE to evaluate the excess risk of each estimator. Due to limited space, we only report the result for the case where the spectral norm $\kappa = 1$. The results for $\kappa = 0.5$ and $\kappa = 2$ are similar and omitted. For each method (RKHS, PFFR and FPCA), Figure 16 visualizes the boxplot of its RMISE across 500 experiments under each simulation setting. Note that the only different between Figure 3 and Figure 16 is that Figure 16 further includes the boxplot of RMISE for FPCA.

As can be seen, the phenomenon is essentially the same as the one observed under the function-on-function regression setting. Specifically, for Scenario A, where the coefficient function $A(s,t)$ is the simple exponential function, FPCA provides reasonable performance. However, its performance is not satisfactory when the coefficient function $A(s,t)$ becomes more complex as in Scenario B. Again, this phenomenon is due to the fact that the functional space spanned by the eigenfunctions associated with the leading principle components may

5. See http://www.stat.ucdavis.edu/PACE for more details.

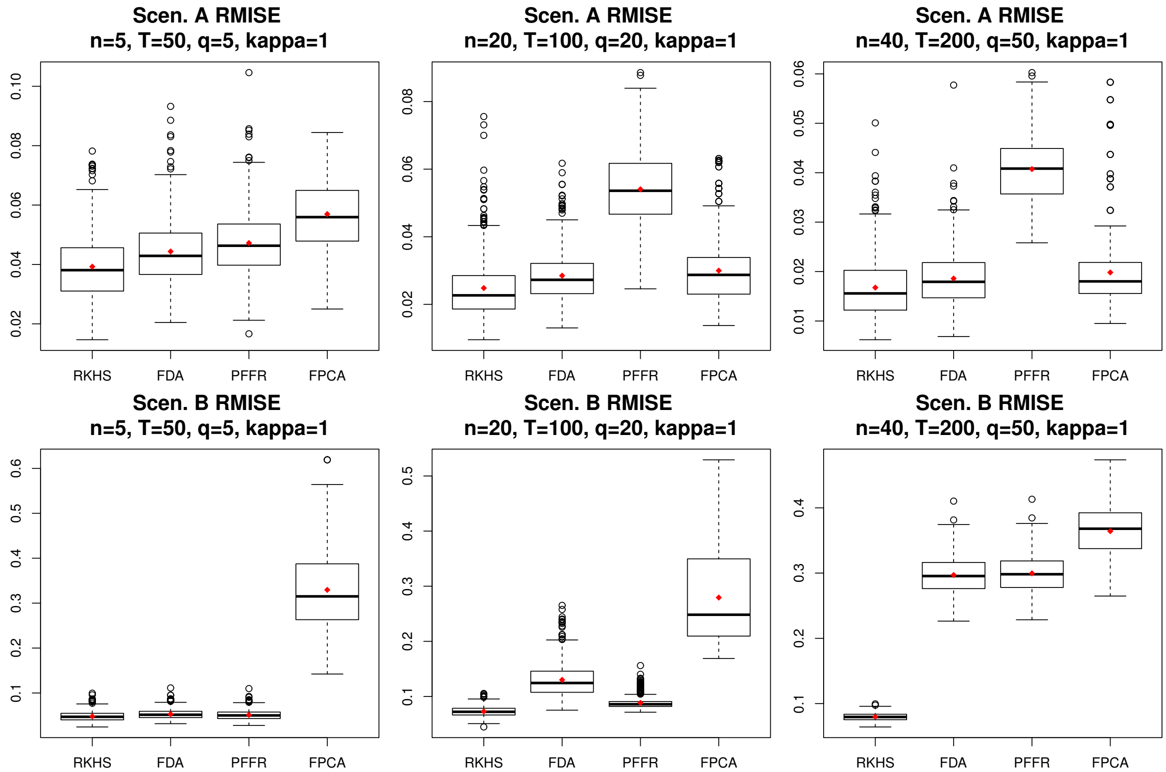

Figure 15: Boxplots of RMISE of RKHS, FDA, PFFR and FPCA across 500 experiments under function-on-function regression with $\kappa = 1$. Red points denote the average RMISE.

not be able to well approximate the coefficient function $A(s,t)$, see Cai and Yuan (2012) and Sun et al. (2018) for similar observations.

Appendix L. Illustration of phase transition

In this section, we provide an indirect numerical illustration for the phase transition phenomenon discovered in Theorem 4. Recall Theorem 4 suggests that putting aside the discretization error ζ_n, there is a phase transition between a nonparametric rate δ_T and a high-dimensional parametric rate $\mathfrak{s}\log(p)/T$.

Here, the nonparametric rate δ_T takes the form $\delta_T = T^{-2r/(2r+1)}$ with some $r > 1/2$ and is associated with the estimation of the bivariate coefficient function A^*. In other words, $\delta_T = T^{-c}$ for some $c > 0$ and c is strictly smaller than 1. The high-dimensional parametric rate $\mathfrak{s}\log(p)/T$ is associated with the estimation of the high-dimensional sparse univariate coefficient functions β^*. In the following, we numerically verify these two rates by considering two special cases of the proposed functional linear model with mixed predictors.

Specifically, we follow the same simulation setting as Scenario B in Section 5.1 of the main text and simulate data from the functional linear regression model

$$Y_t(r) = \int_{[0,1]} A^*(r,s)X_t(s)\,\mathrm{d}s + \sum_{j=1}^{p} \beta_j^*(r)Z_{tj} + \epsilon_t(r),\ \ r \in [0,1]. \tag{87}$$

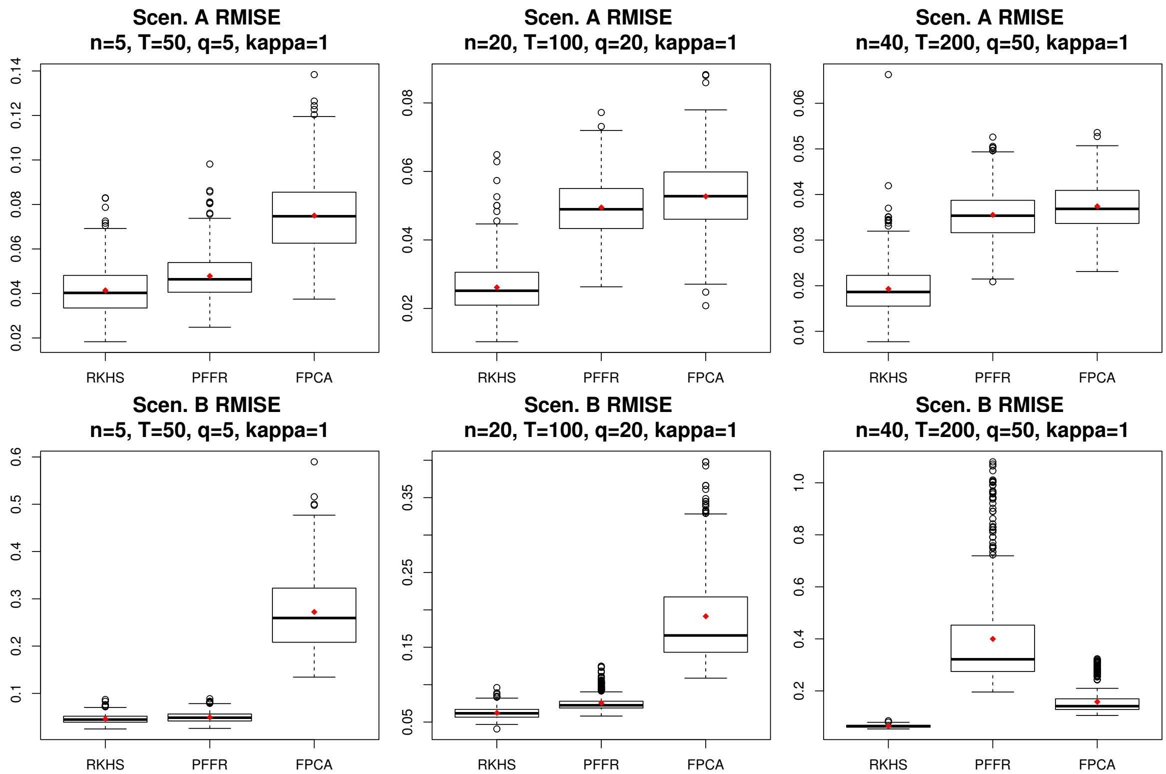

Figure 16: Boxplots of RMISE of RKHS, PFFR and FPCA across 500 experiments under functional regression with mixed predictors with $\kappa = 1$. Red points denote the average RMISE.

The functional predictor X_t, scalar predictor Z_t and the functional error ϵ_t are generated using the same setting in Section 5.1. Following Scenario B, the coefficient functions A^* and $\beta_1^*, \ldots, \beta_p^*$ are generated from a q-dimensional subspace spanned by basis functions $\{u_i(s)\}_{i=1}^q$. We refer to Section 5.1 for more detailed description of Scenario B.

In the following, we fix the dimension $q = 20$ and the SNR parameter $\kappa = 2$. To minimize the impact of the discretization error ζ_n, we set the discrete sample points $\{r_j\}_{j=1}^{n_2}$ for Y_t and $\{s_i\}_{i=1}^{n_1}$ for X_t to be evenly spaced dense grids on $[0, 1]$ with a large number of grids $n = n_1 = n_2 = 50$.

To evaluate the excess risk given by the RKHS estimator, we follow Section 5.1 and use the RMISE defined in (21). To match the result in Theorem 4, we compute MISE, which is the squared RMISE with MISE = RMISE^2.

Nonparametric rate δ_T: We first focus on the nonparametric rate δ_T and consider the special case of (87) by setting $p = 0$, i.e. the function-on-function regression. We vary the sample size $T = 100, 200, 300, \ldots, 2000$ and for each T we conduct 500 experiments. For each sample size T, we compute the average MISE across the 500 experiments and Figure 17(A) gives the plot between log(average MISE) and $\log(T)$, where the relationship is seen to be roughly linear. We further fit an ordinary least squared estimator (OLS) on the points, which gives an estimated slope of -0.811, confirming that δ_T takes the nonparametric rate $O(T^{-c})$ with c being a positive constant less than 1.

High-dimensional parametric rate $\mathfrak{s}\log(p)/T$: We now focus on the high-dimensional parametric rate and consider the special case of (87) by setting $A^* = 0$, i.e. we consider the functional linear regression with only scalar predictors $Z_{t1}, Z_{t2}, \cdots, Z_{tp}$.

We first fix $\mathfrak{s} = 1$ and $T = 200$, and vary the dimension $p = 10, 50, 100, 150, \ldots, 450, 500$. In other words, the coefficient functions $\beta_j^* \equiv 0$ for $j = 2, 3, \ldots, p$ and β_1^* is generated as in Scenario B of Section 5.1 (see more detailed description above). For each dimension p, we conduct 500 experiments and compute the average MISE. Figure 17(B) gives the plot between average MISE and $\log(p)$, where the relationship is seen to be roughly linear. This confirms that fixing the sparsity $\mathfrak{s}$ and sample size T, the excess risk increases in the order of $O\big(\log(p)\big)$.

We next fix $\mathfrak{s} = 1$, $p = 10$, and vary the sample size $T = 100, 200, 300, \ldots, 2000$. For each sample size T, we conduct 500 experiments and compute the average MISE. Figure 17(C) gives the plot between log(average MISE) and $\log(T)$, where the relationship is seen to be roughly linear. We further fit an OLS on the points, which gives an estimated slope of -1.012, confirming that fixing the sparsity $\mathfrak{s}$ and dimension p, the excess risk decreases in the order of roughly $O(1/T)$.

Finally, we fix $T = 200$, $p = 100$, and vary the sparsity $\mathfrak{s} = 1, 2, 5, 10, 15, 20$. In other words, the coefficient functions $\beta_j^* \equiv 0$ for $j = \mathfrak{s}+1, \mathfrak{s}+2, \ldots, p$ and $\beta_1^* = \beta_2^* = \cdots = \beta_{\mathfrak{s}}^*$ are generated as in Scenario B of Section 5.1. For each sparsity $\mathfrak{s}$, we conduct 500 experiments and compute the average MISE. Figure 17(D) gives the plot between log(average MISE) and $\log(\mathfrak{s})$, where the relationship is seen to be roughly linear. We further fit an OLS on the points, which gives an estimated slope of 1.087, confirming that fixing the sample size T and dimension p, the excess risk increases in the order of roughly $O(\mathfrak{s})$.

Thus, we numerically illustrate that the prediction error associated with the estimation of the high-dimensional sparse univariate coefficient functions β^* is of order $O\big(\mathfrak{s}\log(p)/T\big)$. Putting these two cases together, we can conclude that there is a phase transition phenomenon between a nonparametric rate $O(T^{-c})$ and a high-dimensional parametric rate $O\big(\mathfrak{s}\log(p)/T\big)$.

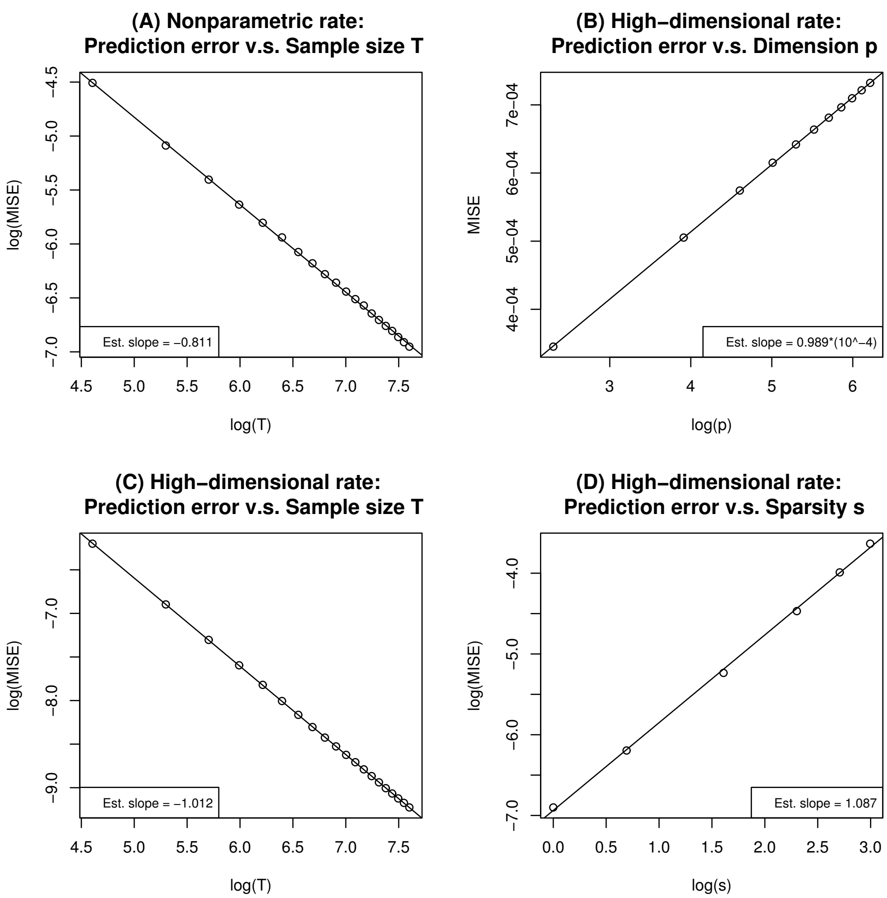

Figure 17: Numerical illustration of the nonparametric rate (A) and the high-dimensional parametric rate (B-D).