Journal of Machine Learning Research 24 (2023) 1-38 Submitted 12/20; Revised 10/22; Published 3/23

# Kernel-based estimation for partially functional linear model: Minimax rates and randomized sketches

**Shaogao Lv** LVSG716@NAU.EDU.CN
*College of Statistics and Data Science*
*Nanjing Audit University*
*Nanjing, China*

**Xin He** HE.XIN17@MAIL.SHUFE.EDU.CN
*School of Statistics and Management*
*Shanghai University of Finance and Economics*
*Shanghai, China*

**Junhui Wang** JUNHUIWANG@CUHK.EDU.HK
*Department of Statistics*
*The Chinese University of Hong Kong*
*Shatin, New Territory, Hong Kong*

**Editor:** Massimiliano Pontil

## Abstract

This paper considers the partially functional linear model (PFLM) where all predictive features consist of a functional covariate and a high dimensional scalar vector. Over an infinite dimensional reproducing kernel Hilbert space, the proposed estimation for PFLM is a least square approach with two mixed regularizations of a function-norm and an $\ell_1$-norm. Our main task in this paper is to establish the minimax rates for PFLM under high dimensional setting, and the optimal minimax rates of estimation are established by using various techniques in empirical process theory for analyzing kernel classes. In addition, we propose an efficient numerical algorithm based on randomized sketches of the kernel matrix. Several numerical experiments are implemented to support our method and optimization strategy.

**Keywords:** Functional linear models, minimax rates, sparsity, randomized sketches, reproducing kernel Hilbert space.

## 1. Introduction

In the problem of functional linear regression, a single functional feature $X(\cdot)$ is assumed to be square-integrable over an interval $\mathcal{T}$, and the classical functional linear regression between the response $Y$ and $X$ is given as

$$Y = \langle X, f^* \rangle_{\mathcal{L}_2} + \varepsilon, \tag{1.1}$$

where the inner product $\langle \cdot, \cdot \rangle_{\mathcal{L}_2}$ is defined as $\langle f, g \rangle_{\mathcal{L}_2} := \int_{\mathcal{T}} f(t)g(t)dt$ for any $f, g \in \mathcal{L}_2(\mathcal{T})$. Here $f^*$ is some slope function within $\mathcal{L}_2(\mathcal{T})$ and $\varepsilon$ denotes an error term with zero-mean. Let $(Y_i, X_i) : i = 1, ..., n$ denote independent and identically distributed (i.i.d.) realizations

[*] Xin He is the corresponding author.

from the population $(Y, X)$, there is extensive literature on estimation of the slope function $f^*$, or the value of $\langle X, f^* \rangle_{\mathcal{L}_2}$.

In practice, it is often the case that a response is affected by both a high-dimensional scalar vector and some random functional variables as predictive features. These scenarios partially motivate us to study PFLM under high dimensional setting. For simplifying the notations, this paper assumes that $Y$ and $X(\cdot)$ are centered. To be more precise, we are concerned with partially functional linear regression with the functional feature $X$ and scalar predictors $\mathbf{Z} = (Z_1, ..., Z_p)^T \in \mathcal{R}^p$, and a linear model links the response $Y$ and predictive features $\mathbf{U} = (X, \mathbf{Z})$ that

$$Y = \langle X, f^* \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma}^* + \varepsilon, \tag{1.2}$$

where $\boldsymbol{\gamma}^* = (\gamma_1^*, ..., \gamma_p^*)^T$ denotes the regression coefficients of the scalar covariates, and $\varepsilon$ is a standard normal variable and independent of $X$ and $\mathbf{Z}$. It is interesting to note that $X$ and $\mathbf{Z}$ are not required to be independent here. Under the sparse high dimensional setting, a standard assumption is that the cardinality of the active set $S_0 := \{j : \gamma_j^* \neq 0,\ j = 1, ..., p\}$ is far less than $p$, while $p$ and $p_0 := |S_0|$ are allowed to diverge as the sample size $n$ increases. It is interesting to point out that PFLM (1.2) is particularly attractive to analyze data consisting of both functional features and many scalar predictors, which commonly appear in many real-world problems as pointed out by Kong et al. 2016. In fact, estimation and variable selection issues for partially functional linear models have been investigated via FPCA methods by Shin and Lee (2012); Lu et al. (2014) and Kong et al. (2016), respectively.

In this paper, we focus on a least square regularized estimation for the slope function and the regression coefficients in (1.2) under a kernel-based framework and high dimension setting. The estimators obtained are based on a combination of the least-squared loss with a $\ell_1$-type penalty and the square of a functional norm, where the former penalty corresponds to the regression coefficients and the latter one is used to control the kernel complexity. The optimal minimax rates of estimation are established by using various techniques in empirical process theory for analyzing kernel classes, and an efficient numerical algorithm based on randomized sketches of the kernel matrix is implemented to verify our theoretical findings.

### 1.1 Our Contributions

This paper makes three main contributions to this functional modeling literature.

Our first contribution is to establish Theorem 1 stating that with high probability, under mild regularity conditions, the prediction error of our procedure under the squared $L_2$-norm is bounded by $\left(\frac{p_0 \log p}{n} + n^{-\frac{2r}{2r+1}}\right)$, where the quantity $r > 1/2$ corresponds to the kernel complexity of one composition kernel $K^{1/2} C K^{1/2}$ with $K$ denoting some bounded kernel and $C(s,t) = \mathbb{E}[X(s)X(t)]$ for any $t,\ s \in \mathcal{T}$. Note that the boundedness of $K$ is required to apply the spectral decomposition and Bousquet inequality in our theoretical analysis. The proof of this upper bound involves two different penalties for analyzing the obtained estimator in high dimensions, and we want to emphasize that it is very hard to prove constraint cone set that has often been used to define a critical condition (constraint eigenvalues constant) for high-dimensional problems (Bickel, Ritov, and Tsybakov, 2009; Verzelen, 2012). To handle this technical difficulty, we combine the methods used in Müller and Van de Geer (2015) for high dimensional partial linear models with various techniques in empirical process theory for analyzing kernel classes (Aronszajn, 1950; Steinwart and Christmann, 2008; Cai and Yuan, 2012; Yuan and Cai, 2010; Zhu et al., 2014).

Our second contribution is to establish algorithm-independent minimax lower bounds under the squared $L_2$ norm. These minimax lower bounds, stated in Theorem 3, are determined in terms of the metric entropy of the composition kernel $K^{1/2}CK^{1/2}$ and the sparsity structure of high dimensional scalar coefficients. For the commonly-used kernels, including the Sobolev classes, these lower bounds match our achievable results, showing optimality of our estimator for PFLM. It is worthy noting that, the lower bound of parametric part does not depend on nonparametric smoothness indices, coinciding with the classical sparse estimation rate in the high dimensional linear models (Verzelen, 2012). By contrast, the lower bound for estimating $f^*$ turns out to be affected by the regression coefficient $\boldsymbol{\gamma}^*$. The proof of Theorem 3 is based on characterizing the packing entropies of the class of nonparametric kernel models, interaction between the composition kernel and high dimensional scalar vector, combined with classical information theoretic techniques involving Fano's inequality and variants (Yang and Barron, 1999; Van. de. Geer, 2000; Tsybakov, 2009).

Our third contribution is to consider randomized sketches for our original estimation with statistical dimension. Despite these attractive statistical properties stated as above, the computational complexity of computing our original estimate prevents it from being routinely used in large-scale problems. In fact, a standard implementation for any kernel estimator leads to the time complexity $O(n^3)$ and space complexity $O(n^2)$ respectively. To this end, we employ the random projection and sketching techniques developed in Yang et al. (2017); Mahoney (2011), where it is proposed to approximate $n$-dimensional kernel matrix by projecting its row and column subspaces to a randomly chosen m-dimensional subspace with $m \ll n$. We give the sketch dimension $m$ proportional to the statistical dimension, under which the resulting estimator has a comparable numerical performance.

### 1.2 Related Work

A class of conventional estimation procedures for functional linear regressions in the statistical literature are based on functional principal components regression (FPCA) or spline functions; see (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Kong, Xue, Yao, and Zhang, 2016) and (Cardot, Ferraty, and Sarda, 2003) for details. These truncation approaches to handle an infinity-dimensional function only depend on the information of the feature $X$. In particular, commonly-used FPCA methods that form an available basis for the slope function $f^*$ are determined solely by empirical covariance of the observed feature $X$, and these basis may not act as an efficient representation to approximate $f^*$, since the slope function $f^*$ and the leading functional components are essentially unrelated. Similar problems also arise when spline-based finite representation are used.

To avoid inappropriate representation for the slope function, reproducing kernel methods have been known to be a family of powerful tools for directly estimating infinity-dimensional functions, and the optimal rates for the regularized least-squares estimation has been provided in Caponnetto and De Vito (2007). When the slope function is assumed to reside in a reproducing kernel Hilbert space (RKHS), denoted by $\mathcal{H}_K$, several existing work (Yuan and Cai, 2010; Cai and Yuan, 2012; Zhu, Yao, and Zhang, 2014) for functional linear or additive regression have proved that the minimax rate of convergence depends on both the kernel

$K$ and the covariance function $C$ of the functional feature $X$. In particular, the alignment of $K$ and $C$ can significantly affect the optimal rate of convergence. However, it is well known that kernel-based methods suffer a lot from storage cost and computational burden. Specially, kernel-based methods need to store a $n \times n$ matrix before running algorithms and are limited to small-scale problems.

### 1.3 Paper Organization

The rest of this paper is organized as follows. Section 2 introduces some notations and the basic knowledge on kernel method, and formulates the proposed kernel-based regularized estimation method. Section 3 is devoted to establish the minimax rate of the prediction problem for PFLM and provide detailed discussion on the obtained results, including the desired convergence rate of the upper bounds and a matching set of minimax lower bounds. In Section 4, a general sketching-based strategy is provided, and an approximate algorithm for solving (2.2) is employed. Several numerical experiments are implemented in Section 5 to support the proposed approach and the employed optimization strategy. A brief summary of this paper is provided in Section 6. Appendix A and B contain several core proof procedures of the main results, including the technical proofs of Theorems 1–3. Some useful lemmas and more technical details are provided in Appendix C.

## 2. Problem Statement and Proposed Method

### 2.1 Notation

Let $u, v$ be two general random variables, and denote the joint distribution of $(u, v)$ by $Q$ and the marginal distribution of $u$(**resp.** $v$) by $Q_u$(**resp.** $Q_v$). For a measurable function $f : u \times v \to \mathcal{R}$, we define the squared $L_2$-norm by $\|f\|^2 := \mathbb{E}_Q f^2(u, v)$, and the squared empirical norm is given by $\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f^2(u_i, v_i)$, where $\{(u_i, v_i)\}_{i=1}^n$ are i.i.d. copies of $(u, v)$. Note that $Q$ may differ from line to line. For a vector $\boldsymbol{\gamma} \in \mathcal{R}^p$, the $\ell_1$-norm and $\ell_2$-norm are given by $\|\boldsymbol{\gamma}\|_1 := \sum_{j=1}^p |\gamma_j|$ and $\|\boldsymbol{\gamma}\|_2 := \left(\sum_{j=1}^p \gamma_j^2\right)^{1/2}$, respectively. With a slight abuse of notation, we write $\|f\|_{\mathcal{L}_2}^2 := \langle f, f\rangle_{\mathcal{L}_2}$ with $\langle f, g\rangle_{\mathcal{L}_2} = \int_{\mathcal{T}} f(t)g(t)dt$. For two sequences $\{a_k : k \geq 1\}$ and $\{b_k : k \geq 1\}$, $a_k \lesssim b_k$ (or $a_k = O(b_k)$) means that there exists some constant $c$ such that $a_k \leq c b_k$ for all $k \geq 1$. Also, we write $a_k \gtrsim b_k$ if there is some positive constant $c$ such that $a_k \geq c b_k$ for all $k \geq 1$. Accordingly, we write $a_k \asymp b_k$ if both $a_k \lesssim b_k$ and $a_k \gtrsim b_k$ are satisfied.

### 2.2 Kernel Method

Kernel methods are one of the most powerful learning schemes in machine learning, which often take the form of regularization schemes in a reproducing kernel Hilbert space (RKHS) associated with a Mercer kernel (Aronszajn, 1950). A major advantage of employing the kernel methods is that the corresponding optimization task over an infinite dimensional RKHS are equivalent to a $n$-dimensional optimization problems, benefiting from the so-called reproducing property, and thus kernel method has become a time-proven popular mainstay in the literature of machine learning (Steinwart and Christmann, 2008; Cai and Yuan, 2012; Yang et al., 2017; Marteau-Ferey et al., 2019; Della Vecchia et al., 2021).

Recall that a kernel $K(\cdot,\cdot):\mathcal{T}\times\mathcal{T}\to\mathcal{R}$ is a continuous, symmetric, and positive semi-definite function. Let $\mathcal{H}_K$ be the closure of the linear span of functions $\{K_t(\cdot):=K(t,\cdot), t\in\mathcal{T}\}$ endowed with the inner product $\langle\sum_{i=1}^n\alpha_iK_{t_i},\sum_{j=1}^n\beta_jK_{t_j}\rangle_K:=\sum_{i,j=1}^n\alpha_i\beta_jK(t_i,t_j)$, for any $\{t_i\}_{i=1}^n,\{t_i\}_{i=1}^n\in\mathcal{T}^n$ and $n\in\mathcal{N}^+$. An important property on $\mathcal{H}_K$ is the reproducing property stating that

$$f(t)=\langle f,K_t\rangle_K,\ \text{ for any } f\in\mathcal{H}_K.$$

This property ensures that an RKHS inherits many nice properties from the standard finite dimensional Euclidean spaces. Throughout this paper, we assume that the slope function $f^*$ resides in a specified RKHS, still denoted by $\mathcal{H}_K$. In addition, another RKHS can be naturally induced by the stochastic process of $X(\cdot)$. Without loss of generality, we assume that $X(\cdot)$ is square integrable over $\mathcal{T}$ with zero-mean, and thus the covariance function of $X$, defining as

$$C(s,t)=\mathbb{E}[X(s)X(t)],\quad\forall\, t,\, s\in\mathcal{T},$$

is also a real and semi-definite kernel.

Note that the kernel complexity is characterized explicitly by a kernel-induced integral operator. Precisely, for any bounded kernel $K(\cdot,\cdot):\mathcal{T}\times\mathcal{T}\to\mathcal{R}$, we define the integral operator $L_K:\mathcal{L}_2(\mathcal{T})\to\mathcal{L}_2(\mathcal{T})$ by

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

By the reproducing property, $L_K$ can be equivalently defined as

$$\langle f,L_K(g)\rangle_K=\langle f,g\rangle_{\mathcal{L}_2},\quad\forall\, f\in\mathcal{H}_K,\, g\in\mathcal{L}_2(\mathcal{T}).$$

Since the operator $L_K$ is linear, bounded and self-adjoint in $\mathcal{L}_2(\mathcal{T})$, the spectral theorem implies that there exist a family of orthonormalized eigenfunctions $\{\phi_\ell^K:\ell\geq1\}$ and a sequence of eigenvalues $\theta_1^K\geq\theta_2^K\geq...>0$ such that

$$K(s,t)=\sum_{\ell\geq1}\theta_\ell^K\phi_\ell^K(s)\phi_\ell^K(t),\quad\forall s,\, t\in\mathcal{T},$$

and thus by definition, it holds

$$L_K(\phi_\ell^K)=\theta_\ell^K\phi_\ell^K,\quad\ell=1,2,...$$

Based on the semi-definiteness of $L_K$, we can always decompose it into the following form

$$L_K=L_{K^{1/2}}\circ L_{K^{1/2}},$$

where $L_{K^{1/2}}$ is also a kernel-induced integral operator associated with a fractional kernel $K^{1/2}$ that

$$K^{1/2}(s,t):=\sum_{\ell\geq1}\sqrt{\theta_\ell^K}\phi_\ell^K(s)\phi_\ell^K(t),\quad s,\, t\in\mathcal{T}.$$

Also, it holds

$$L_{K^{1/2}}(\phi_\ell^K):=\sqrt{\theta_\ell^K}\phi_\ell^K.$$

Given two kernels $K_1, K_2$, we define

$$(K_1K_2)(s,t) := \int_{\mathcal{T}} K_1(s,u)K_2(t,u)du,$$

and then it holds $L_{K_1K_2} = L_{K_1} \circ L_{K_2}$. Note that $K_1K_2$ is not necessarily a symmetric kernel.

In the rest of this paper, we focus on the RKHS $\mathcal{H}_K$ in which the slope function $f^*$ in (1.2) resides. Given the kernel $K$, the covariance function $C$ and by using the above notation, we define the linear operator $L_{K^{1/2}C_kK^{1/2}}$ by

$$L_{K^{1/2}CK^{1/2}} := L_{K^{1/2}} \circ L_C \circ L_{K^{1/2}}.$$

If the both operators $L_{K^{1/2}}$ and $L_C$ are linear, bounded and self-adjoint, so is $L_{K^{1/2}CK^{1/2}}$. By the spectral theorem, there exist a sequence of positive eigenvalues $s_1 \geq s_2 \geq ... > 0$ and a set of orthonormalied eigenfunctions $\{\varphi_\ell : \ell \geq 1\}$ such that

$$K^{1/2}CK^{1/2}(s,t) = \sum_{\ell \geq 1} s_\ell \varphi_\ell(s)\varphi_\ell(t), \quad \forall\, s,t \in \mathcal{T},$$

and particularly

$$L_{K^{1/2}CK^{1/2}}(\varphi_\ell) = s_\ell \varphi_\ell, \quad \ell = 1, 2, ...$$

It is worthwhile to note that the eigenvalues $\{s_\ell : \ell \geq 1\}$ of the linear operator $L_{K^{1/2}CK^{1/2}}$ depend on the eigenvalues of both the reproducing kernel $K$ and the covariance function $C$. To be more precise, it is easy to verify that $s_\ell = \theta_\ell^K \theta_\ell^C$ under the case that $\phi_\ell^K = \phi_\ell^C$. Yet, in general, only the eigenvalues of $K$ and $C$ cannot determine the decay rate of the eigenvalues of $L_{K^{1/2}CK^{1/2}}$, which heavily relies on the alignments of the eigenfunctions of $K$ and $C$. We shall show in Section 3 that the minimax rate of convergence of the excess prediction risk is determined by the decay rate of the eigenvalues $\{s_\ell : \ell \geq 1\}$.

### 2.3 Regularized Estimation and Randomized Sketches

Given the sample $\{Y_i, (X_i, \mathbf{Z}_i)\}_{i=1}^n$ which are independently drawn from (1.2), the proposed estimation procedure for PFLM (1.2) is formulated in a least square regularization scheme by solving

$$(\widehat{f}, \widehat{\gamma}) = \underset{f \in \mathcal{H}_K, \gamma \in \mathcal{R}^p}{\operatorname{argmin}} \Big\{ \frac{1}{n} \sum_{i=1}^n \big( Y_i - \langle X_i, f \rangle_{\mathcal{L}_2} - \mathbf{Z}_i^T \gamma \big)^2 + \mu^2 \|f\|_K^2 + \lambda \|\gamma\|_1 \Big\}, \tag{2.1}$$

where the parameter $\mu^2 > 0$ is used to control the smoothness of the nonparametric component and $\lambda > 0$ associated with the $\ell_1$-type penalty is used to generate sparsity with respect to the scalar covariates.

Note that although the proposed estimation procedure (2.1) is formulated within an infinite-dimensional Hilbert space, the following lemma shows that this optimization task is equivalent to a finite-dimensional minimization problem.

**Lemma 1** *The proposed estimation procedure* (2.1) *defined on* $\mathcal{H}_K \times \mathcal{R}^p$ *is equivalent to a finite-dimensional parametric convex optimization. That is,* $\widehat{f}(t) = \sum_{k=1}^n \alpha_k B_k(t)$ *with unknown coefficients* $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^T$*, for any* $t \in \mathcal{T}$*. Here each basis function* $B_k(t) = \langle X_k, K(t, ) \rangle_{\mathcal{L}_2(\mathcal{T})} \in \mathcal{H}_K$*,* $k = 1, ..., n$*.*

To rewrite the minimization problem (2.1) into a matrix form, we define a $n \times n$ semi-definite matrix $\mathbb{K}^c = (K^c_{ik})^n_{i,k=1}$ with $K^c_{ik} := \langle X_i, B_k \rangle_{\mathcal{L}_2} = \iint X_k(u)X_i(t)K(t,u)dudt$, and by the reproducing property on $K$, we also get $\langle B_i, B_k \rangle_K = K^c_{ik}$, $i,k = 1, ..., n$. Thus, by Lemma 1, the matrix form of (2.1) is given as

$$\min_{\boldsymbol{\alpha} \in \mathcal{R}^n, \boldsymbol{\gamma} \in \mathcal{R}^p} \frac{1}{n} \big\| \mathbf{y} - \mathbb{K}^c \boldsymbol{\alpha} - \mathbb{Z}\boldsymbol{\gamma} \big\|_2^2 + \mu^2 \boldsymbol{\alpha}^T \mathbb{K}^c \boldsymbol{\alpha} + \lambda \|\boldsymbol{\gamma}\|_1, \tag{2.2}$$

where $\mathbb{Z} \in \mathcal{R}^{n \times p}$ denotes the design matrix of $\mathbf{Z}$. Since the unconstrained problem (2.2) is convex for both $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, the standard alternative optimization (Boyd and Vandenberghe, 2004) can be applied directly to approximate a global minimizer of (2.2). Yet, due to the fact that $\mathbb{K}^c$ is a $n \times n$ matrix, both computation cost and storage burden are very heavy in standard implementation, with the orders $O(n^3)$ and $O(n^2)$, respectively. To alleviate the computational issue, we propose an approximate numerical optimization instead of (2.2) in Section 4. Precisely, a class of general random projections are adopted to compress the original kernel matrix $\mathbb{K}^c$ and improve the computational efficiency.

## 3. Main Results: Minimax Rates

In this section, we present the main theoretical results of the proposed estimation in the minimax sense. Specifically, we derive the minimax rates in terms of prediction error for the estimators in (2.1) under high dimension and kernel-based frameworks. The first two theorems prove the convergence of the obtained estimators, while the last one provides an algorithmic-independent lower bound for the prediction error.

### 3.1 Upper Bounds

We denote the short-hand notation

$$\mathcal{G} := \big\{ g_{f,\gamma}(X, \mathbf{Z}) = \langle X, f \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma},\ f \in \mathcal{H}_K,\ \boldsymbol{\gamma} \in \mathcal{R}^p \big\},$$

and the functional $g^*(\mathbf{U}) := \langle X, f^* \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma}^*$ for $\mathbf{U} = (X, \mathbf{Z})$. With a slight confusion of notation, we sometimes also write $\mathcal{G} := \big\{ g = (f, \boldsymbol{\gamma}),\ f \in \mathcal{H}_K,\ \boldsymbol{\gamma} \in \mathcal{R}^p \big\}$. To split the scalar components and the functional component involved in our analysis, we define the projection of a random variable $U$ concerning $\mathcal{H}_K$ as $\Pi(U|\mathcal{H}_K) = \text{argmin}_{f \in \mathcal{H}_K} \|U - \langle X, f \rangle_{\mathcal{L}_2}\|^2$, where $\|\cdot\|^2$ is defined as $\|U\|^2 := \mathbb{E}[U^2]$ for a random variable $U$. For each component of $\mathbf{Z} = (Z_1, ..., Z_p)^T$, let $\Pi(Z_j|X) = \langle X, \Pi(Z_j|\mathcal{H}_K) \rangle_{\mathcal{L}_2}$ and $\Pi_{\mathbf{Z}|X} = (\Pi(Z_1|X), ..., \Pi(Z_p|X))^T$, and then we denote $\widetilde{\mathbf{Z}} := \mathbf{Z} - \Pi_{\mathbf{Z}|X}$ as a random vector of $\mathcal{R}^p$. For any $g_1(\mathbf{U}) := \langle X, f_1 \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma}_1 \in \mathcal{G}$ and $g_2(\mathbf{U}) := \langle X, f_2 \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma}_2 \in \mathcal{G}$, we have the following orthogonal decomposition that

$$\begin{aligned} g_1(\mathbf{U}) - g_2(\mathbf{U}) &= \mathbf{Z}^T(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2) + \langle X, f_2 - f_1 \rangle_{\mathcal{L}_2} \\ &= \widetilde{\mathbf{Z}}^T(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2) + \Pi^T_{\mathbf{Z}^T|X}(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2) + \langle X, f_2 - f_1 \rangle_{\mathcal{L}_2}, \end{aligned}$$

and by the definition of projection, it holds

$$\|g_1 - g_2\|^2 = \|\widetilde{\mathbf{Z}}^T(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2)\|^2 + \|\Pi^T_{\mathbf{Z}|X}(\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2) + \langle X, f_2 - f_1 \rangle_{\mathcal{L}_2}\|^2. \tag{3.1}$$

To establish the refined upper bounds of the prediction and estimation errors, we summarize and discuss the main conditions needed in the theoretical analysis below.
**Condition A**(Eigenvalues condition). The smallest eigenvalue $\Lambda_{min}^2$ of $\mathbb{E}[\widetilde{\mathbf{Z}}\widetilde{\mathbf{Z}}^T]$ is positive, and the largest eigenvalue $\Lambda_{max}^2$ of $\mathbb{E}[\Pi_{\mathbf{Z}|X}\Pi_{\mathbf{Z}|X}^T]$ is finite.
**Condition B**(Design condition). For some positive constants $C_z, C_\pi, C_h$, there holds:

$$|Z_j| \leq C_z, \ \|\Pi(Z_j|X)\|_\infty \leq C_\pi, \text{ and } \|\Pi(Z_j|\mathcal{H}_K)\|_K \leq C_h, \quad \text{for any } j = 1, ..., p.$$

**Condition C**(Light tail condition). There exist two constants $c_1$, $c_2$ such that

$$\mathbb{P}\{\|L_{K^{1/2}}X\|_{\mathcal{L}_2} \geq t\} \leq c_1 \exp(-c_2 t^2), \quad \text{for any } t > 0.$$

**Condition D**(Entropy condition). For some constant $1/2 < r < \infty$, the sequence of eigenvalues $s_\ell$ satisfy that

$$s_\ell = O(\ell^{-2r}), \quad \ell \in \mathcal{N}^+.$$

Condition A is commonly used in literature of semiparametric modelling ; see (Müller and Van de Geer, 2015) for reference. This condition ensures that there is enough information in the data to identify the parameters in the scalar part. Condition B imposes some boundedness assumptions, which are not essential and are used only for simplifying the technical proofs. Note that for the unbounded case, we may need construct a truncation way or assume some exponential-tail decay conditions for theses quantities to apply the empirical processes theory. The readers are referred to Theorems 3 and 4 and the corresponding discussions of Cai and Yuan (2012) for reference. Condition C implies that the random process $L_{K^{1/2}}X$ has an exponential decay rate and the same condition is also considered in Cai and Yuan (2012). Particularly, it is naturally satisfied if $X$ is a Gaussian process. In Condition D, the parameters $s_\ell$ are related to the alignment between $K$ and $C$, which plays an important role in determining the minimax optimal rates. Moreover, the decay of $s_\ell$ characterizes the kernel complexity and has close relation with various covering numbers and Radmeacher complexity. Specially, the polynomial decay assumed in Condition D is satisfied for the classical Sobolev class and Besov class.

The following theorem states that with an appropriately chosen $(\mu, \lambda)$, the predictor $\widehat{g} := \langle X, \widehat{f}\rangle_{\mathcal{L}_2} + \mathbf{Z}^T\widehat{\gamma}$ attains a sharp convergence rate under $L_2$-norm.

**Theorem 1** *Suppose that Conditions A-D hold. With the choice of the tuning parameters $(\mu, \lambda)$, such that*

$$\mu \asymp n^{-\frac{r}{2r+1}} + \sqrt{\log(2p)/n}, \ \lambda \asymp \sqrt{\log(2p)/n}.$$

*Then with probability at least $1 - 2\exp[-n(\delta_1'')^2\mu^2]$, the proposed estimation for PFLM satisfies*

$$\|\widehat{g} - g^*\|^2 \lesssim \Big(n^{-\frac{2r}{2r+1}} + \frac{p_0 \log(2p)}{n}\Big),$$

*where $\delta_1^{''}$ is some appropriately small quantity.*

The proof of Theorem 1 will be given in the first part of Appendix A before Appendix $A_2$. Note that the explicit definition of $\delta_1^{''}$ is provided in Lemma 4 and may depend on $n$. Theorem 1 shows that the proposed estimation (2.1) achieves a fast convergence rate in the term of prediction error. Note that the derived rate depends on the kernel complexity of $K^{1/2}CK^{1/2}$ and the sparsity of scalar components. It is interesting to note that even there exists some underlying correlation structure between the functional feature and the scalar covariates, the choice of hyper-parameter $\mu$ depends on the structural information of all the features, while the sparsity hyper-parameter $\lambda$ only depends on the scalar component.

**Theorem 2** *Suppose that all the conditions in Theorem 1 are satisfied. Then with probability at least* $1-4\exp[-n(\delta_1'')^2\mu^2]-\frac{5}{2p}$*, there holds*

$$\|\widetilde{\mathbf{Z}}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2+\frac{\lambda}{8}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1\lesssim\Big(\frac{p_0}{\Lambda_{min}^2}\frac{\log(2p)}{n}\Big),\tag{3.2}$$

*and in addition, we have*

$$\|\langle X,\widehat{g}-g^*\rangle_{\mathcal{L}_2}\|^2\lesssim\Big(n^{-\frac{2r}{2r+1}}+\frac{p_0\log(2p)}{n}\Big).\tag{3.3}$$

The proof of Theorem 2 will be given in Appendix $A_2$ below. Note that the exponential-type dependence in $n$ or $s$ is characterized in $\delta_1''$, and the term $\frac{5}{2p}$ results from the high dimensional scalar vector. Theorem 2 shows that the parameter estimator and the functional estimator can achieve the fast convergence rate. Specifically, the estimation error of the parametric estimator $\widehat{\boldsymbol{\gamma}}$ can achieve the optimal convergence rate in the high dimensional linear models (Verzelen, 2012), even in the presence of nonparametric components. This result in the functional literature is similar in spirit to the classical high dimensional partial linear models (Müller and Van de Geer, 2015; Yu, Levine, and Cheng, 2019).

### 3.2 Lower Bounds

In this part, we establish the lower bounds on the minimax risk of estimating $\boldsymbol{\gamma}^*$ and $\langle X,f^*\rangle_{\mathcal{L}_2}$ separately. Let $B[p_0,p]$ be a set of $p$-dimensional vectors with at most $p_0$ nonzero coordinates, and $\mathcal{B}_K$ be the unit ball of $\mathcal{H}_K$. Moreover, we define the risk of estimating $\boldsymbol{\gamma}^*$ as

$$R_{\boldsymbol{\gamma}^*}(p_0,p,\mathcal{B}_K):=\inf_{\widehat{\boldsymbol{\gamma}}}\sup_{\boldsymbol{\gamma}^*\in B[p_0,p],f^*\in\mathcal{B}_K}\mathbb{E}[\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2^2],$$

where inf is taken over all possible estimators for $\boldsymbol{\gamma}^*$ in model (1.2). Similarly, we define the risk of estimating $\langle X,f^*\rangle_{\mathcal{L}_2}$ as

$$R_{f^*}(s_0,p,\mathcal{B}_K):=\inf_{\widehat{f}}\sup_{\boldsymbol{\gamma}^*\in B[p_0,p],f^*\in\mathcal{B}_K}\mathbb{E}[\langle X,\widehat{f}-f^*\rangle_{\mathcal{L}_2}^2]=\inf_{\widehat{f}}\sup_{\boldsymbol{\gamma}^*\in B[p_0,p],f^*\in\mathcal{B}_K}\|L_{C^{1/2}}(\widehat{f}-f^*)\|_{\mathcal{L}_2}^2.$$

Note that to derive a sharp lower bound on any minimax error, one focuses on the worst case of the objective function in a hypothesis space, to avoid any meaningless lower bound (e.g. sufficiently close to zero). Technically, Fano inequality and packing entropy are generally adopted to derive a sharp lower bound. Hence, a lower bound of the eigenvalue decay is required to lower bound the interested quantity, which is clearly stated in the following.

**Condition $\widetilde{\mathbf{D}}$**(Entropy condition). For some constant $1/2<r<\infty$, the sequence of eigenvalues $s_\ell$ satisfy that

$$s_\ell\asymp\ell^{-2r},\quad\ell\in\mathcal{N}^+.$$

The following theorem provides the lower bounds of the minimax optimal estimation error for $\boldsymbol{\gamma}^*$ and the predictor error for $f^*$, respectively.

**Theorem 3** *Given $n$ i.i.d. samples from (1.2) with the entropy condition (Condition $\tilde{D}$). When $p$ is diverging and $p_0 \ll p$, the minimax risk for estimating $\boldsymbol{\gamma}^*$ can be bounded from below as*

$$R_{\boldsymbol{\gamma}^*}(p_0, p, \mathcal{B}_K) \gtrsim \frac{p_0 \log(p/p_0)}{n};$$

*the minimax risk for estimating $\langle X, f^* \rangle_{\mathcal{L}_2}$ can be bounded from below as*

$$R_{f^*}(p_0, p, \mathcal{B}_K) \gtrsim \max\Big\{\frac{p_0 \log(p/p_0)}{n}, n^{-\frac{2r}{2r+1}}\Big\}.$$

The proof of Theorem 3 is provided in Appendix B. As mentioned previously, these results indicate that the best possible estimation of $\boldsymbol{\gamma}^*$ is not affected by the existence of nonparametric components, while the minimax risk for estimating the (nonparametric) slope function not only depends on the smoothness itself, but also on the dimensionality and sparsity of the scalar covariates. From the lower bound of $R_{f^*}(p_0, p, \mathcal{B}_K)$, we observe that a rate-switching phenomenon occurring between a sparse regime and a smooth regime. Particularly when $\frac{p_0 \log(p/p_0)}{n}$ dominates $n^{-\frac{2r}{2r+1}}$ corresponding to the sparse regime, the lower bound becomes the classical high dimensional parametric rate $\frac{p_0 \log(p/p_0)}{n}$. Otherwise, this corresponds to the smooth regime and thus has similar behaviors as classical nonparametric models. We also notice that the minimax lower bound obtained for the predictor error generalizes the previous results for the pure functional linar model (Cai and Yuan, 2012).

## 4. Randomized Sketches and Optimization

This section is devoted to considering an approximate algorithm for (2.2), based on constraining the original parameter $\boldsymbol{\alpha} \in \mathcal{R}^n$ to an $m$-dimensional subspace of $\mathcal{R}^n$, where $m \ll n$ is the projection dimension. We define this approximation via a sketch matrix $\mathbb{S} \in \mathcal{R}^{m\times n}$ such that the $m$-dimensional subspace is generated by the row span of $\mathbb{S}$. More precisely, the sketched kernel partial functional estimator is given by first solving

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}_s, \widehat{\boldsymbol{\gamma}}_s) := \arg\min_{\boldsymbol{\alpha}\in\mathcal{R}^m, \boldsymbol{\gamma}\in\mathcal{R}^p} & \frac{1}{n}\boldsymbol{\alpha}(\mathbb{S}\mathbb{K}^c)(\mathbb{S}\mathbb{K}^c)^T\boldsymbol{\alpha} - \frac{2}{n}\boldsymbol{\alpha}^T\mathbb{S}\mathbb{K}^c(\mathbf{y} - \mathbb{Z}\boldsymbol{\gamma}) + \frac{1}{n}\|\mathbf{y} - \mathbb{Z}\boldsymbol{\gamma}\|_2^2 \\ & + \mu^2\boldsymbol{\alpha}^T\mathbb{S}\mathbb{K}^c\mathbb{S}^T\boldsymbol{\alpha} + \lambda\|\boldsymbol{\gamma}\|_1. \end{aligned} \tag{4.1}$$

Then the resulting predictor for the slope function $f^*$ is given as

$$\widehat{f}_s(t) := \sum_{k=1}^n (\mathbb{S}^T\widehat{\boldsymbol{\alpha}}_s)_k B_k(t) = \widehat{\boldsymbol{\alpha}}_s^T\mathbb{S}\mathbf{B}(t), \quad \forall\, t \in \mathcal{T}.$$

where $\mathbf{B}(t) = (B_1(t), ..., B_n(t))^T \in \mathcal{R}^n$, where $B_k(t)$ is defined in Lemma 1. By doing randomized sketches, an approximate form of the kernel estimate $\widehat{\boldsymbol{\alpha}}_s$ can be obtained by solving an $m$-dimensional quadratic program when $\widehat{\boldsymbol{\gamma}}_s$ is fixed, which involves time and space complexity $O(m^3)$ and $O(m^2)$. Computing the approximate kernel matrix is a preprocessing step with time complexity $O(n^2 \log(m))$ for properly chosen projections. It is worthy pointing out that the random sketch techniques are only adopted to facilitate the computational issue which has been well-studied for the nonparametirc regression in a RKHS (Yang et al., 2017; Lin and Cevher, 2020), and some other techniques can also be considered, such as the Nyström type subsampling approach (Rudi et al., 2015).

### 4.1 Alternating Optimization

This section provides the detailed computational issues of the proposed approach. Precisely, we aim to solve the following optimization task that

$$(\widehat{\boldsymbol{\alpha}}_s, \widehat{\boldsymbol{\gamma}}_s) := \underset{\boldsymbol{\alpha}\in\mathcal{R}^m, \boldsymbol{\gamma}\in\mathcal{R}^p}{\operatorname{argmin}} \frac{1}{n}\boldsymbol{\alpha}^T(\mathbb{S}\mathbb{K}^c)(\mathbb{S}\mathbb{K}^c)^T\boldsymbol{\alpha} - \frac{2}{n}\boldsymbol{\alpha}^T\mathbb{S}\mathbb{K}^c(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma}) + \frac{1}{n}(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma})^T(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma}) + \mu^2\boldsymbol{\alpha}^T\mathbb{S}\mathbb{K}^c\mathbb{S}^T\boldsymbol{\alpha} + \lambda\|\boldsymbol{\gamma}\|_1. \tag{4.2}$$

To solve (4.2), a splitting algorithm with proximal operator is applied, which updates the representer coefficients $\boldsymbol{\alpha}$ and the linear coefficients $\boldsymbol{\gamma}$ sequentially. Specifically, at the $t$-th iteration with current solution $(\boldsymbol{\alpha}^t, \boldsymbol{\gamma}^t)$, the following two optimization tasks are solved sequentially to obtain the solution of the $(t+1)$-th iteration

$$\boldsymbol{\alpha}^{t+1} = \underset{\boldsymbol{\alpha}\in\mathcal{R}^m}{\operatorname{argmin}} \Big\{\frac{1}{n}\boldsymbol{\alpha}^T(\mathbb{S}\mathbb{K}^c)(\mathbb{S}\mathbb{K}^c)^T\boldsymbol{\alpha} - \frac{2}{n}\boldsymbol{\alpha}^T\mathbb{S}\mathbb{K}^c(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma}^t) + \mu^2\boldsymbol{\alpha}^T\mathbb{S}\mathbb{K}^c\mathbb{S}^T\boldsymbol{\alpha}\Big\}, \tag{4.3}$$

$$\boldsymbol{\gamma}^{t+1} = \underset{\boldsymbol{\gamma}\in\mathcal{R}^p}{\operatorname{argmin}} \Big\{R_n(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_1\Big\}, \tag{4.4}$$

where $R_n(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\gamma}) := \frac{2}{n}(\boldsymbol{\alpha}^{t+1})^T\mathbb{S}\mathbb{K}^c\mathbb{Z}\boldsymbol{\gamma} + \frac{1}{n}(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma})^T(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma})$.

To update $\boldsymbol{\alpha}$, it is clear that the optimization task (4.3) has an analytic solution that

$$\boldsymbol{\alpha}^{t+1} = \big((\mathbb{S}\mathbb{K}^c)(\mathbb{S}\mathbb{K}^c)^T + n\mu^2\mathbb{S}\mathbb{K}^c\mathbb{S}^T\big)^{-1}\mathbb{S}\mathbb{K}^c(\mathbf{y}-\mathbb{Z}\boldsymbol{\gamma}^t).$$

To update $\boldsymbol{\gamma}$, we first introduce the proximal operator (Moreau, 1962), which is defined as

$$\operatorname{Prox}_{\lambda\|\cdot\|_1}(\mathbf{u}) := \underset{\mathbf{u}}{\operatorname{argmin}} \Big\{\frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{u}\|_1\Big\}. \tag{4.5}$$

Note that the solution of optimization task (4.5) is the well-known soft-thresholding operator with solution that

$$\big(\operatorname{Prox}_{\lambda\|\cdot\|_1}(\mathbf{u})\big)_i = \operatorname{sign}(u_i)(|u_i| - \lambda)_+.$$

Then, for the optimization task (4.4), we have

$$\boldsymbol{\gamma}^{t+1} = \operatorname{Prox}_{\frac{\lambda}{D}\|\cdot\|_1}\Big(\boldsymbol{\gamma}^t - \frac{1}{D}\nabla_{\boldsymbol{\gamma}}R_n(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\gamma}^t)\Big),$$

where $D$ denotes an upper bound of the Lipschitz constant of $R_n(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\gamma}^t)$, and compute $\nabla_{\boldsymbol{\gamma}}R_n(\boldsymbol{\alpha}^{t+1}, \boldsymbol{\gamma}^t) = \frac{2}{n}\mathbb{Z}^T(\mathbb{S}\mathbb{K}^c)^T\boldsymbol{\alpha}^{t+1} + \frac{2}{n}\mathbb{Z}^T\mathbb{Z}\boldsymbol{\gamma}^t - \frac{2}{n}\mathbb{Z}^T\mathbf{y}$. We repeat the above iteration steps until some pre-specified stopping rule is satisfied.

It should be pointed out that the exact value of $D$ is often difficult to determine in large-scale problems. A common way to handle this problem is to use a backtracking scheme (Boyd and Vandenberghe, 2004) as a more efficient alternative to approximately compute an upper bound of it. It is also worthy pointing out that alternating optimization is commonly adopted for solving optimization problem with more than one parameter, and its convergence results have been well-studied in literature (Bezdek and Hathaway, 2003; Li et al., 2019).

### 4.2 Choice of Random Sketch Matrix

In this paper, we consider three random sketch methods, including the sub-Gaussian random sketch (GRS), randomized orthogonal system sketch (ROS) and sub-sampling random sketch (SUB). Precisely, we denote the $i$-th row of the random matrix $\mathbb{S}$ as $\mathbf{s}_i$ and consider three different types of $\mathbf{s}_i$ as follows.

**Sub-Gaussian sketch (GRS):** The row $\mathbf{s}_i$ of $\mathbb{S}$ is zero-mean 1-sub-Gaussian if for any $\mathbf{u} \in \mathcal{R}^n$, we have

$$\mathrm{P}\big(\langle \mathbf{s}_i, \mathbf{u} \rangle \geq t \| \mathbf{u} \|_2\big) \leq e^{-t^2/2}, \quad \forall\, t \geq 0.$$

Note that the row $\mathbf{s}_i$ with independent and identical distributed $N(0,1)$ entries is 1-sub-Gaussian. For simplicity, we further rescale the sub-Gaussian sketch matrix $\mathbb{S}$ such that the rows $\mathbf{s}_i$ have the covariance matrix $\frac{1}{\sqrt{m}}\mathbb{I}_n$, where $\mathbb{I}_n$ denotes a $n$ dimensional identity matrix.

**Randomized orthogonal system sketch (ROS):** The row $\mathbf{s}_i$ of the random matrix $\mathbb{S}$ is formed with i.i.d rows of the form

$$\mathbf{s}_i = \sqrt{\frac{n}{m}} \mathbb{R}\mathbb{H}^T \mathbb{I}_{(i)}, \quad \text{for} \quad i = 1, ..., m,$$

where $\mathbb{R} \in \mathcal{R}^{n \times n}$ is a random diagonal matrix whose diagonal entries are i.i.d. Rademacher variables taking value $\{-1, 1\}$ with equal probability, $\mathbb{H} = \{H_{ij}\}_{i,j=1}^n \in \mathcal{R}^{n \times n}$ is an orthonormal matrix with bounded entries that $H_{ij} \in [-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}]$, and the $n$-dimensional vectors $\mathbb{I}_{(1)}, ..., \mathbb{I}_{(m)}$ are drawn uniformly at random without replacement from the $n$-dimensional identity matrix $\mathbb{I}_n$ .

**Sub-sampling sketches (SUB):** The rows $\mathbf{s}_i$ of the random matrix $\mathbb{S}$ has the form that

$$\mathbf{s}_i = \sqrt{\frac{n}{m}} \mathbb{I}_{(i)},$$

where the $n$-dimensional vectors $\mathbb{I}_{(1)}, ..., \mathbb{I}_{(m)}$ are drawn uniformly at random without replacement from a $n$ dimensional identity matrix. Note that the sub-sampling sketches method can be regarded as a special case of the ROS sketch by replacing the matrix $\mathbb{R}^T\mathbb{H}$ with a $n$-dimensional identity matrix $\mathbb{I}_n$.

### 4.3 Choice of the Sketch Dimension

In practice, we are interested in the $m \times n$ sketch matrices with $m \ll n$ to enhance computational efficiency. Note that the existence of a $n \times n$ kernel matrix in Lemma 1 is only a sufficient condition for equivalent optimization. It has been shown theoretically in the kernel regression (Yang et al., 2017) that the kernel matrix can be compressed to be the one with small size, based on some intrinsic low-dimensional notations. Despite the model difference from Yang et al. (2017), our kernel matrix $\mathbb{K}^c$ does not depend on the scalar covariates $\mathbf{Z}$, and thus those derived results for the kernel regression are still applicable to our case.

Consider the eigen-decomposition $\mathbb{K}^c = \mathbb{U}\mathbb{D}\mathbb{U}^T$ of the kernel matrix, where $\mathbb{U} \in \mathcal{R}^{n\times n}$ is an orthonormal matrix of eigenvectors, and $\mathbb{D} = \text{diag}\{\hat{\mu}_1, ..., \hat{\mu}_n\}$ is a diagonal matrix of eigenvalues, where $\hat{\mu}_1 \geq \hat{\mu}_2 \geq ... \geq \hat{\mu}_n \geq 0$. We define the kernel complexity function as

$$\widehat{\mathcal{R}}(\delta) = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \min\{\delta, \hat{\mu}_j\}}.$$

The critical radius is defined to be the smallest positive solution of $\delta_n > 0$ to the inequality

$$\widehat{\mathcal{R}}(\delta) \leq \delta^2/\sigma.$$

Note that the existence and uniqueness of this critical radius is guaranteed for any kernel class. Based on this, we define the statistical dimension of the kernel is

$$d_n := \min\{j \in [n] : \hat{\mu}_j \leq \delta_n^2\}.$$

Recall that, Theorem 2 in Yang et al. (2017) shows that various forms of randomized sketches can achieve the minimax rate using a sketch dimension proportional to the statistical dimension $d_n$. In particular, for Gaussian sketches and ROS sketches, the sketch dimension $m$ is required satisfy a lower bound of the form

$$m \geq \begin{cases} cd_n & \text{for Gaussian sketches,} \\ cd_n \log^4(n) & \text{for ROS sketches.} \end{cases}$$

Here $c$ is some constant. In this paper, we adopt this specified sketch dimension $m$ to implement our experiments.

## 5. Numerical Experiments

In this part, we examine the numerical performance of the proposed method in several simulated examples and one real-life example. Specifically, we apply the proposed method to some simulated data under various scenarios to verify our theoretical findings in Section 5.1. In Section 5.2, we apply the proposed method to a real dataset from the National Mortality, Morbidity, and Air Pollution Study to illustrate its real application.

### 5.1 Simulated Examples

In this section, we illustrate the numerical performance of the proposed method with random sketches in two numerical examples. Specifically, we assume that the true generating model is

$$Y_i = \int_{\mathcal{T}} f^*(t)X_i(t)dt + \mathbf{Z}_i^T\boldsymbol{\gamma}^* + \varepsilon_i, \tag{5.1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ with $\sigma = 1$, and $\mathcal{T}$ is set as $[0, 1]$. Note that the generating scheme is the same as that in Hall and Horowitz 2007 and Yuan and Cai 2010. In practice, the integrals in calculation of $\mathbb{B}$ and $\mathbb{K}^c$ are approximated by summations, and thus we generate 1000 points in $\mathcal{T} = [0, 1]$ with equal distance and evaluate the integral by using the generated points. As the proper choice of tuning parameters plays a crucial role in achieving the desired performance of the proposed method, we adopt 5-fold cross-validation to select the optimal values of the tuning parameters $\mu$ and $\lambda$.

In all the simulated cases, we consider a RKHS $\mathcal{H}_K$ induced by a reproducing kernel function on $\mathcal{T} \times \mathcal{T}$ that

$$\begin{aligned} K(s,t) &= \sum_{k\geq 1} \frac{2}{(k\pi)^4} \cos(k\pi s)\cos(k\pi t) \\ &= \sum_{k\geq 1} \frac{1}{(k\pi)^4} \cos(k\pi(s-t)) + \sum_{k\geq 1} \frac{1}{(k\pi)^4} \cos(k\pi(s+t)) \\ &= -\frac{1}{3} B_4\Big(\frac{|s-t|}{2}\Big) - \frac{1}{3} B_4\Big(\frac{s+t}{2}\Big), \end{aligned}$$

where $B_{2m}(\cdot)$ denotes the $2m$-th Bernoulli polynomial that

$$B_{2m}(s) = (-1)^{m-1} 2(2m)! \sum_{k\geq 1} \frac{\cos(2\pi k s)}{(2\pi k)^{2m}}, \text{ for any } s \in \mathcal{T}.$$

Note that the RKHS $\mathcal{H}_K$ induced by $K(s,t)$ contains the functions in a linear span of the cosine basis that

$$f(s) = \sqrt{2} \sum_{k\geq 1} g_k \cos(k\pi s), \text{ for any } s \in \mathcal{T}.$$

such that $\sum_{k\geq 1} k^4 g_k^2 < \infty$ and the endowed norm is

$$\|f\|_K^2 = \int_{\mathcal{T}} \big(\sqrt{2} \sum_{k\geq 1} (k\pi)^2 g_k \cos(k\pi t)\big)^2 dt = \sum_{k\geq 1} (k\pi)^4 g_k^2.$$

The performance of the proposed method is evaluated under the following two numerical examples.

**Example 1**. We consider the true slope function $f^*$ and the random function $X$ are

$$f^*(t) = \sum_{k=1}^{50} 4(-1)^{k+1} k^{-2} \sqrt{2} \cos(k\pi t),$$

and

$$X(t) = \xi_1 U_1 + \sum_{k=2}^{50} \xi_k U_k \sqrt{2} \cos(k\pi t),$$

where $U_k \sim U(-\sqrt{3}, \sqrt{3})$ and $\xi_k = (-1)^{k+1} k^{-v/2}$. For the linear part, the true regression coefficients are set as $\boldsymbol{\gamma}^0 = (2, -2, 0, ..., 0)^T$ and the sample $\mathbb{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T \in \mathcal{R}^{n\times p}$ with $\mathbf{Z}_i = (z_{i1}, ..., z_{ip})^T$ are generated i.i.d. as $z_{ij} \sim U(0, 1)$.

**Example 2.** The generating scheme is the same as Example 1, except that

$$\xi_k = \begin{cases} 1, & k = 1, \\ 0.2(-1)^{k+1}(1 - 0.0001k), & 2 \le k \le 4, \\ 0.2(-1)^{k+1}\big[(5\lfloor k/5 \rfloor)^{-v/2} - 0.0001(k \bmod 5)\big], & k \ge 5. \end{cases}$$

Clearly, $\xi_k^2$'s are the eigenvalues of the covariance function $C$ and we choose $v = 1.1, 2$ and 4 to evaluate the effect of the smoothness of $\xi_k$ in the both examples. Note that in Example 1, these eigenvalues are well spaced, and the covariance function $C$ and the reproducing kernel $K$ share the same eigenvalues, while in Example 2, these eigenvalues are closely spaced and the alignment between $K$ and $C$ is considered.

To comprehend the effect of sample size, we consider the same settings as in Yang et al. (2017) that $n = 256, 512, 1024, 2048, 4096, 8192$ and $16384$ and conservatively, take $m = \lfloor n^{1/3} \rfloor$ for the three random sketch methods introduced in Section 4.2. Note that with the choice of $m$, the time and store complexities reduce to $O(n)$ and $O(n^{2/3})$, respectively. Each scenario is replicated 50 times and the performance of the proposed method is evaluated by various measures, including the estimation accuracy of the linear coefficients, the integrated prediction error in terms of the slope function and the prediction error. Specifically, the estimation accuracy of the linear coefficients is evaluated by $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\|_2^2 = \sum_{l=1}^p (\widehat{\gamma}_l - \gamma_l^0)^2$, and Figure 1 shows the estimation accuracy of the coefficients with different choice of $v$.

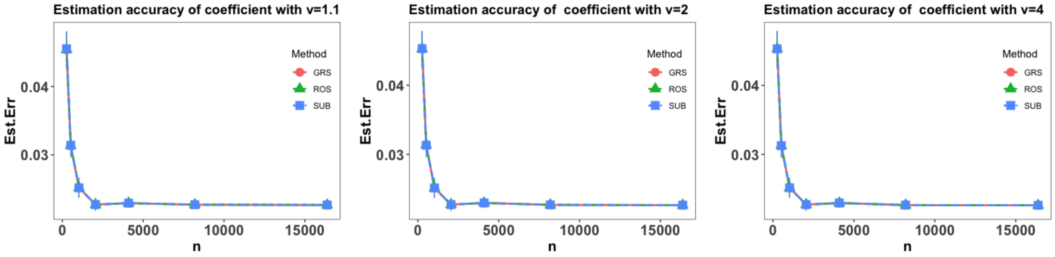

Figure 1: Estimation accuracy of the coefficients in Example 1 under various scenarios.

It is clear that the estimation error of the coefficients converges linearly as sample size $n$ increases and becomes stable when $n$ is sufficiently large, and the three employed sketch methods have similar performance. It is also interesting to notice that the convergence patterns under difference choice of $v$ are almost the same, which concurs with our theoretical findings that estimation of $\boldsymbol{\gamma}^*$ is not affected by the existence of nonparametric components in Theorems 1 and 3.

Let $(Y', X'(\cdot), \mathbf{Z}')$ denotes an independent copy of $(Y, X(\cdot), \mathbf{Z})$ and the integrated prediction error in terms of the slope function is reported by

$$\widehat{\mathbb{E}}_{X'}\|\widehat{f} - f^*\|^2 = \widehat{\mathbb{E}}_{X'}\Big(\int_{\mathcal{T}} (\widehat{f}(t) - f^*(t))X'(t)dt\Big)^2$$

The empirical expectation $\widehat{\mathbb{E}}$ is evaluated by a testing sample with size 10000 and $\widehat{Y}' = \int_{\mathcal{T}} \widehat{f}(t)X_i'(t)dt + (\mathbf{Z}_i')^T\widehat{\boldsymbol{\gamma}}$ and the numerical performance are summarized in Figure 2.

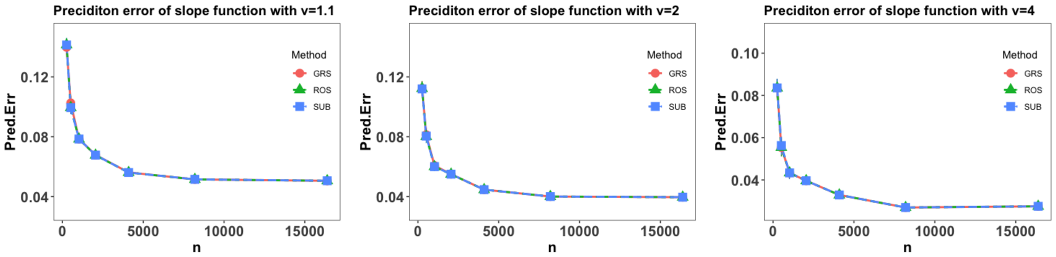

Figure 2: Prediction error of the slope function in Example 1 under various scenarios.

Note that Figure 2 suggests that the prediction error of the slope function converges at some polynomial rate as sample size $n$, which agrees with our theoretical results in Section 3, and the three employed sketch methods yield similar numerical performance. Moreover, it can be seen that with the increase of the value of $v$, the prediction error goes down, which also concurs with our theoretical findings in Theorems 2 and 3 that the faster decay rate of the eigenvalues, the smaller the prediction error.

We also report the integrated prediction error of the response by calculating

$$\widehat{\mathbb{E}}_{Y',X'}\|\widehat{Y}' - Y'\|_2^2.$$

The empirical expectation $\widehat{\mathbb{E}}$ is also evaluated by a testing sample with size 10000 and the numerical performance are summarized in Figure 3.

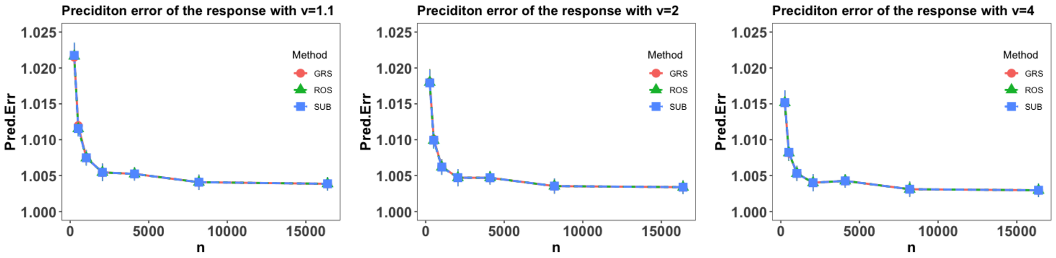

Figure 3: Prediction error of the response in Example 1 under various scenarios.

Clearly, we conclude that prediction error of the response converges at some polynomial rate as sample size $n$ and the prediction error becomes smaller with $v$ increases, which agrees with our theoretical results in Theorem 2. It is also interesting to point out that the three employed sketch methods yield similar numerical performance and the prediction errors tends to converge to 1, which is the variance of $\varepsilon$ in the true modelling. This verifies the efficiency of the proposed estimation and the proper choice of $m$.

Note that the numerical results in Example 2 where the eigenvalues are closely spaced are similar to those in the case with well-spaced eigenvalues in Example 1. Figure 4 shows the numerical performance under the closely spaced eigenvalues setting in Example 2.

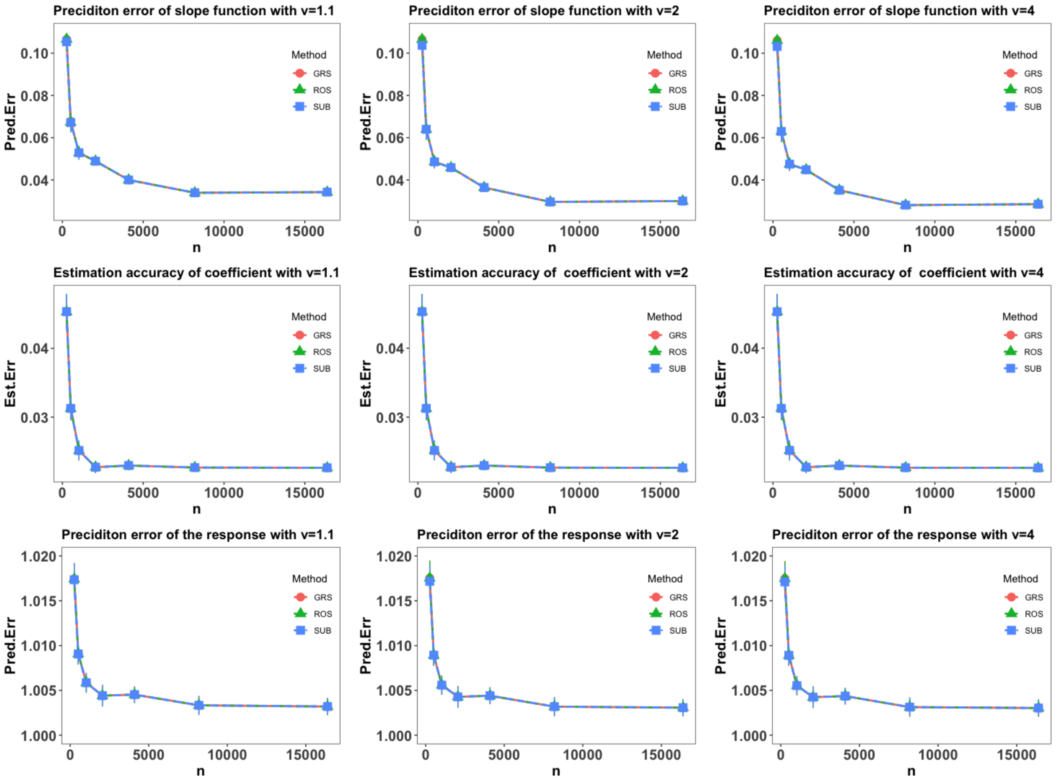

Figure 4: Numerical performance of the proposed method in Example 2 under various scenarios.

## 5.2 Real Data Analysis

In this section, we apply the proposed method to analyze a real dataset from the National Mortality, Morbidity, and Air Pollution Study, where the measurements of air pollution and the counts of mortality for several U.S. cities are collected during the census in 2000. The main interest of this study is to investigate that how the air pollution and some other factors from the U.S. census affect the nonaccidental mortality rate across different cities. We consider the measurements of PM 2.5 (The particulate matter with an aerodynamic diameter of less than 2.5 $\mu$m) collected from 1 April 2000 to 31 August 2000 as the random functional feature, which has attracted tremendous attention due to its association with many adverse health outcomes, and treat 7 factors collected from the U.S. census in 2000 as the scalar predictors, including the household owners proportion, the urban population proportion, the population proportion with at least a high school diploma, the population proportion with at least a university degree, the population proportion below the poverty line, land area per individual and water area per individual. We are interested in studying how the functional feature and scalar predictors affect the log-transformed total nonaccidental mortality rate of individuals with age at least 65 in the next month, September 2000, since this group accounts for the majority of nonaccidental deaths. Following a similar treatment as that in Kong et al., 2016, we remove the records of cities with more than ten consecutive

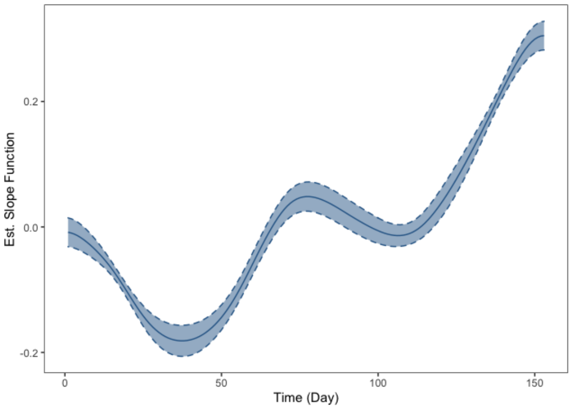

Figure 5: The estimated slope function (solid) and the corresponding 95% confidence interval (dashed) based on 1000 bootstrap samples.

missing PM2.5 measurements, and thus the total number of cities considered in our study is 69.

Since the sample size in the dataset is relatively small, we apply the proposed method without using the random sketch technique by setting $\mathbb{S}$ as the identity matrix. To be more precise, we implement the proposed method with the whole data to estimate the slope function and coefficients of the scalar predictors, and then we refit the selected model by setting $\mu = 0$ and $\lambda = 0$ based on 1000 bootstrap samples to compute the standard errors. Consequently, the proposed method finds that only one scalar predictor, the population proportion with at least a university degree, has a negative effect on the log-transformed total nonaccidental mortality rate with the estimated coefficient -0.19 and standard error 0.004. This indicates that cities with larger proportion of population with high education level have a lower mortality rate. Figure 5 illustrates the estimated slope function as well as the corresponding 95% confidence interval using the bootstrap samples.

From Figure 5, it is clear that the estimated slope function has an increasing trend with Time, especially in July and August, and thus the higher concentrations of PM2·5 in the summer can lead to the increasing of the nonaccidental mortality in the next period, which coincides with the conclusions in Kong et al., 2016 and the reference therein. To further evaluate the effect of functional feature, PM2.5, we further fit two models where one uses the selected scalar predictor and functional feature and the other one only uses the selected scalar predictor, and the obtained values of $R^2$ are 0.3487 and 0.1866, respectively. This further supports the significance of the functional feature, PM2.5, in our analysis.

## 6. Conclusion

This paper establishes the optimal minimax rate for the estimation of partially functional linear model (PFLM) under kernel-based and high dimensional setting. The optimal minimax rates of estimation is established by using various techniques in empirical process theory for analyzing kernel classes, and an efficient numerical algorithm based on randomized sketches of the kernel matrix is implemented to verify our theoretical findings. It is interesting to point out that the random sketch techniques in Section 4 are only adopted to facilitate the computational issue and it would be of great interest to further study the theoretical properties of the approximate estimator in the future work. Moreover, we believe that the current work provides a general routine to investigate the optimal properties of semi-parametric functional models under various settings, and thus it can be extended to other interesting kernel problems, such as combing the semi-parametric functional model with the sparsity-induced kernel methods and mis-specified kernel settings, to establish the minimax rates.

## Acknowledgments

The authors thank the Action Editor and the referees for their constructive suggestions, which significantly improve this paper. Shaogao Lv's research was supported in part by NSFC-11871277, and 2022 Qinglan Project of the Young and Middle-aged Academic Leader of Jiangsu College and University. Xin He's research was supported in part by NSFC-11901375, Shanghai Pujiang Program 2019PJC051, Shanghai Research Center for Data Science and Decision Technology, and Program for Innovative Research Team of Shanghai University of Finance and Economics. Junhui Wang's research was supported in part by HK RGC Grants GRF-11304520, GRF-11301521, GRF-11311022, CUHK Startup Grant 4937091, and CUHK Direct Grant 4053588.

## Appendix A: Technical proofs for upper bounds

For any constant $R > 0$, we define

$$\tau_{R,\mu}(g) := \tau(g) := \frac{\lambda\|\boldsymbol{\gamma}\|_1}{R\sqrt{\delta_0/2}} + \sqrt{\left\|\langle X, f\rangle_{\mathcal{L}_2} + \mathbf{Z}^T\boldsymbol{\gamma}\right\|^2 + \mu^2\|f\|_K^2}, \, \forall\, g \in \mathcal{G}.$$

where $\delta_0$ is a fixed small constant determined later.

For each $R > 0$, we define

$$\mathcal{G}(R) := \{g : \tau_{R,\mu}(g) \leq R\}$$

and the event

$$\mathcal{T} := \mathcal{T}_1(\delta_0, R) \cap \mathcal{T}_2(\delta_0, R)$$

where

$$\mathcal{T}_1(\delta_0, R) := \Big\{(X, Z) : \sup_{g\in\mathcal{G}(R)} \big|\|g\|_n^2 - \|g\|^2\big| \leq \delta_0 R^2\Big\}$$

and

$$\mathcal{T}_2(\delta_0, R) := \Big\{(\mathbb{U}, \boldsymbol{\varepsilon}) : \sup_{g\in\mathcal{G}(R)} \big|\boldsymbol{\varepsilon}^T g(\mathbb{U})/n\big| \leq \delta_0 R^2\Big\},$$

with $g(\mathbb{U}) = (g(\mathbf{U}_1), ..., g(\mathbf{U}_n))^T$ and $\mathbf{U}_i = (X_i(\cdot), \mathbf{Z}_i)$.

**Lemma 2** *Suppose that Condition A holds. Assume that a constant $R$ can be chosen such that*

$$\mu^2 \le \frac{\delta_0 R^2}{8\|f^*\|_K^2},\ \frac{\lambda^2 p_0}{\Lambda_{\min}^2} \le \frac{1}{4}\delta_0 R^2.$$

*Then, on the event $\mathcal{T}$, we have*

$$\tau(\widehat{g} - g^*) \le R.$$

**Proof of Lemma 2.** We first define a linear combination of $\widehat{g}$ and $g^*$ by

$$\tilde{g} = s\widehat{g} + (1-s)g^* = \langle X, \tilde{f}\rangle_{\mathcal{L}_2} + \mathbf{Z}^T\tilde{\boldsymbol{\gamma}}$$

where $\tilde{f} = s\widehat{f} + (1-s)f^*$, $\tilde{\boldsymbol{\gamma}} = s\hat{\boldsymbol{\gamma}} + (1-s)\boldsymbol{\gamma}^*$ and $s = \frac{R}{R+\tau(\widehat{g}-g^*)}$. By convexity and the definition of $(\widehat{f}, \widehat{\boldsymbol{\gamma}})$ in (2.1), we have

$$\begin{aligned}\|\mathbf{y} - \tilde{g}\|_n^2 + \mu^2\|\tilde{f}\|_K^2 + \lambda\|\tilde{\boldsymbol{\gamma}}\|_1 \le\ & s\big(\|\mathbf{y} - \langle X, \widehat{f}\rangle_{\mathcal{L}_2} - \mathbf{Z}^T\hat{\boldsymbol{\gamma}}\|_n^2 + \mu^2\|\widehat{f}\|_K^2 + \lambda\|\hat{\boldsymbol{\gamma}}\|_1\big)\\ & + (1-s)\big(\|\mathbf{y} - \langle X, f^*\rangle_{\mathcal{L}_2} - \mathbf{Z}^T\boldsymbol{\gamma}^*\|_n^2 + \mu^2\|f^*\|_K^2 + \lambda\|\boldsymbol{\gamma}^*\|_1\big)\\ \le\ & \|\mathbf{y} - g^*\|_n^2 + \mu^2\|f^*\|_K^2 + \lambda\|\boldsymbol{\gamma}^*\|_1.\end{aligned}$$

This is referred to as "basic inequality". By plugging the model $\mathbf{y} = g^*(\mathbf{U}) + \boldsymbol{\varepsilon}$ into the above inequality, it can be rewritten as

$$\|\tilde{g} - g^*\|_n^2 + \mu^2\|\tilde{f}\|_K^2 + \lambda\|\tilde{\boldsymbol{\gamma}}\|_1 \le 2\boldsymbol{\varepsilon}^T\big(\tilde{g} - g^*\big)(\mathbb{U})/n + \mu^2\|f^*\|_K^2 + \lambda\|\boldsymbol{\gamma}^*\|_1.$$

Note that

$$\tau(\tilde{g} - g^*) = s\tau(\widehat{g} - g^*) = \frac{R\tau(\widehat{g} - g^*)}{R + \tau(\widehat{g} - g^*)} \le R,$$

which means that $\tilde{g} - g^* \in \mathcal{G}(R)$. Hence, on $\mathcal{T}(\delta_0, R)$, we have

$$\|\tilde{g} - g^*\|^2 + \mu^2\|\tilde{f}\|_K^2 + \lambda\|\tilde{\boldsymbol{\gamma}}\|_1 \le 3\delta_0 R^2 + \mu^2\|f^*\|_K^2 + \lambda\|\boldsymbol{\gamma}^*\|_1.$$

By the identity $\|\tilde{\boldsymbol{\gamma}}\|_1 = \|\tilde{\boldsymbol{\gamma}}_{S_0^c}\|_1 + \|\tilde{\boldsymbol{\gamma}}_{S_0}\|_1$ and the triangle inequality, there holds

$$\|\tilde{g} - g^*\|^2 + \mu^2\|\tilde{f}\|_K^2 + \lambda\|\tilde{\boldsymbol{\gamma}}_{S_0^c}\|_1 \le 3\delta_0 R^2 + \mu^2\|f^*\|_K^2 + \lambda\|(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)_{S_0}\|_1.$$

In addition, since $\widetilde{\mathbf{Z}}$ is orthogonal with $\langle X, f\rangle_{\mathcal{L}_2}$ for any $f \in \mathcal{H}_K$, this leads to

$$\begin{aligned}&\|\widetilde{\mathbf{Z}}^T(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\|^2 + \|\langle X, \tilde{f} - f^*\rangle_{\mathcal{L}_2} + \Pi_{\mathbf{Z}|X}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\|^2 + \mu^2\|\widetilde{f}\|_K^2 + \lambda\|\widetilde{\boldsymbol{\gamma}}_{S_0^c}\|_1\\ &\le 3\delta_0 R^2 + \mu^2\|f^*\|_K^2 + \lambda\|(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)_{S_0}\|_1.\end{aligned} \tag{.1}$$

By the basic inequality $uv \le u^2 + v^2/4$ for any $u, v \in \mathcal{R}$, we also get

$$\begin{aligned}\lambda\|(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)_{S_0}\|_1 &\le \lambda\sqrt{p_0}\|(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)_{S_0}\|_2 \le \lambda\sqrt{p_0}\|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2\\ &\le \frac{\lambda\sqrt{p_0}}{\Lambda_{\min}}\|\widetilde{\mathbf{Z}}^T(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\| \le \frac{\lambda^2 p_0}{\Lambda_{\min}^2} + \frac{\|\widetilde{\mathbf{Z}}^T(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\|^2}{4}.\end{aligned} \tag{.2}$$

Also, we notice that $\|\tilde{f}\|_K^2 \geq \frac{1}{2}\|\tilde{f}-f^*\|_K^2 - \|f^*\|_K^2$. This together with (.1) and (.2) implies that

$$\frac{3}{4}\|\widetilde{\mathbf{Z}}^T(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2 + \|\langle X, \tilde{f}-f^*\rangle_{\mathcal{L}_2} + \Pi_{\mathbf{Z}|X}(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2 + \frac{\mu^2}{2}\|\tilde{f}-f^*\|_K^2$$
$$\leq 3\delta_0 R^2 + 2\mu^2\|f^*\|_K^2 + \frac{\lambda^2 p_0}{\Lambda_{\min}^2} \leq 3\delta_0 R^2 + \frac{1}{4}\delta_0 R^2 + \frac{1}{4}\delta_0 R^2, \tag{.3}$$

where the last inequality follows from the constraints that $\mu^2 \leq \frac{\delta_0 R^2}{8\|f^*\|_K^2}$ and $\frac{\lambda^2 p_0}{\Lambda_{\min}^2} \leq \frac{1}{4}\delta_0 R^2$. From this, it easily follows that

$$\|\tilde{g}-g^*\|^2 + \frac{\mu^2}{2}\|\tilde{f}-f^*\|_K^2 \leq 10\delta_0 R^2. \tag{.4}$$

On the other hand, adding the term $\lambda\|(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1$ on the both sides of (.1), we have

$$\lambda\|(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|_1 \leq 3\delta_0 R^2 + \mu^2\|f^*\|_K^2 + 2\lambda\|(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1$$
$$\leq 3\delta_0 R^2 + \frac{\delta_0 R^2}{8} + \frac{\delta_0 R^2}{2} + \|\widetilde{\mathbf{Z}}^T(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2/2, \tag{.5}$$

where the last inequality follows from the constraints on $(\mu^2, \lambda)$ and (.2). Note that by (.3), it holds $\|\widetilde{\mathbf{Z}}^T(\tilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2 \leq \frac{14}{3}\delta_0 R^2$ . Hence, combining (.4) with (.5), we conclude that

$$\tau(\tilde{g}-g^*) \leq (\sqrt{20\delta_0} + 6\sqrt{2\delta_0})R \leq \frac{R}{2},$$

provided that $\delta_0$ is small properly such that $\delta_0 \leq 1/(4\sqrt{5}+12\sqrt{2})$. Moreover, we have

$$\tau(\tilde{g}-g^*) = s\tau(\widehat{g}-g^*) = \frac{R\tau(\widehat{g}-g^*)}{R+\tau(\widehat{g}-g^*)} \leq \frac{R}{2},$$

which implies that

$$\tau(\widehat{g}-g^*) \leq R.$$

This completes the proof. ■

**Appendix $A_1$: For the event $\mathcal{T}$**

We now show that the event $\mathcal{T}$ has probability close to one. To this end, a concentration inequality will be applied. Lemma 3 is from Bousquet (2002), who improves the results from Ledoux (1997).

**Lemma 3** *(Concentration Theorem (Bousquet, 2002)) Let $U_1, ..., U_n$ be independent random variables with values in some space $\mathcal{U}$ and let $H$ be a class of real-valued functions on $\mathcal{U}$, satisfying for some positive constants $\eta_n$ and $\tau_n$,*

$$\|h\|_\infty \leq \eta_n, \quad \text{and } \frac{1}{n}\sum_{i=1}^n var(h(U_i)) \leq \tau_n^2, \quad \forall\, h \in H.$$

*Define $\mathcal{S} := \sup_{h\in H}\left|\frac{1}{n}\sum_{i=1}^n \big(h(U_i) - \mathbb{E}h(U_i)\big)\right|$. Then for $t > 0$*

$$\mathbb{P}\Big(\mathcal{S} \geq \mathbb{E}(\mathcal{S}) + t\sqrt{2(\tau_n^2 + 2\eta_n\mathbb{E}(\mathcal{S}))} + \frac{2\eta_n t^2}{3}\Big) \leq \exp[-nt^2].$$

**Lemma 4** *Suppose that Conditions A-D hold true, and we take $\mu^2 \geq c(\delta_0)R^2$ with $c(\delta_0) < \delta_0/(8\|f^*\|_K^2)$ and $\lambda^2 \leq R^2 \leq \lambda \leq 1$. For constants $\delta_1, \delta_1''$ and $\kappa_1$ with our suitable choices, we set $\lambda_0 := \sqrt{2\log(2p)/n}$ and*

$$\delta_1 \lambda \geq \lambda_0, \ \mu^2 \geq \kappa_1 n^{-\frac{2r}{2r+1}}.$$

*Then we conclude*

$$\sup_{g \in \mathcal{G}(R)} \left| \|g\|_n^2 - \|g\|^2 \right| \leq \delta_0 R^2$$

*with probability at least $1 - \exp[-n(\delta_1'')^2\mu^2]$.*

**Proof of Lemma 4.** To verify all the conditions of Lemma 3, we denote $\mathcal{S} := \sup_{g \in \mathcal{G}(R)} \left| \|g\|_n^2 - \|g\|^2 \right|$ with $H =: \mathcal{G}(R)$ and $U := (X, \mathbf{Z})$. Direct computation yields that

$$\|g^2\|_\infty = \left\| \left( \langle X, f \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma} \right)^2 \right\|_\infty \leq 2\langle X, f \rangle_{\mathcal{L}_2}^2 + 2C_z^2 \|\boldsymbol{\gamma}\|_1^2 \leq 2\kappa^2 \|X\|_{\mathcal{L}_2}^2 \|f\|_K^2 + 2C_z^2 \|\boldsymbol{\gamma}\|_1^2,$$

where $\kappa := \max_{s,t \in \mathcal{T}} |K(s,t)|$. Note that for $g \in \mathcal{G}(R)$, it follows that

$$\|\boldsymbol{\gamma}\|_1 \leq \frac{\sqrt{\delta_0/2}R^2}{\lambda}, \text{ and } \|f\|_K^2 \leq \frac{R^2}{\mu^2},$$

which implies that

$$\|g^2\|_\infty \leq 2\kappa^2 \|X\|_{\mathcal{L}_2}^2 \frac{R^2}{\mu^2} + \delta_0 C_z^2 \frac{R^4}{\lambda^2} \leq 2\kappa^2 \|X\|_{\mathcal{L}_2}^2 / c(\delta_0) + \delta_0 C_z^2,$$

where the last inequality follows from the fact that $\mu^2 \geq c(\delta_0)R^2$ and that $R^2 \leq \lambda \leq 1$. By taking $\tilde{C} := 2\kappa^2 \|X\|_{\mathcal{L}_2}^2 / c(\delta_0) + \delta_0 C_z^2$, for any $g \in \mathcal{G}(R)$, we also have

$$var(g^2) \leq \mathbb{E}[g^4] \leq \|g^2\|_\infty \mathbb{E}[g^2] \leq \tilde{C} R^2. \tag{.6}$$

We still need to provide an upper bound of $\mathbb{E}[\mathcal{S}]$. Let $\{\sigma_i\}_{i=1}^n$ be a Rademacher sequence independent of $\{(X_i, \mathbf{Z}_i)\}_{i=1}^n$. By symmetrization [see e.g. van der Vaart and Wellner (1996)], we have

$$\begin{aligned} \mathbb{E}[\mathcal{S}] \leq 2\mathbb{E}\Big( \sup_{g \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^n g_i^2 \sigma_i \Big| \Big) &\leq 2\mathbb{E}\Big( \sup_{f \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^n \langle X_i, f \rangle_{\mathcal{L}_2}^2 \sigma_i \Big| \Big) \\ &+ 2\mathbb{E}\Big( \sup_{\boldsymbol{\gamma} \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i^T \boldsymbol{\gamma})^2 \sigma_i \Big| \Big) + 4\mathbb{E}\Big( \sup_{g \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^n (\mathbf{Z}_i^T \boldsymbol{\gamma}) \langle X_i, f \rangle_{\mathcal{L}_2} \sigma_i \Big| \Big). \end{aligned}$$

In the following, we bound the above three quantities respectively. Note that if $\langle X, f \rangle_{\mathcal{L}_2} + Z^T \boldsymbol{\gamma} \in \mathcal{G}(R)$, Condition B leads to

$$\|\mathbf{Z}^T \boldsymbol{\gamma}\|_\infty \leq C_z \frac{\sqrt{\delta_0/2}R^2}{\lambda} \leq C_z \sqrt{\frac{\delta_0}{2}},$$

where the last inequality follows from the fact that $R^2 \leq \lambda$. By the contraction inequality of Rademacher complexity [see Ledoux and Talagrand (1991)], it holds

$$\mathbb{E}\Big( \sup_{\gamma \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Z}_i^T \gamma)^2 \sigma_i \Big| \Big) \leq 4C_z \sqrt{\frac{\delta_0}{2}} \mathbb{E}\Big( \sup_{\gamma \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Z}_i^T \gamma) \sigma_i \Big| \Big).$$

Moreover, we have

$$\begin{aligned} \mathbb{E}\Big( \sup_{\gamma \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Z}_i^T \gamma) \sigma_i \Big| \Big) &\leq \mathbb{E}\Big( \sup_{\gamma \in \mathcal{G}(R)} \Big\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \sigma_i \Big\|_\infty \|\gamma\|_1 \Big) \\ &\leq \frac{\sqrt{\delta_0/2} R^2}{\lambda} \mathbb{E}\Big\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \sigma_i \Big\|_\infty \leq C_z \frac{\lambda_0 R^2}{\sqrt{2/\delta_0} \lambda} \leq \big( \delta_1 C_z \sqrt{\delta_0/2} \big) R^2, \end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, the third inequality follows from the fact that $\mathbb{E}\Big\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \sigma_i \Big\|_\infty \leq \lambda_0 C_z$, and the last inequality follows from the condition that $\lambda \geq \lambda_0 / \delta_1$.

Combining the above two inequalities, it holds

$$\mathbb{E}\Big( \sup_{\gamma \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Z}_i^T \gamma)^2 \sigma_i \Big| \Big) \leq 2\delta_1 C_z^2 \delta_0 R^2. \tag{.7}$$

Next, we provide a sharp bound on $\mathbb{E}\Big( \sup_{f \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, f \rangle_{\mathcal{L}_2}^2 \sigma_i \Big| \Big)$. As above, it is shown that $|\langle X, f \rangle_{\mathcal{L}_2}| \leq \kappa \|X\|_{\mathcal{L}_2} \|f\|_K \leq \frac{\kappa R}{\mu} \|X\|_{\mathcal{L}_2}$. By the contraction property of Rademacher sequences again, we have

$$\begin{aligned} \mathbb{E}\Big( \sup_{f \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, f \rangle_{\mathcal{L}_2}^2 \sigma_i \Big| \Big) &\leq 2\kappa \mathbb{E}[\|X\|_{\mathcal{L}_2}] \frac{R}{\mu} \cdot \mathbb{E}\Big( \sup_{f \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, f \rangle_{\mathcal{L}_2} \sigma_i \Big| \Big) \\ &\leq c_6 \kappa \mathbb{E}[\|X\|_{\mathcal{L}_2}] \kappa_1^{-\frac{2r+1}{4r}} R^2, \end{aligned} \tag{.8}$$

which follows from the obtained result in Appendix. Similarly, we have

$$\begin{aligned} \mathbb{E}\Big( \sup_{g \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} (\mathbf{Z}_i^T \gamma) \langle X_i, f \rangle_{\mathcal{L}_2} \sigma_i \Big| \Big) &\leq \mathbb{E} \sup_{g \in \mathcal{G}(R)} \Big\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{Z}_i \langle X_i, f \rangle_{\mathcal{L}_2} \sigma_i \Big\|_\infty \|\gamma\|_1 \\ &\leq \frac{R^2}{\lambda} \sqrt{\delta_0/2} \mathbb{E} \max_{1 \leq j \leq p} \sup_{f \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} z_{ij} \langle X_i, f \rangle_{\mathcal{L}_2} \sigma_i \Big| \\ &\leq \frac{R^2}{\lambda} C_z \sqrt{\delta_0/2} \mathbb{E}\Big( \sup_{g \in \mathcal{G}(R)} \Big| \frac{1}{n} \sum_{i=1}^{n} \langle X_i, f \rangle_{\mathcal{L}_2}^2 \sigma_i \Big| \Big) \\ &\leq c_6 C_z \sqrt{\delta_0/2} \kappa \mathbb{E}[\|X\|_{\mathcal{L}_2}] \kappa_1^{-\frac{2r+1}{4r}} R^2, \end{aligned} \tag{.9}$$

where the third inequality follows from the contraction property of Rademacher complexity, and the last inequality follows from (.18) in Appendix B. Along the lines of (.7), (.8) and (.9), we get

$$\mathbb{E}[\mathcal{S}] \leq \big( 4\delta_1 C_z^2 \delta_0 + 2c_6 \kappa \mathbb{E}[\|X\|_{\mathcal{L}_2}] \kappa_1^{-\frac{2r+1}{4r}} + 4c_6 C_z \sqrt{\delta_0/2} \kappa \mathbb{E}[\|X\|_{\mathcal{L}_2}] \kappa_1^{-\frac{2r+1}{4r}} \big) R^2.$$

Therefore, by the concentration theorem in Lemma 3, we have

$$\mathbb{P}\Big(\mathcal{S} \geq DR^2 + \sqrt{\frac{t}{n}}\sqrt{2(\tilde{C}R^2 + 2\tilde{C}DR^2)} + \frac{2\tilde{C}^2 t}{3n}\Big) \leq \exp[-t], \quad \forall\, t > 0,$$

where $D := 4\delta_1 C_z^2 \delta_0 + 2c_6\kappa\mathbb{E}[\|X\|_{\mathcal{L}_2}]\kappa_1^{-\frac{2r+1}{4r}} + 4c_6 C_z\sqrt{\delta_0/2}\kappa\mathbb{E}[\|X\|_{\mathcal{L}_2}]\kappa_1^{-\frac{2r+1}{4r}}$. We now take $t = n(\delta_1'')^2\mu^2$. Taking $\delta_1$ and $\delta_1''$ small enough but $\kappa_1$ large enough, such that

$$D + 2\tilde{C}(\delta_1'')^2 + 2\tilde{C}D(\delta_1'')^2 + \frac{2}{3}\tilde{C}^2(\delta_1'')^2 \leq \delta_0.$$

So that

$$\mathbb{P}\Big(\sup_{g\in\mathcal{G}(R)} \big|\|g\|_n^2 - \|g\|^2\big| \geq \delta_0 R^2\Big) \leq \exp[-n(\delta_1'')^2\mu^2].$$

Thus, Lemma 4 is proved and the event $\mathcal{T}_1$ is justified. ■

To verify that the event $\mathcal{T}_2(\delta_0, R)$ occurs with high probability, we make use of some concentration results on Gaussian processes, stated in Lemma 12 of Appendix B.

**Lemma 5** *Suppose that all the conditions in Lemma 4 are satisfied, it holds*

$$\sup_{g\in\mathcal{G}(R)} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i, Z_i)\Big| \leq \delta_0 R^2,$$

*with probability at least* $1 - \exp[-n(\delta_1'')^2\mu^2]$.

### Appendix $A_2$: Optimal parametric rates For Theorem 2

**Proposition 1** *Suppose that Conditions A-D hold. We define some function* $c(\cdot)$ *of* $\delta$ *such that* $c(\delta_0) < \delta_0/(8\|f^*\|_K^2)$, *and* $D''$ *is constant appearing in our proof. For constants* $\delta_1, \delta_1''$ *and* $\kappa_1$ *with suitable choices in our proofs, we set* $\lambda_0 := \sqrt{2\log(2p)/n}$ *and*

$$\max\{\lambda_0/\delta_1, 4D''R^2\} \leq \lambda \leq \min\{1, \frac{1}{2}\sqrt{\delta_0/p_0}\Lambda_{\min}\}R,$$

$$\max\{\kappa_1 n^{-\frac{2r}{2r+1}}, c(\delta_0)R^2\} \leq \mu^2 \leq \delta_0 R^2/(8\|f^*\|_K^2).$$

*Then with probability at least* $1 - 4\exp[-n(\delta_1'')^2\mu^2] - \frac{5}{2p}$, *there holds*

$$\|\widetilde{\mathbf{Z}}^T(\widehat{\gamma} - \gamma^*)\|^2 + \lambda/8\|\widehat{\gamma} - \gamma^*\|_1 \leq p_0\lambda^2/\Lambda_{min}^2. \tag{.10}$$

**Proof of Proposition 1.** Our initial idea of the proof is the first order optimization for convex problems. Define

$$\widehat{\gamma}_j^s = \widehat{\gamma} + se_j,\ \widehat{f}_s^j = \widehat{f} - s\Pi(Z^j|\mathcal{H}_K),\ j = 1, ..., p,$$

where $e_j$ is the $j$th unit vector of $\mathcal{R}^p$. Since $(\widehat{f}, \widehat{\gamma})$ is the minimizer of the penalized least square approach in (2.1), the Karush-Kuhn-Tucker Condition is applied to yield

$$\frac{d}{ds}\Big(\frac{1}{n}\sum_{i=1}^n \big(Y_i - \langle X_i, \widehat{f}_s^j\rangle_{\mathcal{L}_2} - \mathbf{Z}_i^T\widehat{\gamma}_j^s\big)^2 + \mu^2\|\widehat{f}_s^j\|_K^2 + \lambda\|\widehat{\gamma}_j^s\|_1\Big)\Big|_{s=0} = 0,\ j = 1..., p.$$

Hence, we have

$$-\frac{1}{n}\sum_{i=1}^{n}\big(Y_i-\langle X_i,\widehat{f}\rangle_{\mathcal{L}_2}-Z_i^T\widehat{\boldsymbol{\gamma}}\big)\big(\mathbf{Z}_i^Te_j-\Pi(Z^j|X_i)\big)+\frac{\lambda}{2}\hat{\tau}_j-\mu^2\langle\Pi(Z^j|\mathcal{H}_K,\widehat{f}\rangle_K=0,$$

where $\hat{\tau}_j\in[-1,1]$ is a sub-gradient of $|\hat{\gamma}_j|$. Let $\mathbb{Z}=(\mathbf{Z}_1,...,\mathbf{Z}_n)^T$, $\Pi_{\mathbf{Z}|\mathbb{X}}=(\Pi_{\mathbf{Z}|X_1},...,\Pi_{\mathbf{Z}|X_n})^T$, $\hat{\boldsymbol{\tau}}=(\hat{\tau}_1,...,\hat{\tau}_p)$ and $\widetilde{\mathbb{Z}}:=\mathbb{Z}-\Pi_{\mathbf{Z}|\mathbb{X}}$ is an empirical matrix of $\widetilde{\mathbf{Z}}$. Also, we define a map $\Pi_{\mathbf{Z}|\mathcal{H}_K}$ from $\mathcal{H}_K$ to $\mathcal{R}^p$ by $\Pi_{\mathbf{Z}|\mathcal{H}_K}(f)=(\langle\Pi(Z^1|\mathcal{H}_K,f\rangle_K,...,\langle\Pi(Z^p|\mathcal{H}_K,f\rangle_K)$, and similarly $\mathbb{X}(f):=(\langle X_1,f\rangle_{\mathcal{L}_2},...,\langle X_n,f\rangle_{\mathcal{L}_2})^T$. Using differentiating and matrix notation, one gets:

$$-\big(\mathbf{y}-\mathbb{X}(\widehat{f})-\mathbb{Z}\widehat{\boldsymbol{\gamma}}\big)^T\widetilde{\mathbb{Z}}/n+\lambda\hat{\boldsymbol{\tau}}/2-\mu^2\Pi_{\mathbf{Z}|\mathcal{H}_K}(\widehat{f})=0.$$

Recalling the model $Y_i=\langle X_i,f^*\rangle_{\mathcal{L}_2}+\mathbf{Z}_i^T\boldsymbol{\gamma}^*+\varepsilon_i$, we have

$$\big(\mathbb{X}(\widehat{f}-f^*)+\mathbb{Z}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)-\boldsymbol{\varepsilon}\big)^T\widetilde{\mathbb{Z}}/n+\lambda\hat{\boldsymbol{\tau}}/2-\mu^2\Pi_{\mathbf{Z}|\mathcal{H}_K}(\widehat{f})=0.$$

Rearranging the above equality leads to

$$(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}/n+(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\Pi_{\mathbf{Z}|\mathbb{X}}^T\widetilde{\mathbb{Z}}/n+\mathbb{X}(\widehat{f}-f^*)^T\widetilde{\mathbb{Z}}/n-2\boldsymbol{\varepsilon}^T\widetilde{\mathbb{Z}}/n+\lambda\hat{\boldsymbol{\tau}}/2-\mu^2\Pi_{\mathbf{Z}|\mathcal{H}_K}(\widehat{f})=0.$$

Multiplying by $\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*$, we have

$$\begin{aligned}&(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\lambda/2\|\widehat{\boldsymbol{\gamma}}\|_1-\lambda/2\hat{\boldsymbol{\tau}}\boldsymbol{\gamma}^*\\&=-(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\Pi_{\mathbf{Z}|\mathbb{X}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n-\mathbb{X}(\widehat{f}-f^*)^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+2\boldsymbol{\varepsilon}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n\\&+\mu^2\Pi_{\mathbf{Z}|\mathcal{H}_K}(\widehat{f})(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*),\end{aligned}$$

where we use the equality $\|\widehat{\boldsymbol{\gamma}}\|_1=\hat{\boldsymbol{\tau}}\widehat{\boldsymbol{\gamma}}$. Note that $\hat{\boldsymbol{\tau}}\boldsymbol{\gamma}^*\leq\|\boldsymbol{\gamma}^*\|_1$ and thus $\hat{\boldsymbol{\tau}}\boldsymbol{\gamma}^*-\|\widehat{\boldsymbol{\gamma}}_{S_0}\|_1\leq\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1$. Then, we have

$$\begin{aligned}&(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\lambda/2\|\widehat{\boldsymbol{\gamma}}_{S_0^c}\|_1\\&\leq\lambda/2\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1-(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\Pi_{\mathbf{Z}|\mathbb{X}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n-\mathbb{X}(\widehat{f}-f^*)^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n\\&+2\boldsymbol{\varepsilon}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\mu^2\Pi_{\mathbf{Z}|\mathcal{H}_K}(\widehat{f})(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*).\end{aligned}\tag{.11}$$

We will separately provide upper bounds of each term on the right-hand side of (.11).

**Lemma 6** *Suppose that Condition B holds and $\varepsilon$ is a standard Gaussian variable. Then*

$$\boldsymbol{\varepsilon}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n\leq 2(C_z+C_\pi\|X\|_{\mathcal{L}_2})\sqrt{\frac{\log(2p)}{n}}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1,$$

*with probability at least $1-1/p$.*

**Lemma 7** *With the same conditions as Lemma 2. Then on event $\mathcal{T}$*

$$\mu^2\Pi_{Z|\mathcal{H}_K}(\widehat{f})(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\leq\Big(\frac{\delta_0C_h^2}{4\|f^*\|_K^2}+\frac{\delta_0}{8}+\frac{1}{2}\Big)R^2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1.$$

**Lemma 8** *Assume the conditions of lemma 1. With probability at least* $1-p\exp[-n(\delta_1'')^2\mu^2]$, *there holds*

$$\big|\mathbb{X}(\widehat{f}-f^*)^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n\big| \le c_8\delta_0\sqrt{\log p}R^2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1.$$

**Lemma 9** *Assume Conditions A-B. Then with probability at least* $1-1/p$, *there holds*

$$\big|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\Pi^T_{\mathbf{Z}|\mathbb{X}}\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n\big| \le \tilde{D}\sqrt{\delta_0/2}\sqrt{\frac{2(\log 2+3\log p)}{n}}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1^2,$$

*where* $\tilde{D} := C_\pi(C_z+\kappa C_h)$.

**Lemma 10** *Assume that Condition B holds. With probability at least* $1-\frac{1}{2p}$, *there holds*

$$|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n - \|\widetilde{\mathbf{Z}}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2| \le 20(C_z+C_\pi)^2\sqrt{\log(2p)/n}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1^2.$$

Using Lemma 6–Lemma 9 stated as above, we conclude from (.11) that

$$\begin{aligned}&(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\lambda/2\|\widehat{\boldsymbol{\gamma}}_{S_0^c}\|_1\\ &\le \lambda/2\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1 + D''\Big(\sqrt{\frac{\log(2p)}{n}}+R^2+\sqrt{\frac{\log 2}{n}}\Big)\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1, \qquad (.12)\end{aligned}$$

where we use the conclusion in Lemma 2 (e.g. $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 \le \sqrt{\delta_0/2}R^2/\lambda$) and $D'' := 8(C_z+C_\pi\|X\|_{\mathcal{L}_2})+\frac{\delta_0C_h^2}{2\|f^*\|_K^2}+\frac{\delta_0}{4}+1+2c_8\delta_0\sqrt{\log p}+\tilde{D}\sqrt{2\delta_0}$. Adding $\lambda/2\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1$ to both sides of (.12) again, we easily obtain

$$\begin{aligned}&(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\lambda/2\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1\\ \le\ &\lambda\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1 + D''\big(\sqrt{\log(2p)/n}+R^2+\sqrt{\log 2/n}\big)\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1. \qquad (.13)\end{aligned}$$

Provided that $\lambda \ge 4D''\Big(2\sqrt{\frac{\log(2p)}{n}}+R^2\Big)$ is satisfied, (.13) can be simplified to be

$$\begin{aligned}&(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\lambda/4\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1\\ &\le \lambda\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)_{S_0}\|_1 \le \lambda\sqrt{p_0}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_2\\ &\le \lambda\sqrt{p_0}/\Lambda_{min}\|\widetilde{\mathbf{Z}}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\| \le \lambda^2p_0/(2\Lambda_{min}^2)+\frac{1}{2}\|\widetilde{\mathbf{Z}}^T(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)\|^2\\ &\le \lambda^2p_0/(2\Lambda_{min}^2)+\frac{1}{2}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+E''\sqrt{\log(2p)/n}\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1^2, \qquad (.14)\end{aligned}$$

where $E'' := 10(C_z+C_\pi)^2$ for brevity and the third inequality is based on Condition A and the last inequality follows from Lemma 10. Using $\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 \le \sqrt{\delta_0/2}R^2/\lambda$ again, and when $E''\sqrt{\log(2p)/n}\sqrt{\delta_0/2}R^2 \le \lambda^2/8$, from (.14) we conclude that

$$(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*)/n+\lambda/4\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^*\|_1 \le \lambda^2p_0/\Lambda_{min}^2. \qquad (.15)$$

Finally, by Lemma 10 again, rearranging the above inequality and parameters constraints yields our desired results. ■

## Appendix B: Proof for lower bounds

We now turn to the proof of the minimax lower bounds presented in Theorem 3. To this end, we need to introduce a useful lemma from Theorem 2.5 in Tsybakov (2009), that gives a lower bound based on Kullkack divergences.

**Proof of Theorem 3.** Note that any lower bound for a specific case yields immediately a lower bound for the general case. Thus, it is easy to see that the minimax lower bound for estimating $\boldsymbol{\gamma}^*$ is trivial based on the existing results for high dimensional linear models derived from Verzelen (2012) that

$$R_{\boldsymbol{\gamma}^*}(p_0, p, \mathcal{H}_K(1)) \geq \inf_{\hat{\boldsymbol{\gamma}}} \sup_{\boldsymbol{\gamma}^* \in B[p_0,p]} \mathbb{E}[\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2] = \Omega\big(\frac{p_0}{n} \log\big(\frac{p}{p_0}\big)\big).$$

In the following, we focus on the lower bound of the minimax risk for the prediction risk of the nonparametric component. Recall the partial linear functional model

$$Y = \langle X, f^* \rangle_{\mathcal{L}_2} + \mathbf{Z}^T \boldsymbol{\gamma}^* + \varepsilon. \tag{.16}$$

As above, we remove the sup-mum of $\boldsymbol{\gamma}^*$ in $R_{f^*}(p_0, p, \mathcal{H}_K)$ and obtain

$$R_{f^*}(p_0, p, \mathcal{H}_K) \geq \inf_{\widehat{f}} \sup_{f^* \in \mathcal{H}_K(1)} \mathbb{E}[\langle X, \widehat{f} - f^* \rangle_{\mathcal{L}_2}^2] = n^{-\frac{2r}{2r+1}},$$

where the lower bound of the prediction risk for the linear functional model has been established in Cai and Yuan (2012).

The remaining task is to establish the second part of the minimax lower bound, i.e., $\delta_n := p_0/n \log(p/p_0)$. To attain this lower bound, it suffices to consider the specific case of $\delta_n \geq n^{-\frac{2r}{2r+1}}$. This implies that $p$ goes to infinity as $n$ increases, which will be used in our proof.

For some $\boldsymbol{\theta} = (\theta_{M+1}, ..., \theta_{2M}) \in \{0,1\}^M := \Theta_1$. The Varshamov-Gibert bound shows that for any $M \geq 8$, there exists a set $\Theta_1 = \{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N_1)}\} \in \{0,1\}^M$ such that
(a) $\boldsymbol{\theta}^{(0)} = (0, ..., 0)^T$;
(b) $H(\boldsymbol{\theta}, \boldsymbol{\theta}') > M/8$ for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \Theta_1$, where $H(\cdot,\cdot)$ denotes the Hamming distance;
(c) $N_1 \geq 2^{M/8}$.

We now employ the results from Tsybakov (2009) to establish the lower bound that is based upon testing multiple hypotheses.

**Lemma 11** *Assume that $N \geq 2$ and suppose that $\Theta$ with some pseudometric $d$ contains elements $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(N)}$ such that:*
*(i) $d(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(k)}) \geq 2s > 0, \quad \forall\, 0 \leq j \leq k \leq N$;*
*(ii) $P_j \ll P_0$, $\forall\, j = 1, 2, ..., N$, and*

$$\frac{1}{N} \sum_{j=1}^{N} \mathcal{D}_{KL}(P_j, P_0) \leq \alpha \log N$$

*with $0 < \alpha < 1/8$ and $P_j = P_{\boldsymbol{\theta}^{(j)}}$, $j = 0, 1, ..., N$. Then*

$$\inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \Theta} P_{\boldsymbol{\theta}}\big(d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \geq s\big) \geq \frac{\sqrt{N}}{1 + \sqrt{N}}\Big(1 - 2\alpha - \frac{2\alpha}{\log N}\Big) > 0.$$

Fix $\alpha \in (0, 1/8)$. In order to apply Lemma 11, we need to check the following three conditions:
(i) $f_{\boldsymbol{\theta}^{(j)}} \in \mathcal{H}_K(1), \quad j = 0, 1, ..., N,$
(ii) $d(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(k)}) \geq 2s > 0, \ , 0 \leq j \leq k \leq N,$
(iii) $\frac{1}{N} \sum_{j=1}^{N} \mathcal{D}_{KL}(P_j, P_0) \leq \alpha \log N.$

We will now show that these conditions are satisfied for all sufficiently large $n$. Before this, we first need to define a pseudometric between pairs $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\gamma}_1, f_1)$ and $\boldsymbol{\theta}^{(2)} = (\boldsymbol{\gamma}_2, f_2)$ as the $\mathcal{L}_2$-distance between $f_1$ and $f_2$ works out well. To this end, we define the pseudometric $d(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) := d_1(f_1, f_2)$, where $d_1(f_1, f_2) = \|L_{C^{1/2}}(f_1 - f_2)\|_{\mathcal{L}_2}$. It is easy to verify that all of the metric properties are satisfied for $d(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ except that, of course, it is possible to have $d(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) = 0$ while $\boldsymbol{\theta}^{(1)} \neq \boldsymbol{\theta}^{(2)}$. Obviously, $d$ is qualified as a pseudo-metric. In this case, we define $\Theta = (\boldsymbol{\gamma}, f)$ for all $\boldsymbol{\gamma} \in B[p_0, p]$ and $f \in \mathcal{H}_K(1)$. Then we have

$$\inf_{\widehat{f}} \sup_{\boldsymbol{\gamma} \in B[p_0,p], f^* \in \mathcal{H}_K(1)} \mathbb{E}[\langle X, \widehat{f} - f^* \rangle_{\mathcal{L}_2}^2] \geq \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \Theta} d(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})^2.$$

Our second definition is to construct an appropriate finite subset of $\Theta$. To this end, we start with constructing a set of test functions with the form

$$f_{\boldsymbol{\theta}} = M^{-1/2} \sum_{k=M+1}^{2M} \theta_k L_{K^{1/2}} \varphi_k, \quad \boldsymbol{\theta} = (\theta_{M+1}, ..., \theta_{2M})^T \in \{0, 1\}^M.$$

Since $L_{K^{1/2}}(\varphi_k) \in \mathcal{H}_K$ for all $k \in \mathbb{N}$, this implies that $f_{\boldsymbol{\theta}} \in \mathcal{H}_K$. Define a finite subset of $\Theta$ that consists of $\bar{\boldsymbol{\theta}}^{(j)} = (\boldsymbol{\gamma}_j, f_{\boldsymbol{\theta}^{(j)}})$ where $\boldsymbol{\gamma}_j \in B[p_0, p]$ is arbitrary for $0 \leq j \leq N$ and particularly $\bar{\boldsymbol{\theta}}^{(0)} = (0, 0, ..., 0)$. In the following, the second part of $\Theta$ is denoted by $\Theta_1$.

First of all, we will verify condition $(i)$. Note that $\langle L_{K^{1/2}} f, L_{K^{1/2}} g \rangle_K = \langle f, g \rangle_{\mathcal{L}_2}$ for any $f, g \in \mathcal{L}_2$, by orthogonality and $|\theta_k| \leq 1$ we have

$$\|f_{\boldsymbol{\theta}}\|_K^2 = M^{-1} \sum_{k=M+1}^{2M} \theta_k^2 \big\|L_{K^{1/2}} \varphi_k\big\|_K^2 \leq M^{-1} \sum_{k=M+1}^{2M} \big\|\varphi_k\big\|_{\mathcal{L}_2}^2 = 1.$$

So this verifies condition $(i)$. Next, we denote by $P_{\bar{\boldsymbol{\theta}}}$ the joint normal distribution of $\{(Y_i, X_i, \mathbf{Z}_i), i \geq 1\}$ with the conditional mean $f_{\boldsymbol{\theta}}(T_i) + \mathbf{Z}_i^T \boldsymbol{\gamma}$. For any $\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}' \in \Theta$, denote by

$H(\boldsymbol{\theta}, \boldsymbol{\theta}')$ the Hamming distance on $\Theta_1$. Note that

$$\begin{aligned}
d^2(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}') &= \|L_{C^{1/2}}(f_{\boldsymbol{\theta}} - f_{\boldsymbol{\theta}'})\|_{\mathcal{L}_2}^2 \\
&= \left\| M^{-1/2} \sum_{k=M+1}^{2M} (\theta_k - \theta_k') L_{C^{1/2}} L_{K^{1/2}} \varphi_k \right\|_{\mathcal{L}_2}^2 \\
&= M^{-1} \sum_{k=M+1}^{2M} (\theta_k - \theta_k')^2 \left\| L_{C^{1/2}} L_{K^{1/2}} \varphi_k \right\|_{\mathcal{L}_2}^2 \\
&= M^{-1} \sum_{k=M+1}^{2M} (\theta_k - \theta_k')^2 s_k \\
&\geq M^{-1} s_{2M} \sum_{k=M+1}^{2M} (\theta_k - \theta_k')^2 \\
&= M^{-1} s_{2M} H(\boldsymbol{\theta}, \boldsymbol{\theta}').
\end{aligned}$$

Besides, by Varshamov-Gibert bound and the entropy assumption (Condition D), we further have

$$d(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}') \geq \sqrt{s_{2M}/8} \geq c_1 2^{-(r+2)} M^{-r}.$$

Here we take $M = \lfloor \delta_n^{-1/(2r)} \rfloor$, leading to

$$d(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}') \geq c_1 2^{-(r+2)} \sqrt{\delta_n} := 2s.$$

Thus Condition (ii) is verified. Finally, for any $\bar{\boldsymbol{\theta}} \in \Theta$, we notice that

$$\begin{aligned}
\log(P_{\bar{\boldsymbol{\theta}}}/P_{\bar{\boldsymbol{\theta}}^{(0)}}) = &\sum_{i=1}^n \big(Y_i - \langle X_i, f_{\boldsymbol{\theta}} \rangle_{\mathcal{L}_2} - \mathbf{Z}_i^T \boldsymbol{\gamma}\big)\big(\langle X_i, f_{\boldsymbol{\theta}} \rangle_{\mathcal{L}_2}\big) \\
&- \frac{1}{2} \sum_{i=1}^n \big[\langle X_i, f_{\boldsymbol{\theta}} \rangle_{\mathcal{L}_2}\big]^2.
\end{aligned}$$

Hence, the Kullback-Leibler distance between $P_{\bar{\boldsymbol{\theta}}}$ and $P_{\bar{\boldsymbol{\theta}}^{(0)}}$ can be expressed by

$$\mathcal{D}_{KL}(P_{\bar{\boldsymbol{\theta}}}, P_{\bar{\boldsymbol{\theta}}^{(0)}}) = \int \log(P_{\bar{\boldsymbol{\theta}}}/P_{\bar{\boldsymbol{\theta}}^{(0)}}) dP_{\bar{\boldsymbol{\theta}}} = n\big(\|L_{C^{1/2}}(f_{\boldsymbol{\theta}})\|_{\mathcal{L}_2}^2\big).$$

Note that

$$\langle L_{C^{1/2}} L_{K^{1/2}} \varphi_k, L_{C^{1/2}} L_{K^{1/2}} \varphi_j \rangle_{\mathcal{L}_2} = \langle \varphi_k, L_{K^{1/2}} L_C L_{K^{1/2}} \varphi_j \rangle_{\mathcal{L}_2} = \langle \varphi_k, s_j \varphi_j \rangle_{\mathcal{L}_2} = s_j \delta_{kj}.$$

A similar argument as above leads to

$$\|L_{C^{1/2}}(f_{\boldsymbol{\theta}})\|_{\mathcal{L}_2}^2 \leq \frac{s_M}{M} \sum_{k=M+1}^{2M} (\theta_k)^2 = \frac{s_M}{M} H(\boldsymbol{\theta}, \boldsymbol{\theta}') \leq c_2 M^{-2r},$$

where we use Condition D and the fact that $s_\ell$ is a decreasing sequence of $\ell$. Then, we claim that

$$\frac{1}{N}\sum_{j=1}^{N}\mathcal{D}_{KL}(P_j,P_0)\le nc_2M^{-2r}\le \alpha\log N,$$

where $N := N_1\binom{p}{p_0}$ deriving from Conclusion $(c)$ of the Varshamov-Gibert bound and all the possible sparse cases for the parametric part. Indeed, it is easy to check that

$$\log N\ge c_3\big(M+(p-p_0)\log(p/p_0)\big)\ge c_2nM^{-2r},$$

with the choice of $M$ given as above when $\delta_n\ge n^{-\frac{2r}{2r+1}}$. Here we also use the relation $\log\binom{p}{p_0}\simeq(p-p_0)\log(p/p_0)$ when $p$ is diverging and $p\gg p_0$. This means that we verify Condition (iii).

As a consequence, an application of Lemma 11 yields that

$$\inf_{\widehat{f}}\sup_{\boldsymbol{\gamma}\in B[p_0,p],f^*\in\mathcal{H}_K(1)}P_{\bar{\theta}}\Big(\|L_{C^{1/2}}(\widehat{f}-f^0)\|_{\mathcal{L}_2}\ge c\sqrt{\frac{p_0\log(p/p_0)}{n}}\Big)\ge\frac{\sqrt{N}}{1+\sqrt{N}}\Big(1-2\alpha-\frac{2\alpha}{\log N}\Big).$$

As $n$ goes to infinity, so are $M,N$. This means that there exist sufficiently large $n$ while a suitable choice of $\alpha$ (e.g. $\alpha=1/10$), such that $\frac{\sqrt{N}}{1+\sqrt{N}}\Big(1-2\alpha-\frac{2\alpha}{\log N}\Big)>9/10-3\alpha>0$. ■

## Appendix C: Some useful Lemmas

The Gaussian concentration inequality from Theorem 7.1 of Ledoux (2001) is a useful tool in our refined analysis, which provides tighter bounds than the general sub-Gaussian cases. In particular, the super-norm bounds of random variables are not needed, as opposed to Rademacher concentration inequality presented in Lemma 3 as above.

**Lemma 12** *Let $\mathbb{G}=\{G_t\}_{t\in T}$ be a centered Gaussian process indexed by a countable set $T$ such that $\sup_{t\in T}G_t<\infty$ almost surely. Then*

$$\mathbb{P}\Big(\sup_{t\in T}G_t\ge\mathbb{E}[\sup_{t\in T}G_t]+r\Big)\le\exp(-\frac{r^2}{2\sigma^2}),$$

*where $\sigma^2=\sup_{t\in T}\mathbb{E}[G_t^2]<\infty$.*

**Proof of Lemma 5.** Note that for any $g\in\mathcal{G}(R)$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i,\mathbf{Z}_i)\Big|\le\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle X,f\rangle_{\mathcal{L}_2}\Big|+\sup_j\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i z_{ij}\Big|\|\boldsymbol{\gamma}\|_1.$$

On one hand, we conclude from Lemma 14 that

$$\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle X_i,f\rangle_{\mathcal{L}_2}\Big|\Big)\le\frac{R}{\mu}\mathbb{E}[\|G_n\|_{\mathcal{B}(\tilde{\delta})}]\le\frac{\sqrt{2}R}{\mu}\gamma_n(\tilde{\delta}),$$

where $\tilde{\delta} := \big(1+\frac{\Lambda_{\max}}{\Lambda_{\min}}\big)\mu$. In addition, since $\varepsilon_i$'s are standard Gaussian variables and $|Z_i^j| \leq C_z$ by Condition B, Bernstein inequality is applied to yield

$$\mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\sigma_i\Big\|_{\infty} \leq \lambda_0 C_z.$$

Thus, using similar arguments to (.7) and (.8) yields that

$$\mathbb{E}\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i,\mathbf{Z}_i)\Big| \leq \frac{\sqrt{2}R}{\mu}\gamma_n(\tilde{\delta}) + C_z\lambda_0\frac{\sqrt{\delta_0/2}R^2}{\lambda} \leq (c_7\kappa_1^{-\frac{2r+1}{4r}} + C_z\sqrt{\delta_0/2}\delta_1)R^2,$$

which follows from $\|\boldsymbol{\gamma}\|_1 \leq \frac{\sqrt{\delta_0/2}R^2}{\lambda}$ and the derived inequality appearing in (.18). Observe that $\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i,\mathbf{Z}_i)\Big|$ is a centered Gaussian process, and also check that $\sigma^2 \leq \frac{1}{n}R^2$ in Lemma 12. Then, by the Gaussian concentration inequality with $r = 2(\delta_1'')^2\mu^2R^2$, we have

$$\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g(X_i,\mathbf{Z}_i)\Big| \leq 2(\delta_1'')^2\mu^2R^2 + (c_7\kappa_1^{-\frac{2r+1}{4r}} + C_z\sqrt{\delta_0/2}\delta_1)R^2,$$

As long as $\delta_1, \delta_1''$ are small sufficiently and $\kappa_1$ is properly large, we can obtain the desired result. ■

Due to the functional style of $X$, we consider the following Rademacher type of process:

$$R_n(h) = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, h\rangle_{\mathcal{L}_2}\sigma_i.$$

Then, we define the function set

$$\mathcal{B}(\delta) = \{h \in \mathcal{H}_K : \|h\|_K \leq 1 \,\text{and}\, \|L_{C^{1/2}}h\|_{\mathcal{L}_2} \leq \delta\},$$

and the norm

$$\|R_n\|_{\mathcal{B}(\delta)} = \sup_{h\in\mathcal{B}(\delta)}|R_n(h)|.$$

The following lemma, from Lemma 5 of Cai and Yuan (2012), gives some lower bound and an upper bound of the Rademacher complexity $R_n$ over $\mathcal{B}$.

**Lemma 13** *Let $\gamma_n(\delta) := \Big(\frac{1}{n}\sum_{\ell\geq 1}\min\{s_\ell,\delta^2\}\Big)^{1/2}$ for any $\delta > 0$ and assume that Condition C holds true. Then there exist constants $c_3, c_4, c_5 > 0$ such that*

$$c_3\gamma_n(\delta) - c_4 n^{-1}(\log n) \leq \mathbb{E}[\|R_n\|_{\mathcal{B}(\delta)}] \leq c_5\gamma_n(\delta).$$

For any $g(\mathbf{U}) = \langle X, f\rangle_{\mathcal{L}_2} + \mathbf{Z}^T\boldsymbol{\gamma} \in \mathcal{G}(R)$, we have $\|f\|_K \leq \frac{R}{\mu}$ as discussed before. Besides, we also get, $\|\widetilde{\mathbf{Z}}^T\boldsymbol{\gamma}\|^2 + \|\langle X, f\rangle_{\mathcal{L}_2} + \Pi_{\mathbf{Z}|X}^T\boldsymbol{\gamma}\|^2 \leq R^2$. Thus, $\|\boldsymbol{\gamma}\|_2^2 \leq R^2/\Lambda_{\min}^2$, so that $\|\Pi_{\mathbf{Z}|X}^T\boldsymbol{\gamma}\|^2 \leq R^2\Lambda_{\max}^2/\Lambda_{\min}^2$. Hence,

$$\|L_{C^{1/2}}f\|_{\mathcal{L}_2} = \big\|\langle X, f\rangle_{\mathcal{L}_2}\big\| \leq \|\langle X, f\rangle_{\mathcal{L}_2} + \Pi_{\mathbf{Z}|X}^T\boldsymbol{\gamma}\| + \|\Pi_{\mathbf{Z}|X}^T\boldsymbol{\gamma}\| \leq \big(1 + \frac{\Lambda_{\max}}{\Lambda_{\min}}\big)R. \qquad (.17)$$

Thus, we conclude from Lemma 13 that

$$\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\langle X_i,f\rangle_{\mathcal{L}_2}\sigma_i\Big|\Big)\leq\frac{R}{\mu}\mathbb{E}[\|R_n\|_{\mathcal{B}(\tilde{\delta})}]\leq c_5\frac{R}{\mu}\gamma_n(\tilde{\delta}),$$

where $\tilde{\delta}:=\big(1+\frac{\Lambda_{\max}}{\Lambda_{\min}}\big)\mu$. By direct calculation, it follows from Condition D that

$$\gamma_n^2(\delta)\asymp\frac{1}{n}\delta^{2-\frac{1}{r}},\quad\forall\,\delta>0.$$

This along with (.8) implies that, there exists some constant $c_6$ such that

$$\mathbb{E}\Big(\sup_{g\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^{n}\langle X_i,f\rangle_{\mathcal{L}_2}^2\sigma_i\Big|\Big)\leq c_6\kappa\|X\|_{\mathcal{L}_2}\frac{R^2}{\sqrt{n}\mu^{\frac{2r+1}{2r}}}\leq c_6\kappa\|X\|_{\mathcal{L}_2}\kappa_1^{-\frac{2r+1}{4r}}R^2,\tag{.18}$$

where we used the assumption that $\mu^2\geq\kappa_1 n^{-\frac{2r}{2r+1}}$.

We now consider another functional complexity involving the Gaussian variables. To this end, we define Gaussian complexity with the alignment of two kernels:

$$G_n(h):=\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\langle X_i,h\rangle_{\mathcal{L}_2},\quad\forall\,h\in\mathcal{B}(\delta),$$

and

$$\|G_n\|_{\mathcal{B}(\delta)}=\sup_{h\in\mathcal{B}(\delta)}|G_n(h)|.$$

**Lemma 14** *Let $\gamma_n(\delta)$ be defined as that in Lemma 13. For any $\delta>0$, we have*

$$\mathbb{E}[\|G_n\|_{\mathcal{B}(\delta)}]\leq\sqrt{2}\gamma_n(\delta).$$

**Proof of Lemma 14.** For brevity, we define $T=L_{K^{1/2}CK^{1/2}}$ in the following. It is obvious that $\mathcal{B}(\delta)=L_{K^{1/2}}(\mathcal{H}_K(\delta))$, where

$$\mathcal{H}_K(\delta)=\big\{h\in\mathcal{L}_2:\ \|h\|_{\mathcal{L}_2}\leq 1\,\text{and}\,\|T^{1/2}h\|_{\mathcal{L}_2}\leq\delta^2\big\}.$$

Denote

$$\mathcal{G}=\Big\{\sum_{\ell\geq1}\alpha_\ell\varphi_\ell:\sum_{\ell\geq1}\Big(\frac{\alpha_\ell}{\min\{1,\delta/\sqrt{s_\ell}\}}\Big)^2\leq1\Big\}$$

It can be readily shown that $\mathcal{G}\subset\mathcal{H}_K(\delta)\subset\sqrt{2}\mathcal{G}$. This immediately implies that

$$\sup_{h\in\mathcal{G}}|G_n(L_{K^{1/2}}h)|\leq\|G_n\|_{\mathcal{B}(\delta)}\leq\sqrt{2}\sup_{h\in\mathcal{G}}|G_n(L_{K^{1/2}}h)|.$$

By Jensen's inequality, we have

$$\mathbb{E}\sup_{h\in\mathcal{G}}|G_n(L_{K^{1/2}}h)|\leq\Big(\mathbb{E}\sup_{h\in\mathcal{G}}|G_n(L_{K^{1/2}}h)|^2\Big)^{1/2}.$$

By Cauchy-Schwartz inequality, for any $h \in \mathcal{G}$

$$
\begin{aligned}
|G_n(L_{K^{1/2}}h)|^2 &= \Big|\sum_{\ell \geq 1} \alpha_\ell G_n(L_{K^{1/2}}\varphi_\ell)\Big|^2 \\
&\leq \Big(\sum_{\ell \geq 1}\Big(\frac{\alpha_\ell^2}{\min\{1, \delta^2/s_\ell\}}\Big)\Big(\sum_{\ell \geq 1} \min\{1, \delta^2/s_\ell\} G_n^2(L_{K^{1/2}}\varphi_\ell)\Big)
\end{aligned}
$$

which implies that

$$
\sup_{h \in \mathcal{G}} |G_n(L_{K^{1/2}}h)|^2 \leq \sum_{\ell \geq 1} \min\{1, \delta^2/s_\ell\} G_n^2(L_{K^{1/2}}\varphi_\ell).
$$

Note that

$$
\mathbb{E}G_n^2(L_{K^{1/2}}\varphi_\ell) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle X_i, L_{K^{1/2}}\varphi_\ell\rangle_{\mathcal{L}_2}\Big)^2 = \frac{1}{n}\mathbb{E}\langle X_i, L_{K^{1/2}}\varphi_\ell\rangle_{\mathcal{L}_2}^2 = n^{-1}s_\ell.
$$

Thus, we have

$$
\mathbb{E}\sup_{h \in \mathcal{G}} |G_n(L_{K^{1/2}}h)|^2 \leq \sum_{\ell \geq 1} \min\{1, \delta^2/s_\ell\} \mathbb{E}G_n^2(L_{K^{1/2}}\varphi_\ell) = \gamma_n^2(\delta),
$$

which further implies that

$$
\mathbb{E}\|G_n\|_{\mathcal{B}(\delta)} \leq \sqrt{2}\gamma_n(\delta).
$$

This completes the proof. ■

**Proof of Lemma 6.** First of all, we notice that

$$
\boldsymbol{\varepsilon}^T \widetilde{\mathbb{Z}}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)/n \leq \|\boldsymbol{\varepsilon}^T \widetilde{\mathbb{Z}}/n\|_\infty \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1.
$$

When $\widetilde{\mathbb{Z}}$ is fixed, $\boldsymbol{\varepsilon}^T \widetilde{\mathbb{Z}}/n$ is Gaussian. Then for all $r > 0$ and all $j$,

$$
P\Big(|\boldsymbol{\varepsilon}^T \widetilde{\mathbb{Z}}_j/n| \geq \sqrt{\frac{2r}{n}}\|\widetilde{\mathbb{Z}}_j\|_n\Big) \leq 2\exp(-r).
$$

Hence, by the union bound we have

$$
P\Big(\|\boldsymbol{\varepsilon}^T \widetilde{\mathbb{Z}}/n\|_\infty \geq \sqrt{\frac{2(r + \log p)}{n}} \max_{1 \leq j \leq p} \|\widetilde{\mathbb{Z}}_j\|_n\Big) \leq 2\exp(-r).
$$

Moreover, by Condition B, there always holds

$$
\|\widetilde{\mathbb{Z}}_j\|_n \leq |Z_j| + \|X\|_{\mathcal{L}_2}\|\Pi(Z_j|X)\|_{\mathcal{L}_2} \leq C_z + C_\pi \|X\|_{\mathcal{L}_2}, \quad \forall\, j = 1, ..., p.
$$

By the use of total probability principle and taking $r = \log(2p)$, we obtain the desired result. ■

**Proof of Lemma 7.** Note that, it follows from Cauchy-Schwartz inequality that

$$
\begin{aligned}
\mu^2 \Pi_{\mathbf{Z}|\mathcal{H}_K}(\widehat{f})(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) &\leq \mu^2 \max_{1 \leq j \leq p} |\langle \Pi(Z_j|\mathcal{H}_K), \widehat{f}\rangle_K| \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1 \\
&\leq \mu^2/2\big(\|f^*\|_K^2 + 2\|\Pi(Z_j|\mathcal{H}_K)\|_K^2 + \|\widehat{f} - f^*\|_K^2\big)\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_1.
\end{aligned}
$$

Moreover, by Lemma 2, on event $\mathcal{T}$, $\mu^2\|\widehat{f}-f^*\|_K^2 \le R^2$. With the choice of $\mu^2 \le \frac{\delta_0 R^2}{8\|f^*\|_K^2}$ in Lemma 2, we obtain the desired result. ■

**Proof of Lemma 8.** Note that

$$\big|\mathbb{X}(\widehat{f}-f^*)^T\widetilde{\mathbb{Z}}(\widehat{\gamma}-\gamma^*)/n\big| \le \max_{1\le j\le p}\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{Z}_{ij}\langle X_i, \widehat{f}-f^*\rangle_{\mathcal{L}_2}\Big|\|\widehat{\gamma}-\gamma^*\|_1.$$

We now bound the term $\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{Z}_{ij}\langle X_i, \widehat{f}-f^*\rangle_{\mathcal{L}_2}\Big|$ using by concentration inequality. By the definition of projection vector $\widetilde{\mathbf{Z}}$, there holds $\mathbb{E}[\widetilde{Z}^j\langle X, f\rangle_{\mathcal{L}_2}] = 0$ for any $f \in \mathcal{H}_K$. So we define $\mathcal{S} := \sup_{f\in\mathcal{G}(R)}\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{Z}_{ij}\langle X_i, f-f^*\rangle_{\mathcal{L}_2}\Big|$, dropping off the dependence of $j$. To apply Lemma 3, we define $h(U) = \widetilde{Z}_j\langle X, f\rangle_{\mathcal{L}_2}$ with $g = \langle X, f\rangle_{\mathcal{L}_2} + \mathbf{Z}^T\gamma \in \mathcal{G}(R)$. As before, we can obtain

$$\|h\|_\infty \le \kappa(C_z + C_\pi\|X\|_{\mathcal{L}_2})\|X\|_{\mathcal{L}_2}R/\mu \le \kappa(C_z + C_\pi\|X\|_{\mathcal{L}_2})\|X\|_{\mathcal{L}_2}/\sqrt{c(\delta_0)},$$

with the choice of $\mu \ge \sqrt{c(\delta_0)}R$. Besides, it follows from (.17) that

$$var(h(U)) \le \mathbb{E}[h^2(U)] \le (C_z + C_\pi\|X\|_{\mathcal{L}_2})^2\mathbb{E}[\langle X, f\rangle_{\mathcal{L}_2}^2] \le (C_z + C_\pi\|X\|_{\mathcal{L}_2})^2\big(1+\frac{\Lambda_{\max}}{\Lambda_{\min}}\big)^2R^2.$$

In addition, recall $\tilde{\delta} = \big(1+\frac{\Lambda_{\max}}{\Lambda_{\min}}\big)\mu$, we also obtain

$$\mathbb{E}[\boldsymbol{Z}] \le 2(C_z + C_\pi\|X\|_{\mathcal{L}_2})\frac{R}{\mu}\mathbb{E}[\|G_n\|_{\mathcal{B}(\tilde{\delta})}],$$

where we use the symmetrization technique and the contraction property of Rademacher complexity. By similar arguments between (.17) and (.18), we further get

$$\mathbb{E}[\boldsymbol{Z}] \le c_7\kappa_1^{-\frac{2r+1}{4r}}\delta_0 R^2.$$

By the concentration result presented in Lemma 3 and a simple calculation, we obtain from the union bound

$$\max_{1\le j\le p}\Big|\frac{1}{n}\sum_{i=1}^n \widetilde{Z}_{ij}\langle X_i, \widehat{f}-f^*\rangle_{\mathcal{L}_2}\Big| \le c_8\delta_0\sqrt{\log p}R^2,$$

with probability at least $1-\exp[-n(\delta_1'')^2\mu^2]$. This completes the proof. ■

**Proof of Lemma 9**. Observe that

$$(\widehat{\gamma}-\gamma^*)^T\Pi_{\mathbf{Z}|\mathbb{X}}^T\widetilde{\mathbb{Z}}(\widehat{\gamma}-\gamma^*)/n = (\widehat{\gamma}-\gamma^*)^T\Big[\frac{1}{n}\sum_{i=1}^n \Pi_{\mathbf{Z}|X_i}(Z_i-\Pi_{\mathbf{Z}|X_i})^T\Big](\widehat{\gamma}-\gamma^*).$$

This immediately implies that

$$\big|(\widehat{\gamma}-\gamma^*)^T\Pi_{\mathbf{Z}|\mathbb{X}}^T\widetilde{\mathbb{Z}}(\widehat{\gamma}-\gamma^*)/n\big| \le \max_{1\le j,k\le p}\Big|\frac{1}{n}\sum_{i=1}^n \Pi(Z_j|X_i)(Z_{ik}-\Pi(Z_k|X_i))\Big|\|\widehat{\gamma}-\gamma^*\|_1^2.$$

For any given $j, k$, since $\Pi(Z_j|X_i)(Z_{ik} - \Pi(Z_k|X_i))$'s are i.i.d. centered random variables from the definition of protection, and by Condition B there holds:

$$\big|\Pi(Z_j|X)(Z_k - \Pi(Z_k|X))\big| \le C_\pi(C_z + \kappa C_h) = \tilde{D}.$$

Then, by the Hoeffding inequality, we have

$$P\Big(\Big|\frac{1}{n}\sum_{i=1}^n \Pi(Z_j|X_i)(Z_{ik} - \Pi(Z_k|X_i))\Big| > r\Big) \le 2\exp(-\frac{nr^2}{2\tilde{D}^2}),$$

which together with the union bound implies

$$\max_{1\le j,k\le p}\Big|\frac{1}{n}\sum_{i=1}^n \Pi(Z_j|X_i)(Z_{ik} - \Pi(Z_k|X_i))\Big| \le \tilde{D}\sqrt{\frac{2(\log 2 + 3\log p)}{n}},$$

with probability at least $1 - 1/p$, where setting $r := \tilde{D}\sqrt{\frac{2(\log 2+3\log p)}{n}}$. The proof ends with the conclusion of Lemma 2. ■

**Proof of Lemma 10.** By Lemma 14.14 in Bühlmann and Van. de. Geer (2011), it follows that

$$\mathbb{E}\max_{j,k}\Big|\frac{1}{n}\sum_{i=1}^n(\widetilde{Z}_{ij}\widetilde{Z}_{ik} - \mathbb{E}[\widetilde{Z}_{ij}\widetilde{Z}_{ik}])\Big| \le 8\sqrt{\log(2p)/n}(C_z + C_\pi)^2,$$

and by Massart's inequality for all $r$, it holds

$$\mathbb{P}\Big(\max_{j,k}\Big|\frac{1}{n}\sum_{i=1}^n(\widetilde{Z}_{ij}\widetilde{Z}_{ik} - \mathbb{E}[\widetilde{Z}_{ij}\widetilde{Z}_{ik}])\Big| \ge 8(C_z + C_\pi)^2\big[\sqrt{\log(2p)/n} + \sqrt{2r/n}\big]\Big) \le \exp(-r).$$

By taking $r = \log(2p)$, we have

$$\mathbb{P}\Big(\max_{j,k}\Big|\frac{1}{n}\sum_{i=1}^n(\widetilde{Z}_{ij}\widetilde{Z}_{ik} - \mathbb{E}[\widetilde{Z}_{ij}\widetilde{Z}_{ik}])\Big| \ge 20(C_z + C_\pi)^2\sqrt{\log(2p)/n}\Big) \le 1/(2p).$$

Finally, we observe that

$$\begin{aligned}
&(\widehat{\boldsymbol\gamma} - \boldsymbol\gamma^*)^T\widetilde{\mathbb{Z}}^T\widetilde{\mathbb{Z}}(\widehat{\boldsymbol\gamma} - \boldsymbol\gamma^*)/n - \|\widetilde{\mathbf{Z}}^T(\widehat{\boldsymbol\gamma} - \boldsymbol\gamma^*)\|^2 \\
&= \sum_{j,k}(\widehat{\gamma}_j - \gamma_j^0)\Big(\frac{1}{n}\sum_{i=1}^n(\widetilde{Z}_{ij}\widetilde{Z}_{ik} - \mathbb{E}[\widetilde{Z}_{ij}\widetilde{Z}_{ik}])\Big)(\gamma_k - \gamma_k^0) \\
&\le \max_{j,k}\Big|\frac{1}{n}\sum_{i=1}^n(\widetilde{Z}_{ij}\widetilde{Z}_{ik} - \mathbb{E}[\widetilde{Z}_{ij}\widetilde{Z}_{ik}])\Big|\|\widehat{\boldsymbol\gamma} - \boldsymbol\gamma^*\|_1^2.
\end{aligned}$$

Thus, the desired result is obtained. ■

**Proof of Lemma 1**. By the first optimality of convex optimization within the RKHS $\mathcal{H}_K$, taking partial derivative of $f$ for (2.1) but fixing $\boldsymbol\gamma$, we have

$$\frac{1}{n}\sum_{i=1}^n\big(Y_i - \langle X_i, \widehat{f}\rangle_{\mathcal{L}_2} - \mathbf{Z}_i^T\boldsymbol\gamma\big)\langle X_i, K(\cdot,)\rangle_{\mathcal{L}_2} - \lambda\widehat{f}(\cdot) = 0,$$

where the reproducing property of RKHS is applied. Letting $\theta_i = Y_i - \langle X_i, \widehat{f}\rangle_{\mathcal{L}_2} - \mathbf{Z}_i^T\boldsymbol{\gamma}$ be a sequence of scalars, we can rewrite the last formula as:

$$\widehat{f}(\cdot) = \frac{1}{\lambda n}\sum_{i=1}^{n}\theta_i\langle X_i, K(\cdot,)\rangle_{\mathcal{L}_2},\ \lambda \neq 0,$$

meaning that $\widehat{f}(\cdot)$ can be expressed by a finite basis expansion, where each basis function $B_i(\cdot)$ is generated naturally by $B_i(\cdot) = \langle X_i, K(\cdot,)\rangle_{\mathcal{L}_2}$, $i = 1, ..., n$. Since $X_i$'s are available and $K$ is specified in advance, $B_i(\cdot)$'s can be obtained from a simple integration. In other words, $\widehat{f}(\cdot)$ can be expressed as a linear combination of $n$ basis functions.

In the end, plugging the above finite formula on $\widehat{f}$ into our original objective (2.1) yields our desired finite optimization. This completes the proof of Lemma 1. ■

## References

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

J. Bezdek and R. Hathaway. Convergence of alternating optimization. *Neural, Parallel & Scientific Computations*, **11**:351–368, 2003.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **37**:1705–1732, 2009.

O. Bousquet. A bennet concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique Academie des Sciences Paris*, **334**:495–550, 2002.

S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, Cambridge, 2004.

P. Bühlmann and S. Van. de. Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, Heidelberg, 2011.

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

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

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, **13**:571–591, 2003.

A. Della Vecchia, J. Mourtada, E. De Vito, and L. Rosasco. Regularized erm on random subspaces. *In International Conference on Artificial Intelligence and Statistics*, pages 4006–4014, 2021.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice.* Springer, New York, 2006.

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

D. Kong, K. Xue, F. Yao, and H. Zhang. Partially functional linear regression in high dimensions. *Biometrika*, **103**:1–13, 2016.

M. Ledoux. On talagrand's deviation inequalities for product measures. *Probability and Statistics*, **1**:63–87, 1997.

M. Ledoux. *The Concentration of Measure Phenomenon (Mathematical Surveys and Monographs).* American Mathematical Society, Providence, RI, 2001.

Q. Li, Z. Zhu, and G. Tang. Alternating minimizations converge to second-order optimal solutions. *International Conference on Machine Learning*, pages 3935–3943, 2019.

J. Lin and V. Cevher. Convergences of regularized algorithms and stochastic gradient methods with random projections. *Journal of Machine Learning Research*, **21**:1–44, 2020.

Y. Lu, J. Du, and Z. Sun. Functional partially linear quantile regression model. *Metrika*, **77**:17–32, 2014.

M. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, **3**:1–54, 2011.

U. Marteau-Ferey, D. Ostrovskii, F. Bach, and A. Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. *In Conference on learning theory*, pages 2294–2340, 2019.

P. Müller and S. Van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, **42**:580–608, 2015.

J. Ramsay and B. Silverman. *Functional Data Analysis.* Springer, New York, 2005.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *In Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

H. Shin and M. Lee. On prediction rate in partial functional linear regression. *Journal of Multivariate Analysis*, **103**:93–106, 2012.

I. Steinwart and A. Christmann. *Support Vector Machines.* Springer Science & Business Media, 2008.

A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, New York, 2009.

S. Van. de. Geer. *Emprical Processes in M-Estimation.* Cambridge University Press, New York, 2000.

N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, **6**:38—90, 2012.

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

Y. Yang, M. Pilanci, and M. Wainwright. Randomized sketches for kernels: fast and optimal non-parametric regression. *Annals of Statistics*, **45**:991–1023, 2017.

Z. Yu, M. Levine, and G. Cheng. Minimax optimal estimation in partially linear additive models under high dimension. *Bernoulli*, **25**:1289–1325, 2019.

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

H. Zhu, F. Yao, and H. Zhang. Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society, Series B*, **76**:581–603, 2014.