LEARNING RATES FOR PARTIALLY LINEAR FUNCTIONAL MODELS WITH HIGH DIMENSIONAL SCALAR COVARIATES

YIFAN XIA AND YONGCHAO HOU
School of Statistics, Southwestern University of Finance and Economics
Chengdu, 611130, China

XIN HE
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai, China

SHAOGAO LV*
College of Statistics and Mathematics, Nanjing Audit University
Nanjing, China

ABSTRACT. This paper is concerned with learning rates for partial linear functional models (PLFM) within reproducing kernel Hilbert spaces (RKHS), where all the covariates consist of two parts: functional-type covariates and scalar ones. As opposed to frequently used functional principal component analysis for functional models, the finite number of basis functions in the proposed approach can be generated automatically by taking advantage of reproducing property of RKHS. This avoids additional computational costs on PCA decomposition and the choice of the number of principal components. Moreover, the coefficient estimators with bounded covariates converge to the true coefficients with linear rates, as if the functional term in PLFM has no effect on the linear part. In contrast, the prediction error for the functional estimator is significantly affected by the ambient dimension of the scalar covariates. Finally, we develop the proposed numerical algorithm for the proposed penalized approach, and some simulated experiments are implemented to support our theoretical results.

1. Introduction. Learning with functional data has received considerable attention over the past years due to its broad real-world applications, including medicine, linguistics, chemometrics as well as signal processing applications; see, for example, [10] and [16] for case studies. More concretely, [22] for applications in electroencephalogram (EEG) signals in brain computer interfacing, and [11] for analysis of music/speech via integration of audio content and functional brain response. For introduction to the concepts and applications of functional data analysis, we refer the interested reader to the book written by [17].

2010 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Functional data, proximal method, high dimensions, sparsity, reproducing kernel Hilbert spaces.

This work is partly supported by the National Social Science Fund of China (NSSFC-16BTJ013, NSSFC-16ZDA010), Sichuan Social Science Fund (SC14B091) and Sichuan Project of Science and Technology (2017JY0273). Shaogao Lv is the corresponding author and his research is partially supported by NSFC-11871277.

* Corresponding author.
In the problem of functional linear regression, the predictor $X(\cdot)$ is a random function defined on an interval $\mathcal{T}$, and the classical functional linear model between the response $Y$ and $X$ is

$$Y = \int_{\mathcal{T}} X(t) f_0(t) \, dt + \epsilon,$$

where $f_0$ is called the slope function and is assumed to be square-integrable. $\epsilon$ is a random error independent of $X(t)$. Let $(Y_i, X_i) : i = 1, \ldots, n$ denote independent and identically distributed realizations from the population $(Y, X)$, there is much of the practical interest in estimation of the slope function itself, or its application for the purpose of prediction $\int_{\mathcal{T}} X(t) f_0(t) \, dt$. It is worth noting that, the problem of slope-function estimation is significantly different from prediction of the slope function. The former is intrinsically nonparametric, by contrast, the latter can be either nonparametric or semiparametric. In particular, the optimal mean-square convergence rate of predictors can be attained at $n^{-1}$. The greater detailed discussion can be found in [6]. Functional linear regression has been widely studied over the last twenty years in the fields of statistics and machine learning, see [8, 26, 18].

In practice, it is often the case that a response is affected by both a vector of finite length and a function-valued random variable as predictor variables. For instance, in growth data analysis, a child’s adult height is related to his (her) past height curve and a vector that consists of the variables such as sex, heights of parents and others. These scenarios partially motivate us to study PLFM under the high dimensional setting. For simplifying the notations, we suppose that $Y$ and $X$ are centered. This paper now considers functional linear regression with single functional and scalar predictors $Z = (z_1, \ldots, z_p)$, which is a $p$-dimensional vector of scalar covariates. Partially functional linear regression between $Y$ and $(X, Z)$ is of the form

$$Y = \int_{\mathcal{T}} X(t) f^0(t) \, dt + Z^T \gamma^0 + \epsilon,$$

where $\gamma^0 = (\gamma_1^0, \ldots, \gamma_p^0)$ is referred to be the regression coefficients of nonfunctional covariates, and $\epsilon$ is the standard normal variable independent of $X(\cdot)$ and $Z$. Under the sparse high dimensional setting, a standard assumption is that the cardinality of the active set $S_0 := S(\gamma^0) := \{ j, \gamma_j^0 \neq 0, j \in \{1, \ldots, p\} \}$ is far less than $p$. Note that $p$ and $s_0 := |S_0|$ are allowed to diverge as $n$ increases. A primary purpose for (1.1) is to extract useful information from $Z$ and $X(\cdot)$, whereas the classical functional linear model focuses on a single functional covariate alone.

Recent years have witnessed a great deal of work on the high dimensional problems for linear or additive models with scalar covariates only [9, 25, 5], while a small set of partially linear functional models with high dimensional scalar covariates have been previously studied. For example, [19] considered two different approaches for the PLFM under fixed dimensional setting, which are proved to achieve the same convergence rate of the mean squared prediction error under several regularity assumptions. [21] investigated the asymptotic behaviors of their proposed estimators for the partial quantile linear regression via the truncation tool of functional principal component analysis (FPCA), and the associated application problem for neuroimaging data has been considered in [28]. However, the above mentioned work focus only on the PLFM with fixed dimensional scalar covariates, and in this case we can not know how the ambient dimensions or sparsity structure affect estimation performance induced by the slope function. Though we notice that [13] has recently considered the PLFM with high dimensional scalar covariates, they only
provided some weak learning rates of their estimators for the parametric part and the slope function. Precisely, the convergence rate of the parametric part is the same as that of the slope function, which is of the order \((s_0 + s_n)/n\) in Theorem 1 of [13]. Note that \(s_n\) over there represents the number of truncated FPCA, which diverges with a polynomial order of \(n\). So their derived rates are suboptimal in the minimax sense. In addition, their proposed method is based on the classical FPCA, which involves computationally expensive numerical optimization. Moreover, this standard truncation way via the FPCA leads to a difficult choice problem for the number of the main FPCA components.

In this paper, we analyze the high dimensional PLFM. Our main contribution is to combine the methods used in [7] for functional linear regression over RKHS with the Lasso [23] for linear models, and thus obtain an estimator that can capture the sparse structure and the functional information.

We establish theoretical results for the error bounds of both the prediction risk for the functional covariate and the estimation error for the scalar covariates. Our derived rates are sharper than the existing work in [13, 12] and our kernel-based penalized method enjoys some specific advantages with comparison in classical FPCA technique. Precisely, the convergence rate of the excess prediction risk is of the order \(s_0 \log p/n\), which is determined by the ambient dimension of the scalar covariates and the sparsity parameter \(s_0\). Surprisingly, such rate for the prediction risk is independent of the functional complexity of the RKHS, which coincides with those discussed in [6] and [7]. Intuitively, the prediction error for functional covariate is parameterized by the integral operator when \(X(\cdot)\) is sufficiently smooth. In our case of bounded process on \(X\), it is reasonable that the learning rate for functional-based prediction problem is the same rate as that of the linear part. On the other hand, the oracle rate of the linear part is the same as that for the pure linear model up to the sparsity \(s_0\), see [2], as if the functional component in the PLFM has no effect on the linear part. We conjecture that our derived rates may be optimal in the minimax sense up to the sparsity \(s_0\), which is of independent interest as our future work.

Finally, we develop the proposed numerical algorithm for the proposed penalized approach, and some simulated experiments are implemented to support our theoretical results.

2. Theoretical results.

2.1. Notation and definitions. Let \(u, v\) be two general random variables, and denote by \(Q\) the joint distribution of \((u, v)\) and by \(Q_u(Q_v)\) the marginal distribution of \(u(v)\). For \(f : u \times v \rightarrow \mathbb{R}\) a measurable function, we define the squared \(L_2\)-norm by \(\|f\|_2^2 := \mathbb{E}_Q f^2(u, v)\), and the square 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 \(\gamma \in \mathbb{R}^p\), the \(\ell_1\)-norm and \(\ell_2\)-norm are given by \(\|\gamma\|_1 := \sum_{j=1}^p |\gamma_j|\) and \(\|\gamma\|_2 := \left(\sum_{j=1}^p \gamma_j^2\right)^{1/2}\), respectively. With a little notational confusion as above, we write \(\|f\|_{L_2} := \langle f, f \rangle_{L_2}\) with \(\langle f, g \rangle_{L_2} = \int_T f(t)g(t)dt\). For two sequences \(\{a_k : k \geq 1\}\) and \(\{b_k : k \geq 1\}\), we write \(a_k \asymp b_k\) if there are positive constants \(c\) and \(c'\) such that \(c \leq a_k/b_k \leq c'\) for all \(k \geq 1\). Also, \(a_k = O(b_k)\) means that there exists some constant \(c\) such that \(a_k \leq cb_k\) for all \(k \geq 1\).

Kernel methods are one of the most powerful learning schemes in machine learning, which often take the form of the regularization schemes in a reproducing kernel...
Hilbert space (RKHS) associated with a Mercer kernel \([1]\). Recall that a kernel \(K\) is a continuous, symmetric, and positive semi-definite function on \(T \times T\). Let \(\mathcal{H} := \mathcal{H}_K\) denote be the closure of the linear span of the set of functions \(\{K_t := K(t, \cdot) : t \in T\}\) endowed with the inner product \((\sum_{i=1}^{m} \alpha_i K_{t_i}, \sum_{j=1}^{m} \beta_j K_{t_j})_K := \sum_{i,j=1}^{m} \alpha_i \beta_j K(t_i, t_j)\), for any \(\{t_i\}_{i=1}^{m}, \{t_j\}_{j=1}^{m} \in T^m\) and \(m \in \mathbb{N}^+\).

An important property on \(\mathcal{H}\) is the so-called reproducing property:

\[
\langle f, K_t \rangle_K, \text{ for any } f \in \mathcal{H}.
\]

The reproducing property ensures that any kernel-based approach defined on an infinite dimensional Hilbert pace \(\mathcal{H}\) can be equivalently transformed into a finite dimensional parametric optimization. Thus, some standard finite truncated ways such as functional PCA or spline-based approaches are avoided, and moreover the associated estimation bias can be reduced significantly. In practice, those kernel-based methods are less sensitive to the dimension of covariates in comparison with local approximated methods, for instance, local linear (constant) methods in non-parametric statistics.

Let \(X(\cdot)\) be a square integrable centered stochastic process defined over \(T\). The covariance function of \(X\) is also a real, symmetric and nonnegative definite function defined as

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

The covariance function \(C(\cdot)\) is square integrable if \(\mathbb{E}\|X\|_{L_2}^2 < \infty\). For any given reproducing kernel \(R : T \times T\), let \(L_R : L_2 \rightarrow L_2\) denote an integral operator defined by

\[
L_R(f)(\cdot) = \int_T R(s, \cdot) f(s) ds.
\]

By the reproducing property of RKHS, \(L_R\) can be equivalently defined through

\[
\langle f, L_R(g) \rangle_K = \langle f, g \rangle_{L_2}.
\]

The spectral theorem implies that there exists a family of orthonormalized eigenfunctions \(\{\phi_\ell^R : \ell \geq 1\}\) and a sequence of eigenvalues \(\theta_\ell^R \geq \theta_{\ell+1}^R \geq \cdots > 0\) such that

\[
R(s, t) = \sum_{\ell \geq 1} \theta_\ell^R \phi_\ell^R(s) \phi_\ell^R(t), \quad s, t \in T,
\]

and

\[
L_R(\phi_\ell^R) = \theta_\ell^R \phi_\ell^R, \quad \ell = 1, 2, \ldots,
\]

Let \(L_{R^{1/2}}\) be a linear operator defined by

\[
L_{R^{1/2}}(\phi_\ell^R) = \sqrt{\theta_\ell^R} \phi_\ell^R,
\]

where

\[
R^{1/2}(s, t) = \sum_{\ell \geq 1} \sqrt{\theta_\ell^R} \phi_\ell^R(s) \phi_\ell^R(t), \quad s, t \in T.
\]

It is obvious that \(L_{R^{1/2}} = L^{1/2} R\). Moreover, define

\[
R_1 R_2(s, t) = \int_T R_1(s, u) R_2(t, u) du.
\]

Then, we have \(L_{R_1 R_2} = L_{R_1} \circ L_{R_2}\).

For a given reproducing kernel \(K\) and a covariance function \(C\), define the linear operator \(L_{K^{1/2} C K^{1/2}}\) by \(L_{K^{1/2} C K^{1/2}} = L_{K^{1/2}} (L_C (L_{K^{1/2}} (f)))\). If both \(L_{K^{1/2}}\) and \(L_C\) are bounded linear operators, so is \(L_{K^{1/2} C K^{1/2}}\). By the spectral theorem, there
exist a sequence of positive eigenvalues \( s_1 \geq s_2 \geq ... > 0 \) and a set of orthonormalised eigenfunctions \( \{ \varphi_\ell : \ell \geq 1 \} \) such that
\[
K^{1/2}C K^{1/2}(s, t) = \sum_{\ell \geq 1} s_\ell \varphi_\ell(s) \varphi_\ell(t), \quad \forall s, t \in \mathcal{T},
\]
and
\[
L_{K^{1/2}C K^{1/2}}(\varphi_\ell) = s_\ell \varphi_\ell, \quad \ell = 1, 2, ....
\]
It is seen that the eigenvalues \( \{ s_\ell : \ell \geq 1 \} \) of the linear operator \( L_{K^{1/2}C K^{1/2}} \) rely on the eigenvalues of both the reproducing kernel \( K \) and the covariance function \( C \). We shall show in Section 2 that the minimax rate of convergence of the excess prediction risk is determined by the decay rate of the eigenvalues \( \{ s_\ell : \ell \geq 1 \} \).

2.2. Upper Bound. Given the available sample \( \{ Y_i, (X_i, Z_i) \}_{i=1}^n \) which are independently drawn from (1.1), the least squares regularization scheme for PLFM (1.1) is proposed by
\[
\min_{f \in \mathcal{H}, \gamma \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \int_{\mathcal{T}} X_i(t) f(t) dt - Z_i^T \gamma)^2 + \lambda \| f \|_R^2 + \rho \| \gamma \|_1 \right\}, \tag{2.1}
\]
whose minimizer is denoted by \((\hat{f}, \hat{\gamma})\). Here the tuning parameter \( \lambda > 0 \) is used to control the functional complexity of the nonparametric component, while \( \rho > 0 \) associated with the Lasso penalty [23] is used for generating sparse coefficients with respect to scalar covariates. In our theory, the choices of \( \lambda \) and \( \rho \) will be given explicitly.

Even though the proposed approach (2.1) is directly defined in an infinite-dimensional Hilbert space, the following lemma shows that this optimization is equivalent to a finite-dimensional minimization problem.

**Lemma 2.1.** Our proposed algorithm (2.1) defined on \( \mathcal{H} \times \mathbb{R}^p \) is equivalent to a finite-dimensional parametric convex optimization.

The proof of Lemma 2.1 is deferred to the Appendix. This current article focuses on the prediction problem with the functional covariate, and the oracle rate of the high dimensional parametric estimator. The goal of prediction is to recover the functional \( \zeta_X(f^0) \), where
\[
\zeta_X(f) := \int_{\mathcal{T}} X(t) f(t) dt.
\]

Before establishing our main theoretical results for the functional and the parametric estimation, we introduce several basic notation and assumptions used in the proof. We first define a mixed set by
\[
\mathcal{G} := \{ g(X, Z) = \zeta_X(f) + Z^T \gamma, \ f \in \mathcal{H}, \gamma \in \mathbb{R}^p \},
\]
and \( g^0(X, Z) := \zeta_X(f^0) + Z^T \gamma^0 \). Define a sequence of projections of \( Z = (z_1, ..., z_p) \) into the integral-type functional of \( X(\cdot) \) and \( \mathcal{H} \) by
\[
\pi_j(\cdot) := \arg \min_{f \in \mathcal{H}} \| z_j - \zeta_X(f) \|^2, \quad j = 1, ..., p.
\]
We denote \( \Pi := (\zeta_X(\pi_1), ..., \zeta_X(\pi_p)) \in \mathbb{R}^p \) and \( \tilde{\Pi} := Z - \Pi \). For every \( g, g^0 \in \mathcal{G} \),
\[
g - g^0 = \tilde{\Pi}^T (\gamma - \gamma^0) + [\zeta_X(f - f^0) + \Pi^T (\gamma - \gamma^0)].
\]
This together with the definition of projection implies that
\[
\| g - g^0 \|^2 = \| \tilde{\Pi}^T (\gamma - \gamma^0) \|^2 + \| \zeta_X(f - f^0) + \Pi^T (\gamma - \gamma^0) \|^2, \forall g \in \mathcal{G}. \tag{2.2}
\]
For some fixed constant $\delta_0$, we define
\[
\tau_{\rho,R} := \tau(g) = \rho \|\gamma\|_1/(R\sqrt{\delta_0/7}) + \left(\|\zeta_X(f) + Z^T\gamma\|^2 + \lambda \|f\|_K^2\right)^{1/2},
\]
where $R$ is a small positive quantity to be determined later (depending on $n$). Under some regularity conditions, we will show that $\tau(\hat{g} - g^0) \leq R$ with a high probability. Then conditional on the event $\{(X, Z), \tau(\hat{g} - g^0) \leq R\}$, consistency and optimality of our estimator $(\hat{f}, \hat{\gamma})$ can be proved accordingly.

Next, we present several technical assumptions used in our theory.

**Condition A** (Eigenvalue condition). The smallest eigenvalue $\Lambda_{\min,\hat{H}}^2$ of $E[\hat{H}^T \hat{H}]$ is positive uniformly.

Assumption A is standard in semiparametric literature, see [15], ensuring sufficient information in estimating $\gamma$.

**Condition B** (Design condition). (1) For some constant $C_z$, there holds $|z_{ij}| \leq C_z$, for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. (2) $\|X\|_{\mathcal{L}_2} \leq C_x < \infty$ almost surely. (3) $\kappa := \sup_{t \in \mathcal{T}} |K(t, t)| < \infty$.

(1) and (2) on Condition B seem to be strong assumptions, while we can often approximate a non-bounded distribution or process with its truncated version. (3) on Condition B is mild and the commonly-used Gaussian kernel satisfies such condition.

**Condition C** (Positiveness condition). For some constant $C_K$, there holds $\int_{\mathcal{T}} f^2(t) dt \leq C_K E\left(\int_{\mathcal{T}} X(t)f(t) dt\right)^2$ for all $f \in \mathcal{H}$.

Condition C looks strong a bit but it holds when $E[X(t)X(t')] > c > 0$ for any $t, t' \in \mathcal{T}$. For example, any centered independent difference process with $E[X^2(t)] > 0$ uniformly on $t \in \mathcal{T}$. In the case of Gaussian process and $\mathcal{T} = [1, 2]$, we know that $E[X(t)X(t')] = \min\{t, t'\} \geq 1$. This condition is also viewed as an identification condition in the sense of square integral.

In order to present our general theoretical results, we introduce some events under which our results are stated explicitly. For any given $R > 0$
\[
\mathcal{G}_R := \{g : \tau(g) \leq R\}, \quad \text{and} \quad \mathcal{T}(\delta_0, R) := \mathcal{T}_1(\delta_0, R) \cap \mathcal{T}_2(\delta_0, R),
\]
where
\[
\mathcal{T}_1(\delta_0, R) := \{(X, Z) : \sup_{g \in \mathcal{G}_R} \left\|\|g\|_n^2 - \|g\|^2\right\| \leq \delta_0 R^2\}
\]
and
\[
\mathcal{T}_2(\delta_0, R) := \{(X, Z, \epsilon) : \frac{1}{n} \sup_{g \in \mathcal{G}_R} \left|\sum_{i=1}^n \epsilon_i g(X_i, Z_i)\right| \leq \delta_0 R^2\},
\]

**Theorem 2.2.** Given $R > 0$ and $0 < \delta_0 \leq 1/700$. Under Conditions A and that
\[
\frac{\rho^2 \delta_0}{\Lambda_{\min,\hat{H}}^2} \leq \frac{\delta_0 R^2}{4}, \quad \lambda \leq \frac{\delta_0 R^2}{8\|f^0\|_K^2},
\]
Then on $\mathcal{T}(\delta_0, R)$, we have $\tau(\hat{g} - g^0) \leq R$.

Note that $\delta_0$ is just a technical quantity and we do not pursuit the optimality of constants. Besides, Theorem 1 can not provide the optimal choices for the parameters $\rho, \lambda$ and $R$, since we shall show that $\mathcal{T}(\delta_0, R)$ holds under some additional constraints on these three parameters.

It is seen from Theorem 1 that, the learning rates for the prediction estimation and the linear part can be obtained easily once $\mathcal{T}(\delta_0, R)$ holds with a high probability. In the literature of empirical processes, $\mathcal{T}_1(\delta_0, R)$ is related to uniform law
of large number with the constrained set $\mathcal{G}_R$, which is often characterized by the so-called Rademacher complexity, defined in Section 3. By contrast, $\mathcal{F}_s(\delta_0, R)$ is closely related to Gaussian complexity, which has been studied well in [14].

**Theorem 2.3.** Under Conditions A-C. Assume that the triple parameters $(R, \rho, \lambda)$ can be chosen such that

$$R \asymp \sqrt{\frac{s_0 \log p}{n}}, \quad \rho \asymp \sqrt{\frac{\log p}{n}}, \quad \lambda \asymp \frac{s_0 \log p}{n}.$$  

Then, with probability at least $1 - \delta'$,

$$\left\| \int_{\mathcal{T}} X(t)(\hat{f}(t) - f_0(t))dt \right\|^2 = O\left(\frac{s_0^2 \log p}{n}\right),$$

and

$$\left\| \hat{\gamma} - \gamma_0 \right\|_1 = O\left(s_0 \sqrt{\frac{\log p}{n}}\right).$$

Remark that, the convergence rate of the excess prediction risk is of the order $\frac{s_0^2 \log p}{n}$, which is determined by the ambient dimension of the scalar covariates and the sparsity parameter $s_0$. However, such rate for the prediction risk is independent of the functional complexity of the RKHS, which coincides with those discussed in [6] and [7]. More precisely, the prediction error for functional covariate is parameterized by the integral operator when $X(\cdot)$ is sufficiently smooth. In our case of bounded process on $X$, it is reasonable that the learning rate for functional-based prediction problem is achieved as the same rate as the linear part. On the other hand, the oracle rate of the linear part is the same as that for the pure linear model, see [2], as if the functional component in the PLFM has no affect on the linear part.

Note also that the choice of the nonparametric tuning parameter $\lambda$ is of the order $\frac{s_0 \log p}{n}$, which is strictly less than the ordinary choice in nonparametric literature (i.e. $n^{-2r/(2r+1)}$) with the $r$-times differentiable functions. This under-smoothing requirement is often encountered in two-step estimation in statistics. Due to the smoothness of integration and bounded process of $X$, we allow more complex functions to be entered into our algorithm, which corresponds to a smaller quantity of $\lambda$.

### 3. Computing algorithm.

The proposed framework aims to solve the following optimization problem

$$\min_{f \in H, \gamma \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \int_{\mathcal{T}} X_i(t)f(t)dt - Z_i^T \gamma \right)^2 + \lambda \| f \|_{\mathcal{K}} + \rho \| \gamma \|_1 \right\}. \quad (3.1)$$

By Lemma 2.1, the minimizer of (3.1) must have a finite form that

$$\hat{f}(t) = \frac{1}{n\lambda} \sum_{i=1}^{n} \theta_i \int_{\mathcal{T}} X_i(s) K_{s}(t)ds, \quad (3.2)$$

Then plug (3.2) into (3.1), we have

$$\min_{\theta \in \mathbb{R}^n, \gamma \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \frac{1}{n\lambda} \theta^T B(X_i) - Z_i^T \gamma \right)^2 + \frac{1}{n^2\lambda} \theta^T \Sigma \theta + \rho \| \gamma \|_1 \right\}, \quad (3.3)$$

where $B(X_i) = B_1(X_i), ..., B_n(X_i))^T \in \mathbb{R}^n$ with $B_j(X_i) = \int_{\mathcal{T}} \int_{\mathcal{T}} X_j(s)K(s,t)X_i(t)dsdt$ and $\Sigma \in \mathbb{R}^{n \times n}$ with components $\Sigma_{ij} = \int_{\mathcal{T}} \int_{\mathcal{T}} X_i(t)K(t,s)X_j(s)dsdt$. Hence,
the optimization problem (3.1) over a probably infinite function space changes to the optimization problem in a n-dimensional vector space.

To solve (3.3), we consider a splitting algorithm with Proximal operator, which updates θ and γ sequentially. For simplicity, we denote the first term of (3.3) as $R_n(\theta, \gamma)$ and the last term as $\Omega(\gamma)$. Then at the t-th iteration with solution $(\theta^t, \gamma^t)$, we have

$$\theta^{t+1} = \min_{\theta \in \mathbb{R}^n} \left\{ R_n(\theta, \gamma^t) + \frac{1}{n^2 \lambda} \theta^T \Sigma \theta + \rho \|\gamma^t\|_1 \right\},$$ (3.4)

$$\gamma^{t+1} = \min_{\gamma \in \mathbb{R}^p} \left\{ R_n(\theta^{t+1}, \gamma) + \frac{1}{n^2 \lambda} (\theta^{t+1})^T \Sigma \theta^{t+1} + \rho \|\gamma\|_1 \right\},$$ (3.5)

Specifically, we have

$$\theta^{t+1} = \left( \frac{1}{n \lambda} \sum_{i=1}^n B(X_i)B(X_i)^T + \Sigma \right)^{-1} \left( \sum_{i=1}^n (y_i - Z_i^T \gamma^t) B(X_i) \right)$$

$$\gamma^{t+1} = \text{Prox}_\rho \left\{ \gamma^t - \frac{1}{D} \nabla \gamma R_n(\theta^{t+1}, \gamma^t) \right\},$$

where Prox is a proximal operator that is explicitly expressed as $\text{Prox}_\rho \|\cdot\|_1 (u) = \text{sign}(u)(|u| - \frac{\rho}{2})_+$, $D$ is the Lipschitz upper bound of $R_n(\theta^{t+1}, \gamma^t)$ and $\nabla R_n(\theta, \gamma) = -\frac{2}{n} \sum_{i=1}^n (Y_i - \frac{1}{n} \theta^T B(X_i) - Z_i^T \gamma) Z_i$. Note that in practice, the exact value of the Lipschitz upper bound $D$ is often difficult to determine in large-scale problems. And thus we recommend a backtracking scheme [4] as a more efficient alternative to approximate the value of $D$.

4. Numerical analysis. In this section, we illustrate the performance of the proposed method in simulated cases. We consider the simulation setting used by [27]. Specifically, we assume that $\mathcal{T} = [0, 1]$ and the true slope function and the random function $X(\cdot)$ are

$$f^0(t) = 4 + \sum_{k=2}^{50} 4(-1)^{k+1}k^{-2}\sqrt{2}\cos(k\pi t), \text{ and } X(t) = \xi_1 U_1 + \sum_{k=2}^{50} \xi_k U_k \sqrt{2}\cos(k\pi t),$$

where $U_k \sim U(-\sqrt{3}, \sqrt{3})$ and $\xi_k$ are pre-defined as $\xi_k = (-1)^{k+1}k^{-v/2}$ with some pre-specified positive constant $v$. For the linear part, the true regression coefficients $\gamma^0 = (-5, 5, 0, ..., 0)^T$ and random variable $Z = (Z_1, ..., Z_p)^T \in \mathbb{R}^{n \times p}$ with $Z_i = (Z_{i1}, ..., Z_{ip})^T$ is generated as $Z_{ij} \sim N(0, 1)$. Then, the response is generated as

$$Y_i = \int_0^1 f^0(t)X_i(t)dt + Z_i^T \gamma^0 + \epsilon_i,$$ (4.1)

where $\epsilon_i \sim N(0, 1)$ denotes the random noise.

The proper choice of the tuning pairs $(\lambda, \rho)$ plays a crucial role in achieving the desired performance of the proposed method. Here, we apply the ABIC criterion used in [13]. Specifically, given a fixed pair $(\lambda, \rho)$, we consider the criterion that

$$ABIC(\lambda, \rho) = \log \text{RSS}(\lambda, \rho) + \frac{\text{Card}(\hat{\gamma}) \log n}{n},$$

where $\text{RSS}(\lambda, \rho) = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{T} \sum_{t=1}^T \hat{f}(t) x_i(t) + Z_i^T \hat{\gamma})^2$ and $\text{Card}(\gamma)$ denotes the number of nonzero components of $\hat{\gamma}$. Then, we prefer to choose the pair $(\lambda, \rho)$ which minimizes this criterion via grid search, where the grid is set as $\{10^{-2+s0.1s}: s = 0, ..., 40\}$. 

For the numerical example, we use the kernel function
\[ K(s, t) = \frac{1}{(2!)^2} B_2(s) B_2(t) - \frac{1}{4!} B_4(|s - t|), \]
where \( B_m(\cdot) \) is the \( m \)-th Bernoulli polynomial. The scenarios with \( n = 100, 200 \) and \( v = 1, 2, 4 \) have been considered. Each scenario was replicated 50 times. The averaged performance measures are summarized in Table 1, including \( \| \hat{\gamma} - \gamma^0 \|_2 \), \( E\| \hat{f} - f^0 \| \) and \( \| \hat{y} - y \|_2 \). Note that the expectation and the prediction error are evaluated by using 1000 independently and identically generated testing samples, and similar measures have been adopted by [27] and [13].

The averaged performance measures are summarized in Table 1. Note that the integrals in calculation of \( \Sigma, B(X_i) \) and \( f(t) \) can be approximated by summations for practical purposes. Specifically, in our method, we generate 1000 points in \([0, 1]\) with equal distance and evaluate the integral by using the generated points. Clearly, we conclude that with the increase of \( n \) and \( v \), the performance of the proposed method is getting better in the sense that all the errors decrease significantly.

### Table 1. The averaged performance measures of the proposed method in simulated example.

| (n, v)   | \( \| \hat{\gamma} - \gamma^0 \|_2 \) | \( E\| \hat{f} - f^0 \| \) | \( \| \hat{y} - y \|_2 \) |
|----------|---------------------------------|------------------------------|------------------|
| (100,1.1)| 0.3401 (0.0166)                 | 4.0851 (0.0085)              | 4.3526 (0.0151)  |
| (100,2)  | 0.3338 (0.0166)                 | 4.0578 (0.0085)              | 4.3226 (0.0151)  |
| (100,4)  | 0.3232 (0.0166)                 | 4.0329 (0.0085)              | 4.2887 (0.0151)  |
| (200,1.1)| 0.2235 (0.0136)                 | 4.0797 (0.0088)              | 4.2708 (0.0137)  |
| (200,2)  | 0.2230 (0.0134)                 | 4.0540 (0.0089)              | 4.2460 (0.0134)  |
| (200,4)  | 0.2166 (0.0119)                 | 4.0313 (0.0087)              | 4.2185 (0.0126)  |

5. Proofs. In order to prove Theorem 1, we need to introduce the classical Talagrand’s Concentration Inequality, which can be found in [3] or [20].

**Lemma 5.1** (Talagrand’s Concentration Inequality). Let \( \mathcal{G} \) be a function class on \( Z \) that is separable with respect to \( \infty \)-norm, and \( \{z_i\}_{i=1}^n \) be i.i.d. random variables with values in \( Z \). Furthermore, let \( B > 0 \) and \( U \geq 0 \) be \( B := \sup_{g \in \mathcal{G}} \mathbb{E}[(g - \mathbb{E}[g])^2] \) and \( U := \sup_{g \in \mathcal{G}} \|g\|_\infty \), then there exists a universal constant \( N_0 \) such that, for \( Z = \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(z_i) - \mathbb{E}[g] \right| \), we have
\[
P\left( Z \geq N_0 \left[ \mathbb{E}[Z] + \sqrt{\frac{Br}{n}} + \frac{Ur}{n} \right] \right) \leq e^{-r}, \quad \forall r > 0.
\]

**Lemma 5.2.** We choose the tuning parameters as the following way:
\[
R/\rho \leq c_1 \sqrt{\delta_0}, \quad \sqrt{n} \rho \geq \sqrt{2 \log p / \delta_1}, \quad \lambda n \geq \sqrt{\delta_0 / \delta_2}, \quad R^2 \leq c_2 \lambda,
\]
where \( c_1, c_2 \) are fixed constants, while \( \delta_1, \delta_2 \) are sufficiently small constants and mainly depend on \( \delta_0 \). Then, with probability at least \( 1 - \exp(-\delta^0 n \rho^2) \) with small enough constant \( \delta^0 > 0 \),
\[
\sup_{g \in \mathcal{G}_R} \left| \|g\|_n^2 - \|g\|^2 \right| \leq \delta_0 R^2.
\]
Proof. In order to apply Talagrand’s Concentration Inequality as above, it suffices to bound these key quantities: $\mathbb{E}[\mathcal{Z}]$, $B$, $U$. To this end, we introduce \( \{\sigma_i\}_{i=1}^n \) as a Rademacher sequence independent of \( \{(X_i, Z_i)\}_{i=1}^n \). By symmetrization, it is shown in Corollary 3.4 of [24] that $\mathbb{E} \sup_{\gamma \in \mathcal{G}_R} \|g\|_n^2 - \|g\|^2$ can be upper bounded by

$$2\mathbb{E} \sup_{g \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g_i^2 \right| = 2\mathbb{E} \sup_{g \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i ((Z_i^T \gamma)^2 + 2(Z_i^T \gamma) \zeta_i(f) + \zeta_i^2(f)) \right|,$$

where we write $\zeta_i(f) := \zeta_{X,i}(f)$ for the ease of notation. We next consider these three terms respectively.

Note that for $\zeta_X(f) + Z^T \gamma \in \mathcal{G}_R$, by Condition B, we have

$$|Z^T \gamma| \leq C_2 \|\gamma\|_1 \leq \frac{C_2 R^2}{\sqrt{\lambda} / \delta_0} \leq \frac{C_2 c_1}{\sqrt{\lambda} / \delta_0},$$

where we used the fact $R^2/(c_1 \rho) \leq 1$ due to $R \leq 1/\sqrt{\delta_0}$ and $R/\rho \leq c_1 \sqrt{\delta_0}$. By the contraction property of Rademacher complexity, we have

$$\mathbb{E} \sup_{\gamma: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i ((Z_i^T \gamma)^2 + 2(Z_i^T \gamma) \zeta_i(f) + \zeta_i^2(f)) \right| \leq \frac{4C_2 c_1}{\sqrt{\lambda} / \delta_0} \mathbb{E} \sup_{\zeta: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i (Z_i^T \gamma) \right|.$$ 

Note that

$$\mathbb{E} \left\| \sum_{i=1}^n \sigma_i Z_i \right\|_{\infty} \leq \max_{1 \leq j \leq p} \left( \sum_{i=1}^n \left| z_{ij} \right| \right)^{1/2} \sqrt{2 \log p} \leq C_2 \sqrt{2 \log p} \sqrt{n}.$$

This follows that

$$\mathbb{E} \sup_{\gamma: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i (Z_i^T \gamma) \right| \leq \mathbb{E} \sup_{\gamma: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i Z_i \right| \|\gamma\|_1 \leq \frac{R^2}{\sqrt{\lambda} / \delta_0} \mathbb{E} \left| \sum_{i=1}^n \sigma_i Z_i \right| / n \leq \frac{C_2 \sqrt{2 \log p} R^2}{\sqrt{\lambda} / \delta_0} \leq \frac{C_2 \delta_1 R^2}{\sqrt{\lambda} / \delta_0},$$

where we used the assumption $\sqrt{n} \rho \geq \sqrt{2 \log p} / \delta_1$. Thus, we obtain

$$\mathbb{E} \sup_{\gamma: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i (Z_i^T \gamma)^2 \right| \leq \frac{4C_2 \delta_1 C_2^2 R^2}{\sqrt{\lambda}}. \tag{5.1}$$

We now consider the term: $\mathbb{E} \sup_{f: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \zeta_i^2(f) \right|$. Note that for $\zeta_X(f) + Z^T \gamma \in \mathcal{G}_R$, we have $\|f\|_\infty \leq \kappa \|f\|_K \leq \kappa R / \sqrt{\lambda}$. This implies $|\zeta_X(f)| \leq \|X\|_\mathcal{L}_2 \|f\|_\infty \leq \kappa C_2 R / \sqrt{\lambda}$. By the contraction property again, one gets

$$\mathbb{E} \sup_{f: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \zeta_i^2(f) \right| \leq \frac{4\kappa C_2 R}{\sqrt{\lambda}} \mathbb{E} \sup_{f: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \zeta_i(f) \right|$$

and by Jensen’s inequality

$$\mathbb{E} \sup_{f: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \zeta_i(f) \right| \leq \frac{1}{n} \left( \int_{-T}^T \mathbb{E} \left( \sum_{i=1}^n \sigma_i X_i(t)^2 \right) dt \right)^{1/2} \sup_{f: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \|f\|_{\mathcal{L}_2} \leq \frac{C_x}{\sqrt{n}} \sup_{f: \zeta_X(f) + Z^T \gamma \in \mathcal{G}_R} \|\zeta_X(f)\|, \tag{5.2}$$
where we used the assumption on Condition C. Note now that for any \( f \) satisfying \( \xi_X(f) + Z^T\gamma \in \mathcal{G}_R \), we have \( \| \xi_X(f) + Z^T\gamma \| \leq R \) and \( \| Z^T\gamma \| \leq C_z \| \gamma \|_1 \leq C_z \sqrt{\delta_0/T^2/\rho} \). The triangle inequality is applied to imply that

\[
sup_{f: \xi_X(f) + Z^T\gamma \in \mathcal{G}_R} \| \xi_X(f) \| \leq (1 + C_z \sqrt{\delta_0/T} R/\rho) R \leq (1 + c_1 \sqrt{s_0}) R,
\]

since \( R/\rho \leq c_1 \sqrt{s_0} \) by assumption. Hence, this follows from (5.2) that

\[
E \sup_{f: \xi_X(f) + Z^T\gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \xi_i(f) \right| \leq \left( \frac{4 \sqrt{\lambda n}}{\sqrt{\delta_0 \rho}} \right) \sup_{f: \xi_X(f) + Z^T\gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \xi_i(f) \right| \| \gamma \|_1
\]

provided that \( \lambda n \geq \sqrt{s_0/\delta_2} \). It remains to bound the term \( E \sup_{g \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i (Z^T\gamma)_{\xi_i}(f) \right| \). By a direct calculation, we have

\[
E \sup_{g \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i (Z^T\gamma)_{\xi_i}(f) \right| \leq \frac{C_z \sqrt{2 \log \rho} R^2}{\sqrt{\delta_0 \rho}} \sup_{f: \xi_X(f) + Z^T\gamma \in \mathcal{G}_R} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \xi_i(f) \right| \| \gamma \|_1
\]

where the second inequality is based on the contraction property of Rademacher complexity, and the third one follows from the above derived results from (5.2) to (5.3). Thus, combining (5.1), (5.3) and (5.4), we conclude that

\[
E \mathcal{Z} = E \sup_{g \in \mathcal{G}_R} \left( ||g||_n^2 - ||g||^2 \right) \leq \left( \frac{8 c_1 \delta_0 \delta_1 C_z^2}{7} + 8 \kappa (1 + c_1) \delta_2 C_x^2 + \frac{2 (1 + c_1) C_x C_z \delta_1}{\sqrt{T/\delta_0}} \right) R^2.
\]

Besides, we also have that

\[
U = \sup_{g \in \mathcal{G}_R} \| g \|_n \leq \sup_{f: \xi_X(f) + Z^T\gamma \in \mathcal{G}_R} \left( \kappa C_z \| f \|_{\mathcal{X}} + C_z \| \gamma \|_1 \right) \leq \frac{\kappa C_z R}{\sqrt{\lambda}} + \frac{C_z}{\sqrt{T/\delta_0}}.
\]

where we used the assumption \( R^2 \leq c_2 \lambda \). By the definition of \( \mathcal{G}_R \)

\[
B = \sup_{g \in \mathcal{G}_R} \mathbb{E}[ (g - \mathbb{E}[g])^2 ] \leq \sup_{g \in \mathcal{G}_R} \mathbb{E}[g^2] \leq R^2.
\]

Then, an application of Talagrand’s Concentration Inequality yields that

\[
\sup_{g \in \mathcal{G}_R} \left( ||g||_n^2 - ||g||^2 \right) \leq 2 N_0 \delta''_1 R^2 + N_0 R \sqrt{\lambda \delta''} + N_0 c'' \rho^2 \delta''^2,
\]

with probability at least \( 1 - \exp(-n \rho^2 \delta'') \) by taking \( r := n \rho^2 \delta'' \). Here

\[
\delta''_1 := \frac{8 c_1 \delta_0 \delta_1 C_z^2}{7} + 8 \kappa (1 + c_1) C_z^2 \delta_2 + \frac{2 (1 + c_1) C_x C_z \delta_1}{\sqrt{T/\delta_0}}.
\]

With the choice of \( \rho^2 \leq \frac{\sqrt{\lambda} \delta_0 \ln n}{4 s_0} R^2 \) := \( c' R^2 \) as Lemma 2.2, and taking now \( \delta_1, \delta_2, \delta'' \) small enough, such that

\[
2 N_0 \delta''_1 + N_0 \sqrt{c' \delta''} + N_0 c' \delta'' \leq \delta_0,
\]

so we obtain our desired result with probability at least \( 1 - \exp(-n \rho^2 \delta'') \).
To consider the term $T_2(\delta_0, R)$, we introduce the following lemma. Besides the Rademacher processes, The Gaussian concentration inequality plays an important role in our analysis as well, which is given in Theorem 7.1 of [14].

**Lemma 5.3.** Let $G = \{G_t\}_{t \in \mathcal{T}}$ be a centered Gaussian process indexed by a countable set $\mathcal{T}$ such that $\sup_{t \in \mathcal{T}} G_t < \infty$ almost surely. Then

$$\mathbb{P}\left( \sup_{t \in \mathcal{T}} G_t \geq \mathbb{E}[\sup_{t \in \mathcal{T}} G_t] + r \right) \leq \exp(-\frac{r^2}{2\bar{\sigma}^2}),$$

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

**Lemma 5.4.** If the following bounds hold with small enough $\delta_3 > 0$:

$$\sqrt{s_0}/n \leq \delta_3 R, \quad \sqrt{2}p/\delta_3, \quad \rho^2 \leq \frac{\delta_0 \Lambda_{\min, \tilde{n}}^2 R^2}{4s_0}.$$

Then, with probability at least $1 - \exp(-\delta''n\rho^2)$,

$$\frac{1}{n} \sup_{g \in \mathcal{G}_R} \left| \sum_{i=1}^n \epsilon_i g(X_i, Z_i) \right| \leq \delta_0 R.$$

**Proof.** For any $g \in \mathcal{G}_R$, we observe that

$$\left\| \sum_{i=1}^n \epsilon_i g(X_i, Z_i) \right\|_2 \leq \left\| \sum_{i=1}^n \epsilon_i X_i(t) \right\|_2 \| f \|_2 + \sup_j \left\| \sum_{i=1}^n \epsilon_i z_{ij} \right\| \| \gamma \|_1.

First, by Condition C and the same arguments as Lemma 5.2, one gets

$$\| f \|_2 \leq (1 + c_1 \sqrt{s_0}) R, \quad \forall f : \xi_x(f) + Z^T \gamma \in \mathcal{G}_R,$$,

as well as $\| \gamma \|_1 \leq \frac{R^2}{\sqrt{\delta_0} R}$. Since $\epsilon_i$'s are i.i.d. sub-gaussian random variables, by the Jensen’s inequality we obtain

$$\mathbb{E}\left( \sum_{i=1}^n \epsilon_i X_i(t) \right)^2 \leq \left( \int_{\mathcal{T}} \mathbb{E}\left( \sum_{i=1}^n \epsilon_i X_i(t) \right)^2 dt \right)^{1/2} = \sqrt{n \mathbb{E}[\| X \|_{L_2}^2]},$$

and conditional on any fixed $z_{ij}$

$$\mathbb{E} \max_{1 \leq j \leq p} \left| \sum_{i=1}^n \epsilon_i z_{ij} \right| \leq \left( \sum_{i=1}^n z_{ij}^2 \right)^{1/2} \sqrt{2} \log p \leq C_x \sqrt{2} \log p \sqrt{n}.

All the mentioned results are combined to imply that

$$\mathbb{E} \frac{1}{n} \sup_{g \in \mathcal{G}_R} \left| \sum_{i=1}^n \epsilon_i g(X_i, Z_i) \right| \leq (1 + c_1 \sqrt{s_0}) C_x R/\sqrt{n} + \frac{C_x \sqrt{2} \log p R^2}{\sqrt{\delta_0} \rho \sqrt{n}} \leq (1 + c_1) C_x \delta_3 R^2 + \frac{C_x \delta_1 R^2}{\sqrt{\delta_0} \rho \sqrt{n}}.

Moreover, we have

$$\bar{\sigma}^2 = \sup_{t \in \mathcal{T}} \mathbb{E}[G_t^2] \leq \frac{1}{n} \left\| g \right\|^2 \leq \frac{1}{n} R^2.$$

Let $r := \sqrt{\delta''} \rho R$, by Lemma 5.3 we conclude that, with probability at least $1 - \exp(-\delta''n\rho^2)$,

$$\frac{1}{n} \sup_{g \in \mathcal{G}_R} \left| \sum_{i=1}^n \epsilon_i g(X_i, Z_i) \right| \leq (1 + c_1) C_x \delta_3 R^2 + \frac{C_x \delta_1 R^2}{\sqrt{\delta_0} \rho \sqrt{n}} + \sqrt{\delta_0 \delta'' \Lambda_{\min, \tilde{n}} R^2}.$$
It is seen from the above inequality, and we can get our desired result as long as \( \delta_3, \delta_1, \delta'' \) are sufficiently small with respect to \( \delta_0 \).

**Appendix.**

**Proof of Lemma 2.1.** By the first optimality of convex optimization, taking partial derivative of \( \beta \) for (2.1) but \( \gamma \) is fixed, we have

\[
\frac{1}{n} \sum_{i=1}^{n} (Y_i - \int_T X_i(t)f(t)dt - Z^T_i \gamma) \int_T X_i(t)K_i(\cdot)dt - \lambda \hat{f}(\cdot) = 0,
\]

where the reproducing property of RKHS is applied. Letting \( \theta_i = (Y_i - \int_T X_i(t)f(t)dt - Z^T_i \gamma) \) be a sequence of scalars, we can rewrite the last formula as:

\[
\hat{f}(\cdot) = \frac{1}{\lambda n} \sum_{i=1}^{n} \theta_i \int_T X_i(t)K_i(\cdot)dt, \: \lambda \neq 0,
\]

which means that \( \hat{f}(\cdot) \) can be expressed by a finite basis expansion, where each basis function \( B_i(\cdot) \) is generated naturally by \( B_i(\cdot) = \int_T X_i(t)K_i(\cdot)dt, \: i = 1, \ldots, n. \)

Since \( X_i \)'s are available and \( K_i \) is specified in advance, \( B_i(\cdot) \)'s can be obtained from a simple integration. In other words, \( \hat{f}(\cdot) \) can be expressed as a linear combination of the \( n \) number of basis functions. This completes the proof of Lemma 2.1.

**Proof of Theorem 1.** Define

\[
\tilde{g} := s\tilde{g} + (1-s)g^0 = \int_T X(t)\tilde{f}(t)dt + Z^T\tilde{\gamma},
\]

where \( \tilde{f} = s\hat{f} + (1-s)f^0 \) and \( s\tilde{\gamma} + (1-s)\gamma^0 \) and \( s := R/(R + \tau(\hat{g} - g^0)) \). We now prove the key conclusion: \( \tau(\hat{g} - g^0) \leq R/2 \), which easily results in our desired result.

By convexity and definition of \( (\hat{f}, \tilde{\gamma}) \), we have that

\[
\|Y - \tilde{g}\|_n^2 + \lambda\|\tilde{f}\|_K^2 + \rho\|\tilde{\gamma}\|_1 \leq s(\|Y - \hat{g}\|_n^2 + \lambda\|\hat{f}\|_K^2 + \rho\|\hat{\gamma}\|_1) + (1-s)(\|Y - g^0\|_n^2 + \lambda\|f^0\|_K^2 + \rho\|\gamma^0\|_1)
\]

\[
\leq \|Y - g^0\|_n^2 + \lambda\|f^0\|_K^2 + \rho\|\gamma^0\|_1.
\]

This is the basic inequality among oracle analysis. Using the model \( y = g^0 + \epsilon \), the above inequality can be rewritten as

\[
\|\tilde{g} - g^0\|_n^2 + \lambda\|\tilde{f}\|_K^2 + \rho\|\tilde{\gamma}\|_1 \leq \frac{2}{n} \sum_{i=1}^{n} \epsilon_i(\tilde{g} - g^0)(X_i, Z_i) + \lambda\|f^0\|_K^2 + \rho\|\gamma^0\|_1.
\]

Observe that

\[
\tau(\tilde{g} - g^0) = s\tau(\hat{g} - g^0) = \frac{R\tau(\hat{g} - g^0)}{R + \tau(\hat{g} - g^0)} \leq R.
\]

Hence, on \( \mathcal{T}(\delta_0, R) \), we have

\[
\|\tilde{g} - g^0\|_n^2 + \lambda\|\tilde{f}\|_K^2 + \rho\|\tilde{\gamma}\|_1 \leq 3\delta_0 R^2 + \lambda\|f^0\|_K^2 + \rho\|\gamma^0\|_1.
\]
Using (2.2) and the identity $\|\gamma\|_1 = \|\gamma S_0\|_1 + \|\gamma S_0\|_1$, we obtain from the triangle inequality
\[
\|\tilde{\Pi}^T(\tilde{\gamma} - \gamma^0)\|^2 + \|\int_T X(t)(\tilde{f} - f^0)(t)dt + \Pi^T(\tilde{\gamma} - \gamma^0)\|^2 + \rho\|\tilde{\gamma} S_0\|_1 + \lambda\|\tilde{f}\|_K^2
\leq 3\delta_0 R^2 + \lambda\|f^0\|_K^2 + \rho\|\tilde{\gamma} - \gamma^0\|_1.
\] (5.5)
By the basic inequality $xx' \leq x^2 + x'^2/4$ and Condition A, we get
\[
\rho\|\tilde{\gamma} - \gamma^0\|_1 \leq \rho\sqrt{s_0} \|\tilde{\gamma} - \gamma^0\|_2 \leq \frac{\rho\sqrt{s_0}}{\Lambda_{\text{min},1}} \|\tilde{\Pi}(\tilde{\gamma} - \gamma^0)\|
\leq \frac{\rho^2 s_0}{\Lambda_{\text{min},1}} + \frac{\|\tilde{\Pi}(\tilde{\gamma} - \gamma^0)\|^2}{4} \leq \frac{\delta_0 R^2}{4} + \frac{\|\tilde{\Pi}(\tilde{\gamma} - \gamma^0)\|^2}{4},
\] (5.6)
where the last inequality follows from the assumption of this lemma. On the other hand, noting that $\|\tilde{f} - f^0\|_K^2 \leq 2(\|\tilde{f}\|_K^2 + \|f^0\|_K^2)$, we conclude from (5.6) that
\[
\frac{3}{4} \|\tilde{\Pi}^T(\tilde{\gamma} - \gamma^0)\|^2 + \|\int_T X(t)(\tilde{f} - f^0)(t)dt + \Pi^T(\tilde{\gamma} - \gamma^0)\|^2 + \lambda/2\|\tilde{f} - f^0\|_K^2
\leq \frac{14}{4} \delta_0 R^2 - \rho\|\tilde{\gamma} S_0\|_1 \leq \frac{14}{4} \delta_0 R^2,
\]
where we use the assumption on parameter constraint: $\lambda \leq \delta_0 R^2/(8\|f^0\|_K^2)$. Hence, using the equality (2.2) again, we get
\[
\|\tilde{g} - g^0\|^2 + 2\lambda/3\|\tilde{f} - f^0\|_K^2 \leq \frac{14}{3} \delta_0 R^2,
\]
which implies that
\[
\left(\|\int_T X(t)(\tilde{f} - f^0)(t)dt + Z^T(\tilde{\gamma} - \gamma^0)\|^2 + 2\lambda/3\|\tilde{f} - f^0\|_K^2\right)^{1/2} \leq \left(\frac{14\delta_0}{3}\right)^{1/2} R. \tag{5.7}
\]
We now turn back to the term (5.5). Since $\|\tilde{\gamma} S_0\|_1 = \|\tilde{\gamma} - \gamma^0\|_1$ by sparsity assumption, adding $\rho\|\tilde{\gamma} - \gamma^0\|_1$ on both sides of (5.5), we have
\[
\|\tilde{\Pi}^T(\tilde{\gamma} - \gamma^0)\|^2 + \|\int_T X(t)(\tilde{f} - f^0)(t)dt + \Pi^T(\tilde{\gamma} - \gamma^0)\|^2 + \rho\|\tilde{\gamma} - \gamma^0\|_1 + \lambda\|\tilde{f}\|_K^2
\leq 3\delta_0 R^2 + \lambda\|f^0\|_K^2 + 2\rho\|\tilde{\gamma} - \gamma^0\|_1 \leq 4\delta_0 R^2 + \|\tilde{\Pi}^T(\tilde{\gamma} - \gamma^0)\|^2 \tag{5.8}
\]
where the last inequality follows from (5.6) and the assumption on $(\lambda, \rho)$. Then, this follows from (5.8) that
\[
\rho\|\tilde{\gamma} - \gamma^0\|_1 \leq 4\delta_0 R^2,
\]
and particularly
\[
\frac{\rho\|\tilde{\gamma} - \gamma^0\|_1}{R\sqrt{\delta_0/7}} \leq 4\sqrt{7\delta_0} R. \tag{5.9}
\]
Consequently, we see from (5.7) and (5.9) that
\[
\tau(\tilde{g} - g^0) \leq 4\sqrt{7\delta_0} R + \sqrt{7\delta_0} R = 5\sqrt{7\delta_0} R \leq \frac{R}{2},
\]
since $\delta_0 \leq 1/700$ by assumption. This immediately implies that
\[
\tau(\tilde{g} - g^0) = 5\tau(\tilde{g} - g^0) = \frac{R\tau(\tilde{g} - g^0)}{R + \tau(\tilde{g} - g^0)} \leq \frac{R}{2}.
\]
Hence, by a simple algebra yields our desired result.
Proof of Theorem 2. We can check that the following chosen quantities for $(R, \rho, \lambda)$ $R \asymp \sqrt{\frac{2n \log p}{n}}$, $\rho \asymp \sqrt{\frac{\log p}{n}}$ and $\lambda \asymp \frac{2n \log p}{n}$ satisfy all the corresponding constraints appearing in Theorem 1, Lemma 5.2 and Lemma 5.4. Then, following the definition of $\tau(\cdot)$ and Theorem 2.2 yields our desired results in Theorem 2.

REFERENCES

[1] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc., 68 (1950), 337–404.
[2] P. J. Bickel, Y. Ritov and A. B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist., 37 (2009), 1705–1732.
[3] O. Bousquet, A Bennett concentration inequality and its application to suprema of empirical processes, C. R. Math. Acad. Sci. Paris, 334 (2002), 495–550.
[4] S. Boyd and L. Vandenberghe, Convex optimization, Cambridge University Press, Cambridge, 2004.
[5] P. Buhlmann and S. van de Geer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer-Verlag Berlin Heidelberg, 2011.
[6] T. T. Cai and P. Hall, Prediction in functional linear regression, Ann. Statist., 34 (2006), 2159–2179.
[7] T. T. Cai and M. Yuan, Minimax and adaptive prediction for functional linear regression, J. Amer. Statist. Assoc., 107 (2012), 1201–1216.
[8] A. Cuevas, M. Febrero and R. Fraiman, Linear functional regression: The case of fixed design and functional response, Canadian J. Statist., 30 (2002), 285–300.
[9] D. Donoho, Compressed sensing, IEEE Trans. Inform. Theory, 52 (2006), 1289–1306.
[10] F. Ferraty and P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer, New York, 2006.
[11] X. Ji, J. W. Han, X. Jiang, X. T. Hu, L. Guo, J. G. Han, L. Shao and T. M. Liu, Analysis of music/speech via integration of audio content and functional brain response, Inform. Sci., 297 (2015), 271–282.
[12] D. H. Kong, J. G. Ibrahim, E. Lee and H. Zhu, FLCRM: Functional linear cox regression model, Biometrics, 74 (2018), 109–117.
[13] D. H. Kong, K. J. Xue, F. Yao And H. H. Zhang, Partially functional linear regression in high dimensions, Biometrika, 103 (2016), 147–159.
[14] M. Ledoux, The Concentration of Measure Phenomenon, Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 2001.
[15] P. Muller and S. Van de Geer, The partial linear model in high dimensions, Scand. J. Statist., 42 (2015), 580–608.
[16] C. Preda and G. Saporta, PLS regression on a stochastic process, Comput. Statist. Data Anal., 48 (2005), 149–158.
[17] J. O. Ramsay and B. W. Silverman, Functional Data Analysis, 2nd edition, Springer, New York, 2005.
[18] E. Sanchez-Lozano, G. Tzimiropoulos, B. Martinez, F. De la Torre and M. Valstar, A functional regression approach to facial landmark tracking, IEEE Trans. Pattern Anal. Mach. Intell., 40 (2018), 2037–2050.
[19] H. Shin and M. H. Lee, On prediction rate in partial functional linear regression, J. Multi. Anal., 103 (2012), 93–106.
[20] M. Talagrand, New concentration inequalities in product spaces, Invent. Math., 126 (1996), 505–563.
[21] Q. G. Tang and L. S. Cheng, Partial functional linear quantile regression, Sci. China Math., 57 (2014), 2589–2608.
[22] J. P. Thivierge, Functional data analysis of cognitive events in EEG, IEEE Int. Confer. Syst. Man Cyber., (2007), 2473–2478.
[23] R. J. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. B, 58 (1996), 267–288.
[24] S. Van. de. Geer, Empirical Processes in M-Estimation, Cambridge University Press, New York, 2000.
[25] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming, IEEE Trans. Inf. Theory, 55 (2009), 2183–2202.
[26] F. Yao, H. G. Muller and J. L. Wang, Functional data analysis for sparse longitudinal data,  
*J. Amer. Statist. Assoc.*, 100 (2005), 577–590.

[27] Y. Yuan and T. T. Cai, A reproducing kernel Hilbert space approach to functional linear  
regression, *Ann. Statist.*, 38 (2006), 3412–3444.

[28] D. D. Yu, L. L. Kong and I. Mizera, Partial functional linear quantile regression for neuroimaging data analysis,  
*Neurocomputing*, 195 (2016), 74–87.

Received January 2019; revised April 2019.

E-mail address: xiayifan@163.com
E-mail address: email2@aimSciences.org
E-mail address: @ece.pdx.edu