Physica Scripta

PAPER

Growing-dimensional partially functional linear models: non-asymptotic optimal prediction error

Huiming Zhang[1,2] **and Xiaoyu Lei**[3]

[1] Institute of Artificial Intelligence, Beihang University, Beijing, People's Republic of China
[2] Zhuhai UM Science & Technology Research Institute, Zhuhai, People's Republic of China
[3] Department of Statistics, University of Chicago, United States of America

E-mail: leixy@uchicago.edu

Keywords: partially functional linear models, sub-Gaussian and Berstein concentration inequalities in Hilbert space, reproducing kernel Hilbert space, non-asymptotic bound, minimax rate, diverging number of covariates

OPEN ACCESS

RECEIVED
5 April 2023

REVISED
7 July 2023

ACCEPTED FOR PUBLICATION
25 July 2023

PUBLISHED
9 August 2023

Abstract

Under the reproducing kernel Hilbert spaces (RKHS), we focus on the penalized least-squares of the partially functional linear models (PFLM), whose predictor contains both functional and traditional multivariate parts, and the multivariate part allows a divergent number of parameters. From the non-asymptotic point of view, we study the rate-optimal upper and lower bounds of the prediction error. An exact upper bound for the excess prediction risk is shown in a non-asymptotic form under a more general assumption known as the effective dimension to the model, by which we also show the prediction consistency when the number of multivariate covariates p slightly increases with the sample size n. Our new finding implies a trade-off between the number of non-functional predictors and the effective dimension of the kernel principal components to ensure prediction consistency in the increasing-dimensional setting. The analysis in our proof hinges on the spectral condition of the sandwich operator of the covariance operator and the reproducing kernel, and on sub-Gaussian and Berstein concentration inequalities for the random elements in Hilbert space. Finally, we derive the non-asymptotic minimax lower bound under the regularity assumption of the Kullback-Leibler divergence of the models.

1. Introduction

Statistical analysis of functional data has become an important and challenging part in modern statistics since the leading work Ramsay (1982) and pioneering paper Grenander (1950). Due to technological innovation, the progress in data storage enables scientists to acquire complex data sets with the structures of curves, images, or other data with functional structures, referred to as functional data. Functional data analysis has a wide range of applications, including chemometrics, econometrics, and biomedical studies Ramsay and Silverman (2007), Kokoszka and Reimherr (2017). There has been a large amount of works now focusing on many different non-parametric aspects of functional data such as kernel ridge regressions Cai and Hall (2006), Preda (2007), Du and Wang (2014), Reimherr *et al* (2018), penalized B-spline regressions Cardot *et al* (2003), functional principal component regressions Yao *et al* (2005), local linear regressions Baíllo and Grané (2009), and reader can refer to the review paper Wang *et al* (2016) for more details.

Many existing works related to the estimation and prediction problems of functional data are based on the framework of functional principal component analysis (FPCA), see Yao *et al* (2005), Cai and Hall (2006), Hall and Horowitz (2007), Zhou *et al* (2023). However, the predictive power of FPCA-based methods is weakened when the functional principal components cannot form an effective basis for the slope function, which often occurs in practice. A similar phenomenon also appears in principal component regressions, see Jolliffe (1982). An alternative method for the functional data is based on the reproducing kernel Hilbert space (RKHS) framework, which assumes the slope function is contained in an RKHS. It is shown in Cai and Yuan (2012) that the RKHS-based method performs better than the FPCA-based method when the slope function does not align

well with the eigenfunctions of the covariance kernel. In fact, FPCA strongly relies on the leading principal scores with large eigenvalues correspondingly, and the eigenfunctions for representing the slope function inevitably lose some information for the response. From the machine learning theory point of view, the FPCA is essentially a non-supervised method that often performs poorly in data analysis. For example, the analysis of Canadian weather data mentioned in Cai and Yuan (2012) and the section 3 of Cui *et al* (2020).

In this paper, we study the partially functional linear models (FPLM) containing both functional and multivariate parts in the predictor, which is originally considered in Shin (2009). Let $\boldsymbol{X} := (X_1,\cdots,X_p)^T$ be a p-dimensional multivariate predictor, $Y(t)$ be a functional predictor, ε be a random noise and Z be the scalar response. In our work, we consider the PFLM taking the semi-parametric form

$$Z = \boldsymbol{X}^T\boldsymbol{\alpha}_0 + \int_{\mathcal{T}} Y(t)\beta_0(t)\mathrm{d}t + \varepsilon, \tag{1}$$

where the $\beta_0(t)$ is the slope function for functional predictor and the $\boldsymbol{\alpha}_0$ is the regression coefficient for multivariate predictor. The model (1) contains both parametric and non-parametric part, which belongs to semi-parametric statistics.

We assume the predictor to be the random design where $\boldsymbol{X}$ and $Y(t)$ are independent. Because the intercepts of predictor are easy to estimate by centralizing, for simplicity, we assume

$$\mathbb{E}\boldsymbol{X} = 0 \text{ and } \mathbb{E}Y(t) = 0.$$

Moreover, we require that the random noise ε has conditional zero mean and finite variance provided the predictor $\boldsymbol{X}$ and Y. For real data, suppose that we collect data $\{(Z_i,\ \boldsymbol{X}_i,\ Y_i(t),\ t \in \mathcal{T})\}_{i=1}^n$ that is i.i.d. (independent and identically distributed) drawn from $(Z, \boldsymbol{X}, Y(t))$.

In some situations, many non-functional predictors are often collected for practical data analysis, and this increasing-dimensional setting has been considered in Aneiros *et al* (2015), Kong *et al* (2016). Moreover, our work can also be applied to deal with divergent number of parameters. Theoretically, this setting requires assuming that the number of scalar covariates grows with the sample size, i.e., $p = p_n \to \infty$, and the convergence rate of the desired estimator becomes totally different from the case where the dimension of the non-functional predictors is fixed.

Dealing with the functional data as a stochastic process is a significant challenge in functional data analysis. Obviously, a functional covariate $Y(t)$ has an infinite number of predictors over the time domain (observed as discrete-time points) that are all highly correlated. The covariance function characterizes the correlation of the functional covariate. The estimation of the slope function in functional regressions is connected to ill-posed inverse problems. To handle the infinite-dimensionality of $\beta_0(t)$, people often impose certain regularity conditions on the hypothesized space of the slope function to ensure that the infinite-dimensional problems are tractable as a finite-dimensional approximation solution. Notwithstanding, the convergence rate of the slope estimators depends directly on the assumptions of the covariance operator's eigenvalue decay and the slope function's restricted space. Thus, the convergence rate cannot be parametric due to the infinite-dimensionality of the model (1).

Some recent developments in PFLM include Zhang and Lian (2019), Zhu *et al* (2019), Cui *et al* (2020). Because of the shortcoming of the FPCA-based methods, we apply penalized least squares under the framework of RKHS here. There are few efforts on the non-asymptotic upper and lower bounds in the existing literature. For FPCA-based method, Brunel *et al* (2016) considers the adaptive estimation procedure of functional linear models under a non-asymptotic framework; Wahl (2018) analyses the prediction error of functional principal component regression (FPCR) and proves a non-asymptotic upper bound for the corresponding squared risk. For the RKHS-based method, many works focus on the asymptotic results, such as Cai and Hall (2006), Cai and Yuan (2012). Under RKHS-based kernel ridge regressions, Liu and Li (2020) recently studies the non-asymptotic RKHS-norm error bounds (called oracle inequalities) for the plug-in KRR estimator of f_0 in Gaussian non-parametric regression $Y = f_0(X) + \varepsilon,\quad \varepsilon \sim N(0,\ \sigma^2)$, where f_0 belongs to L_2. By applying the Matérn kernel and supposing f_0 in a Hölder space with the polynomial decay rate of eigenvalues $\lambda_n = O(n^{-2a})$, Liu and Li (2020) derives the nearly minimax optimal convergence rate $\left(\frac{\log n}{n}\right)^{\frac{a}{2a+1}}$ (up to a $\log n$ factor) for L_2-norm estimation error of the estimation of derivatives using plug-in kernel ridge regression (KRR) estimator.

To analyze the PFLM, the main innovation of our work is that we provide non-asymptotic upper and lower bounds for the excess prediction risk under the assumption of the effective dimension, which is equivalent to assuming the eigenvalues of the sandwich operator decay (see remark 2). Tong and Ng (2018) establishes the upper bound for the excess prediction risk for the RKHS-based slope estimator of the functional linear models, but they do not consider the PFLM. Their result on the upper bound is a special case of our more general result because FLM is a special case of PFLM ($p = 0$). If we let $p = 0$ in the theorem 1 and suppose $0 \cdot \ln 0 = 0$, we obtain the same non-asymptotic upper bound for FLM as the theorem 3.6 in Tong and Ng (2018) up to a constant. We also derive a minimax lower bound for the excess prediction risk under a general assumption

concerning the Kullback-Leibler divergence of the model. In Cai and Yuan (2012), the minimax lower bound is derived for the FLM in an asymptotic sense in theorem 1, which is also a corollary of our result on the non-asymptotic minimax lower bound when $n \to \infty$. See the last paragraph in section 4 for detailed derivation. Moreover, the optimal convergence rate (both upper and lower) of the excess prediction risk of PFLM is the same as that of FLM, which means the convergence of the functional part dominates the convergence of the PFLM.

The specific theoretical contributions of our work are listed as

- A significant contribution is that we obtain the non-asymptotic upper and lower bounds of PFLM, which have not been well studied in the existing literature. We provide an exact non-asymptotic minimax lower bound on the excess prediction risk in PFLM. Moreover, a particular application of the proposed non-asymptotic version of the optimal prediction upper bound is that it allows analyzing the PFLM with a divergent number of non-functional predictors, which leads to the prediction consistency under the setting $p^7 \log^6(p) = o(n)$.
- We derive the non-asymptotic upper bound of the excess prediction risk for the RKHS-based least squares estimation in PFLM, and the obtained optimal bound we obtain is more exact than that of Cui *et al* (2020) which only obtains the stochastic order of the convergence rate without the definite multiplying constants relevant to the high probability events. Our derivation for the optimal bound does not need the *inverse Cauchy-Schwarz inequality*

$$\mathbb{E}\left(\int Y(t)f(t)\,dt\right)^4 \leqslant C\left[\mathbb{E}\left(\int Y(t)f(t)\,dt\right)^2\right]^2, \text{ for } f \in L^2(\mathcal{T})$$

  as a moment assumption of the functional predictor. This condition is imposed in Cui *et al* (2020) and Cai and Yuan (2012) to attain minimax prediction bounds for (partially) functional linear regressions. Our proof does not directly rely on the well-known representation lemma for the smoothing splines; see Wahba (1990) and Cucker and Smale (2001).
- The proof for the theorem 1 is divided into three steps, and it relies on new non-trivial results. First, we prove the difference of the functional part between the true parameter and our least squares estimate is bounded. Second, based on the boundedness, we show the excess prediction risk contributed by the multivariate part of the predictor is convergent at n^{-1}-rate. Finally, according to the convergence of the multivariate part, we obtain the convergence of the prediction risk corresponding to the functional part in $n^{-\frac{1}{1+\theta}}$-rate, where θ is related to the effective dimension in the Assumption 6. Specifically, the novelty of the proof lies in the lemma 2, which is a crucial lemma for the theorem 1. In the lemma 2, to show the concentration property of the random elements in Banach space, we use the methods in functional analysis and convert the random elements in Banach space to other relevant random elements in Hilbert space.

The outline of this paper is constructed as follows. In section 2, we provide the notations and definitions we need and a brief introduction to the RKHS and the PFLM. Section 3 shows our main theorem about the non-asymptotic upper bound for the excess prediction risk and two relevant corollaries. In section 4, we state the minimax lower bound for the excess prediction risk. In section 5.1, we provide the proof of the theorem 1 in section 3. In section 5.2, we show the proof the theorem 2 in section 4. In section 5.3 and 6, we prove the lemmas we need for the proofs in section 5.1 and 5.2. In section 7, we summarize our conclusions and point out some future directions for research.

2. Preliminaries

2.1. Notations and definitions

Define $\|\boldsymbol{v}\|_2 := (\sum_{i=1}^p v_i^2)^{\frac{1}{2}}$ to be the ℓ_2-norm of vector $\boldsymbol{v} \in \mathbb{R}^p$. Let $\mathcal{T} \subset \mathbb{R}$ be a compact set. Denote by $L^2(\mathcal{T})$ the Hilbert space composed by square integrable functions on $\mathcal{T}$, whose inner product and norm are respectively denoted by $\langle f, g\rangle$ and $\|f\|$ for any $f, g \in L^2(\mathcal{T})$.

Consider T a bounded linear operator from a Banach space A to a Banach space B respectively endowed with the norms $\|\cdot\|_A$ and $\|\cdot\|_B$. Define the operator norm of T as

$$\|T\|_{\text{op}} := \sup_{x \in A: \|x\|_A = 1} \|T(x)\|_B.$$

Let T^* be the adjoint of T from B^* to A^* defined by $T^*(f)(x) := f(T(x))$, for any $f \in B^*$. Notice the adjoint of an operator does not change the operator norm, and thus we have $\|T^*\|_{\text{op}} = \|T\|_{\text{op}}$.

For a matrix $E = (e_{ij})_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}$, when writing $\|E\|_{\text{op}}$, we actually view E as a bounded linear operator from $\mathbb{R}^p$ to $\mathbb{R}^p$ endowed with ℓ_2-norm defined by $\boldsymbol{v} \mapsto E\boldsymbol{v}$, which is also called the spectral norm. Let

$\|E\|_\infty := \max_{1\leqslant i,j\leqslant p}|e_{ij}|$ be the ℓ_∞-norm of the matrix E and $\lambda_{\max}(E)$ be the largest eigenvalue of the matrix E. Moreover, we have $\|E\|_{\text{op}} \leqslant p\|E\|_\infty$ from 5.6. P23 in Page 365 of Horn and Johnson (2012).

For a real, symmetric, square integrable on the domain $\mathcal{T}\times\mathcal{T}$ such that $\int_{\mathcal{T}\times\mathcal{T}} R^2(s,t)\mathrm{d}s\mathrm{d}t < \infty$. The nonnegative definite function R: $\mathcal{T}\times\mathcal{T}\to\mathbb{R}$, is called reproducing kernel in the following. Let L_R: $L^2(\mathcal{T})\to L^2(\mathcal{T})$ be an integral operator (also a bounded linear operator) defined by

$$L_R(f)(t) := \langle R(s,t), f(s)\rangle = \int_{\mathcal{T}} R(s,t)f(s)\mathrm{d}s.$$

According to the Hilbert-Schmidt theorem, there exists a set of orthonormalized eigenfunctions $\{\psi_k^R: k\geqslant 1\}$ and a sequence of eigenvalues $\theta_1^R \geqslant \theta_2^R \geqslant \cdots > 0$ such that

$$R(s,t) = \sum_{k=1}^{+\infty} \theta_k^R \psi_k^R(s)\psi_k^R(t), \quad \forall s,t\in\mathcal{T}, \tag{2}$$

see theorem 4.6.5 in Hsing and Eubank (2015) for the proof of such series representation (2).

Noticing the orthonormality of the eigenfunctions $\{\psi_k^R\}_{k\geqslant 1}$, we have

$$L_R(\psi_k^R)(s) = \langle R(s,t), \psi_k^R(t)\rangle = \left\langle \sum_{i=1}^{+\infty}\theta_i^R\psi_i^R(s)\psi_i^R(t), \psi_k^R(t)\right\rangle = \sum_{i=1}^{+\infty}\theta_i^R\psi_i^R(s)\langle\psi_i^R(t),\psi_k^R(t)\rangle = \theta_k^R\psi_k^R(s).$$

In what follows, let $\{(\theta_k^R, \psi_k^R)\}_{k\geqslant 1}$ be the eigenvalue-eigenfunction pairs corresponding to the operator (or the equivalent bivariate function) R. Define $L_R^{\frac{1}{2}}$ as a operator satisfying $L_R^{\frac{1}{2}}(\psi_k^R) = \sqrt{\theta_k^R}\psi_k^R$. For two bivariate functions R_1, R_2: $\mathcal{T}\times\mathcal{T}\to\mathbb{R}$, define

$$(R_1R_2)(s,t) := \langle R_1(s,\cdot), R_2(\cdot,t)\rangle = \int_{\mathcal{T}} R_1(s,u)R_2(u,t)\mathrm{d}u.$$

Then we have the relation $L_{R_1R_2} = L_{R_1}\circ L_{R_2}$, where $\circ$ means the composition of mappings. To show $L_{R_1R_2} = L_{R_1}\circ L_{R_2}$, we notice

$$\begin{aligned} L_{R_1}\circ L_{R_2}(f)(t) &= \int_{\mathcal{T}} R_1(t,s)L_{R_2}(f)(s)\mathrm{d}s = \int_{\mathcal{T}} R_1(t,s)\left(\int_{\mathcal{T}} R_2(s,u)f(u)\mathrm{d}u\right)\mathrm{d}s \\ &= \int_{\mathcal{T}}\left(\int_{\mathcal{T}} R_1(t,s)R_2(s,u)\mathrm{d}s\right)f(u)\mathrm{d}u = L_{R_1R_2}(f)(t). \end{aligned}$$

Let $\mathrm{HS}(\mathcal{T})$ be the Hilbert space of the Hilbert-Schmidt operators on $L^2(\mathcal{T})$ with the inner product $\langle A,B\rangle_H := \mathrm{Tr}(B^*A)$ and the norm $\|A\|_{\text{HS}}^2 = \sum_{k=1}^{+\infty}\|A(\phi_k)\|^2$ where $\{\phi_k\}_{k\geqslant 1}$ is an orthonormal basis of $L^2(\mathcal{T})$. The space $\mathrm{HS}(\mathcal{T})$ is a subspace of the bounded linear operators on $L^2(\mathcal{T})$, with the norm relations $\|A\|_{\text{op}}\leqslant\|A\|_{\text{HS}}$ and $\|AB\|_{\text{HS}}\leqslant\|A\|_{\text{op}}\|B\|_{\text{HS}}$.

Given a reproducing kernel K, we can uniquely identify a RKHS $\mathcal{H}(K)$ composed by a subspace of $L^2(\mathcal{T})$ satisfying $K(t,\cdot)\in\mathcal{H}(K)$ for any $t\in\mathcal{T}$, which is endowed with an inner product $\langle\cdot,\cdot\rangle_K$ such that

$$f(t) = \langle K(t,\cdot), f\rangle_K, \quad \text{for any} \quad f\in\mathcal{H}(K).$$

There is a well-known fact

$$L_K^{\frac{1}{2}}(L^2(\mathcal{T})) = \mathcal{H}(K),$$

i.e. the RKHS $\mathcal{H}(K)$ can be characterized as the range of $L_K^{\frac{1}{2}}$ equipped with the norm $\|L_K^{1/2}(f)\|_K = \|f\|_{L^2(\mathcal{T})}$, see corollary 1 in Sun (2005) for details and extensions. For simplicity, let $\mathcal{H}(K)$ be dense in $L^2(\mathcal{T})$, which means $L_K^{\frac{1}{2}}$ is injective. The definition of $L_K^{\frac{1}{2}}$ directly yields the compactness of $L_K^{\frac{1}{2}}$.

Assumption 1 Assumption on integral operator. Assuming that $\mathcal{T}$ is bounded and compact, the reproducing kernel $K\in\mathcal{K}$ is continuous, where $\mathcal{K}$ a family of continuous kernels (on the compact $\mathcal{T}$). Define κ as

$$\|L_K^{\frac{1}{2}}\|_{\text{op}} \leqslant \kappa < \infty \text{ for } K\in\mathcal{K}. \tag{3}$$

Readers can refer to Wahba (1990), Cucker and Smale (2001), Hsing and Eubank (2015) for more discussions on RKHS.For the i.i.d. data $\{(Z_i, \boldsymbol{X}_i, Y_i(t), t\in\mathcal{T})\}_{i=1}^n$, define the empirical covariance matrix D_n and the covariance matrix D for the multivariate part of the predictor to be

$$D_n := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T \quad \text{and} \quad D := \mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T),$$

where we assume the expectations of random variables $\boldsymbol{X}\boldsymbol{X}^T$ exist. Let $\lambda_{\max} := \lambda_{\max}(D)$ and $\lambda_{\min} := \lambda_{\min}(D)$ be the largest and smallest eigenvalues of the covariance matrix D. Similarly, when the expectation of $Y(s)Y(t)$ exists, define the empirical covariance function $C_n(s,t)$ and the covariance function $C(s,t)$ for the functional part of the predictor to be

$$C_n(s,\,t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} Y_i(s)\,Y_i(t) \quad \text{and} \quad C(s,\,t) \coloneqq \mathbb{E}(Y(s)\,Y(t)).$$

Given the asymmetric, square-integrable, and non-negative definite covariance function $C(s, t)$, define the *sandwich operator* of the covariance operator C and the reproducing kernel K by

$$T \coloneqq L_K^{\frac{1}{2}} \circ L_C \circ L_K^{\frac{1}{2}} \text{ and its empirical version } T_n \coloneqq L_K^{\frac{1}{2}} \circ L_{C_n} \circ L_K^{\frac{1}{2}},$$

see Cai and Yuan (2012) for details. For simplicity of the following technique analysis, define

$$g_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i L_K^{\frac{1}{2}} Y_i \text{ and } a_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \boldsymbol{X}_i,$$

which are key quantities to derive the convergence rate of the desired estimator. Define the bounded linear operators G_n: $L^2(\mathcal{T}) \to \mathbb{R}^p$ and H_n: $\mathbb{R}^p \to L^2(\mathcal{T})$ by

$$G_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \langle Y_i,\, L_K^{\frac{1}{2}} f\rangle \boldsymbol{X}_i, \quad \forall f \in L^2(\mathcal{T}) \quad \text{and} \quad H_n(\boldsymbol{\alpha}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i^T \boldsymbol{\alpha}) L_K^{\frac{1}{2}} Y_i, \quad \forall\ \boldsymbol{\alpha} \in \mathbb{R}^p.$$

Because of the compactness of operator T and the Hilbert-Schmidt theorem on compact operator, see theorem 11.3 in Schechter (2001), there exist a set of eigenvalue-eigenfunction pairs $\{(\tau_k, \varphi_k)\colon k \geqslant 1\}$ such that the operation of T can be decomposed in the following way

$$T(f) = \sum_{k=1}^{\infty} \tau_k \langle f,\, \varphi_k\rangle\, \varphi_k,$$

where $\{\varphi_k\colon k \geqslant 1\}$ is orthonormal basis and $\{\tau_k\colon k \geqslant 1\}$ decrease to 0. Define the trace of the operator $(T + \lambda I)^{-1}T$ as

$$D(\lambda) \coloneqq \mathrm{Tr}((T + \lambda I)^{-1}T),$$

which is also called the effective dimension introduced to measure the convergence rate of the functional part; see Zhang (2005), Caponnetto and De Vito (2007).

2.2. The penalized least square for PFLM

The goal of prediction given the predictor $\boldsymbol{X}$ and $Y(t)$ is to recover the prediction: $\eta_0(\boldsymbol{X},\, Y(t)) \coloneqq \boldsymbol{X}^T\boldsymbol{\alpha}_0 + \int_{\mathcal{T}} Y(t)\beta_0(t)\mathrm{d}t$. i.e., the right side of (1) without the random noise ε. To estimate the true parameter $(\boldsymbol{\alpha}_0, \beta_0)$, the penalized least square is defined as

$$(\hat{\boldsymbol{\alpha}}_n,\, \hat{\beta}_n) \coloneqq \underset{(\boldsymbol{\alpha},\beta)\in\mathbb{R}^p\times\mathcal{H}(K)}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \boldsymbol{X}_i^T\boldsymbol{\alpha} - \int_{\mathcal{T}} Y_i(t)\beta(t)\mathrm{d}t\right)^2 + \lambda_n\|\beta\|_K^2. \tag{4}$$

Noticing $L_K^{\frac{1}{2}}(L^2(\mathcal{T})) = \mathcal{H}(K)$, there exists $\hat{f}_n \in L^2(\mathcal{T})$ such that $L_K^{\frac{1}{2}}(\hat{f}_n) = \hat{\beta}_n$. So the (4) is replaced by

$$(\hat{\boldsymbol{\alpha}}_n,\, \hat{f}_n) \coloneqq \underset{(\boldsymbol{\alpha},f)\in\mathbb{R}^p\times L^2(\mathcal{T})}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Z_i - \boldsymbol{X}_i^T\boldsymbol{\alpha} - \langle Y_i,\, L_K^{\frac{1}{2}} f\rangle)^2 + \lambda_n\|f\|^2. \tag{5}$$

For the Euclidean predictor $\boldsymbol{X}$, we assume the dimension of the multivariate parameter p less than or equal to the number of the training samples n ($p \leqslant n$), by which the empirical covariance matrix D_n is invertible a.s. according to the theorem in Okamoto (1973), if the distribution of $\{\boldsymbol{X}_i\}_{i=1}^n$ is absolutely continuous (with respect to Lebesgue measure).

Assumption 2 Assumption on high-dimensions. We assume D_n and D are positive definite for the given data with $p \leqslant n$.

According to the definition of the penalized least squares $(\hat{\boldsymbol{\alpha}}_n, \hat{f}_n)$, which minimize (5), the difference between the penalized least squares $(\hat{\boldsymbol{\alpha}}_n, \hat{f}_n)$ and the true parameter $(\boldsymbol{\alpha}_0, f_0)$ can be represented as

$$\{\hat{f}_n - f_0 = -\lambda_n(T_n + \lambda_n I)^{-1}f_0 - (T_n + \lambda_n I)^{-1}H_n(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) + (T_n + \lambda_n I)^{-1}g_n, \tag{6}$$

$$\{\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 = -D_n^{-1}G_n(\hat{f}_n - f_0) + D_n^{-1}a_n, \tag{7}$$

where T_n, H_n, G_n, D_n, g_n and a_n are given in the previous section. The derivation of (6) and (7) is left to section 5.3.1, where we use the method of the calculus of variations. Thus the existence of the minimizer of the penalized least squares (5) becomes to find a tuple $(\hat{\boldsymbol{\alpha}}_n, \hat{f}_n)$ satisfying the equations above.

Let $\hat{\eta}_n(\boldsymbol{X},\, Y(t)) \coloneqq \boldsymbol{X}^T\hat{\boldsymbol{\alpha}}_n + \int_{\mathcal{T}} Y(t)\hat{\beta}_n(t)\mathrm{d}t$ be the prediction rule induced by the penalized least square estimator $(\hat{\boldsymbol{\alpha}}_n,\, \hat{\beta}_n)$. For a prediction rule $\eta(\boldsymbol{X}, Y(t))$, define the prediction risk to be

$$\mathcal{E}(\eta) := \mathbb{E}[Z^* - \eta(\boldsymbol{X}^*, Y^*(t))]^2,$$

where $(Z^*, \boldsymbol{X}^*, Y^*(t))$ is an independent copy of $(Z, \boldsymbol{X}, Y(t))$. We measure the accuracy of the prediction $\hat{\eta}_n$ by the excess prediction risk

$$\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) = \mathbb{E}[\hat{\eta}_n(\boldsymbol{X}^*, Y^*(t)) - \eta_0(\boldsymbol{X}^*, Y^*(t))]^2.$$

Let $f_0 \in L^2(\mathcal{T})$ satisfying $L_K^{\frac{1}{2}} f_0 = \beta_0$ and rewrite:

$$\eta_0(\boldsymbol{X}, Y(t)) := \boldsymbol{X}^T\boldsymbol{\alpha}_0 + \int_{\mathcal{T}} Y(t)(L_K^{\frac{1}{2}} f_0)(t)\mathrm{d}t \text{ and } \hat{\eta}_n(\boldsymbol{X}, Y(t)) := \boldsymbol{X}^T\hat{\boldsymbol{\alpha}}_n + \int_{\mathcal{T}} Y(t)(L_K^{\frac{1}{2}} \hat{f}_n)(t)\mathrm{d}t,$$

by which we can bound the excess prediction risk

$$\begin{aligned}
\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) &= \mathbb{E}\left[\boldsymbol{X}^{*T}(\boldsymbol{\alpha}_0 - \hat{\boldsymbol{\alpha}}_n) + \int_{\mathcal{T}} Y^*(t)(L_K^{\frac{1}{2}}(f_0 - \hat{f}_n))(t)\mathrm{d}t\right]^2 \\
&\leqslant 2\mathbb{E}[\boldsymbol{X}^{*T}(\boldsymbol{\alpha}_0 - \hat{\boldsymbol{\alpha}}_n)]^2 + 2\mathbb{E}\left[\int_{\mathcal{T}} Y^*(t)(L_K^{\frac{1}{2}}(f_0 - \hat{f}_n))(t)\mathrm{d}t\right]^2 \\
&= 2(\boldsymbol{\alpha}_0 - \hat{\boldsymbol{\alpha}}_n)^T\mathbb{E}(\boldsymbol{X}^*\boldsymbol{X}^{*T})(\boldsymbol{\alpha}_0 - \hat{\boldsymbol{\alpha}}_n) + 2\iint_{\mathcal{T}\times\mathcal{T}} \mathbb{E}[Y^*(t)Y^*(s)](L_K^{\frac{1}{2}}(f_0 - \hat{f}_n))(t)(L_K^{\frac{1}{2}}(f_0 - \hat{f}_n))(s)\mathrm{d}t\mathrm{d}s.
\end{aligned}$$

Notice the equality

$$\mathbb{E}\left[\int_{\mathcal{T}} Y(t)f(t)\mathrm{d}t\right]^2 = \iint_{\mathcal{T}\times\mathcal{T}} \mathbb{E}[Y(s)Y(t)]f(s)f(t)\mathrm{d}s\mathrm{d}t = \int_{\mathcal{T}} f(t)\left(\int_{\mathcal{T}} C(s,t)f(s)\mathrm{d}s\right)\mathrm{d}t = \langle f, L_C f\rangle. \tag{8}$$

With the definitions above, we reformulate the upper bound for the excess prediction risk to

$$\begin{aligned}
\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) &\leqslant 2\lambda_{\max}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|_2^2 + 2\langle L_K^{\frac{1}{2}}(\hat{f}_n - f_0), L_C L_K^{\frac{1}{2}}(\hat{f}_n - f_0)\rangle \\
&= 2\lambda_{\max}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|_2^2 + 2\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\|^2,
\end{aligned} \tag{9}$$

which is relatively easy to analyze.

3. The analysis and main results

3.1. The analysis

Based on the RKHS framework, more regularity assumptions are needed to ensure our main results. First, Assumptions 3–5 are on the moment condition of data. A centered random variable X is called sub-Gaussian if

$$\mathbb{E}e^{tX} \leqslant e^{t^2\sigma^2/2}, \quad \forall \ t \in \mathbb{R},$$

where the quantity $\sigma^2 > 0$ is named as the sub-Gaussian *variance proxy* [see Zhang and Chen 2021, denote as $X \sim \mathrm{subG}(\sigma^2)$]. Let $\|X\|_G = \sup_{k\geqslant 1}\left[\frac{\mathbb{E}X^{2k}}{(2k-1)!!}\right]^{1/(2k)}$ be the sub-Gaussian norm, and X is sub-Gaussian if $\|X\|_G < \infty$; see Buldygin and Kozachenko (2000).

Assumption 3. We assume $\{X_j\}_{j=1}^p$ satisfy the sub-Gaussian growth of moments condition, i.e.

$$\max_{1\leqslant j\leqslant p} \|X_j\|_G \leqslant M_1 < \infty \tag{10}$$

and $v^2 := \max_{1\leqslant j\leqslant p}\mathbb{E}|X_j|^2 < \infty$.

Assumption 4. $Y(t)$ is a bounded square integrable stochastic process: there exists $M_2 > 0$ such that $\|Y(\cdot)\|_{L^2(\mathcal{T})} \leqslant M_2$ (a.s.).

Assumption 5. The random noise ε has conditional zero mean and finite variance: $\mathbb{E}(\varepsilon|\boldsymbol{X}, Y) = 0$ and $\mathbb{E}(\varepsilon^2|\boldsymbol{X}, Y) < \sigma^2$ given $\boldsymbol{X}$ and Y.

Assumption 6 Assumption on effective dimensions. The effective dimension of T satisfies

$$D(\lambda) := \mathrm{tr}((T + \lambda I)^{-1}T) \leqslant c\lambda^{-\theta} \text{ for constants } c > 0 \text{ and } 0 < \theta \leqslant 1.$$

The sub-Gaussian data Assumption 3 has been adopted in many high-dimensional statistics references; see Zhang and Chen (2021) for a review. We assume that $Y(\cdot)$ is bounded in L^2 norm a.s. in the Assumption 4. The Assumption 5 follows the general assumptions of conditional zero mean and finite variance on random noise given the observations $\boldsymbol{X}$ and $Y(t)$. The Assumption 6 on the effective dimension has been adopted in Tong and Ng (2018), which reflects the convergence of eigenvalues of L_C and L_K and how their eigenfunctions align. The Assumption 6 is also equal to the common assumption on the decay rate of the eigenvalues of T (see remark 2).

Define a random variable ξ taking values in $\mathrm{HS}(\mathcal{T})$ by

$$\xi(f) := (T + \lambda_n I)^{-\frac{1}{2}} \langle L_K^{\frac{1}{2}} Y, f \rangle L_K^{\frac{1}{2}} Y.$$

Actually, we can weaken the boundedness in Assumption 4 by assuming the Bernstein's growth of moments condition of ξ in $\mathrm{HS}(\mathcal{T})$: there exist $\tilde{M}$, $\tilde{\nu} > 0$ such that

$$\mathbb{E}(\|\xi - \mathbb{E}\xi\|_{\mathrm{HS}}^l) \leqslant \frac{D(\lambda_n)\tilde{M}^2}{2} \tilde{\nu}^{l-2} l!, \quad \text{for all integer } l \geqslant 2. \tag{11}$$

Then using the lemma 6, we obtain a similar result as in the lemma 1, by which we have an analogous result for the non-asymptotic optimal prediction error as in the theorem 1. But the condition (11) is challenging to verify, and a similar situation is also provided for the FPCA method in H2 of Brunel *et al* (2016). Here, we do not offer the complete proof under the condition (11).

Before getting to our main results, we need two important lemmas, of which the proofs are left to section 5.3 and 6. The following lemma 1 and lemma 2 can be viewed as the concentration inequalities for the operator-valued random variables T_n, G_n and H_n. The concentration inequalities for the random variable taking values in Hilbert space, as stated in the lemmas 6 and 7 play an important role in the proofs of lemmas.

Lemma 1. *Under the Assumption 4, for any $\delta_1 \in (0, 2e^{-1})$, with probability at least $1 - \delta_1$, we have*

$$\|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}} \leqslant c_1 \log\left(\frac{2}{\delta_1}\right) B_n, \text{ where } c_1 := \kappa M_2 \text{ and } B_n := \frac{2c_1}{n\sqrt{\lambda_n}} + \sqrt{\frac{2D(\lambda_n)}{n \log(2/\delta_1)}}.$$

Lemma 2 $\|G_n\|_{\mathrm{op}} = \|H_n\|_{\mathrm{op}}$. *Under the Assumptions 1, 3 and 4, for any $\delta_2 \in (0, 1)$, with probability at least $1 - \delta_2$, we have*

$$\|G_n\|_{\mathrm{op}} = \|H_n\|_{\mathrm{op}} \leqslant \frac{c_2}{\sqrt{n}},$$

where $c_2 =: c_1 p[\nu + 8M_1 \log^{1/2}(p/\delta_2)]$ and ν is constant in Assumption 3.

3.2. Main results

With all the preparations above, we can state this paper's main result. The following theorem provides a non-asymptotic upper bound for the excess prediction risk.

Theorem 1. *Under the Assumptions 1- 6, for any δ_1, δ_3, δ_4, $\delta_5 \in (0, 1)$, $\delta_2 \in (0, 2e^{-1})$, let $\lambda_n := \omega n^{-\frac{1}{1+\theta}}$, $\omega > 0$, $N_1 := 24p \|D^{-1}\|_{\mathrm{op}}(48p \|D^{-1}\|_{\mathrm{op}} M_1^4 + M_1^2) \log\left(\frac{2p^2}{\delta_5}\right)$ and*

$$N_2 = \left(2\kappa^2 M_2^2 [\nu + 8M_1 \log^{1/2}(p/\delta_2)]^2 \frac{3\|D^{-1}\|_{\mathrm{op}}}{2\omega}\right)^{\frac{1+\theta}{\theta}},$$

such that for $n \geqslant n_0 := \lceil \max\{N_1, N_2\} \rceil$, we have with probability at least $1 - \sum_{i=1}^5 \delta_i$

$$\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) \leqslant (2\lambda_{\max}(2c_4c_6 + c_5)^2 + 2c_9^2) n^{-1} + (4(c_7 + c_8)c_9\sqrt{\omega}) n^{-\frac{2+\theta}{2+2\theta}} + (2(c_7 + c_8)^2 \omega) n^{-\frac{1}{1+\theta}}, \tag{12}$$

where $\{c_i\}_{i=4}^9$ are specific constants given in the proof that depend on the true parameters and the assumptions, and can be written as

$$c_4 := p\kappa M_2 [\nu + 8M_1 \log^{1/2}(p/\delta_2)] \frac{3\|D^{-1}\|_{\mathrm{op}}}{2}, \; c_5 = \frac{3\sqrt{p}\sigma\nu\|D^{-1}\|_{\mathrm{op}}}{2\sqrt{\delta_4}},$$

$$c_6 := \|f_0\| + \frac{c_2 c_5}{\omega} + \frac{\sigma}{\sqrt{\delta_3}}\left(2c_1\omega^{-1} + \omega^{-\frac{1+\theta}{2}}\sqrt{\frac{2c}{\log(2/\delta_1)}}\right)$$

$$c_7 := \|f_0\|\left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right) + \omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right) + 1\right]$$

$$c_8 := \left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right) + \omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right) + 1\right]^2 \frac{\sigma(\omega^{-\frac{1+\theta}{2}}\sqrt{c})}{\sqrt{\delta_3}}. \text{ and}$$

$$c_9 := \frac{\kappa M_2(2c_4c_6 + c_5)}{\sqrt{\omega}}\left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right) + \omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right) + 1\right]\left(\nu + 8M_1\log^{1/2}\left(\frac{p}{\delta_2}\right)\right)$$

where c is constant in Assumption 6, and $c_1 := \kappa M_2$ in lemma 1.

Equation (12) presents an exact upper bound of the excess perdition risk with all precise constants determined by the regularity conditions. The first term on the right side of (12) is ascribed to the parametric part of the PFLM. The second term is a mixed bound consisting of both the parametric and the functional part since the prediction risk is a square function composed of both the functional and non-functional predictors. The last term is a dominated term, which reveals that the signal strength $||f_0||$, the operator norm of the reproducing

kernel, and the variation of functional predictor play a crucial role in the non-asymptotic upper bound of prediction risk. Unfortunately, these assumption-dependent constants are always ignored in most references of asymptotic analysis for functional regressions.

Remark 1. At first glance, the upper bound on the excess risk is independent of α_0. From (7),

$$\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\langle Y_i, L_K^{\frac{1}{2}}(\hat{f}_n - f_0)\rangle\boldsymbol{X}_i + \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i\right),$$

and thus $\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0$ has no relation to $\boldsymbol{\alpha}_0$. We illustrate this fact by setting $\beta_0(t) = 0$ in our model (1). Then the PFLM degenerates to the classical linear models with a divergence number of Euclidean predictors. We have $\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)^{-1}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i\right)$, which is free of $\boldsymbol{\alpha}_0$.

From the proof of the theorem 1, one can obtain the non-asymptotic upper bounds of functional and the non-functional parameters regressions below.

Corollary 1. *Under the conditions in theorem 1, for $n > n_0$ we have,*

$$P\left(\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|_2 \leqslant \frac{2c_4c_6 + c_5}{\sqrt{n}}\right) \geqslant 1 - \sum_{i=2}^{5}\delta_i, \tag{13}$$

$$P\left(\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\| \leqslant (c_7 + c_8)\sqrt{\lambda_n} + \frac{c_9}{\sqrt{n}}\right) \geqslant 1 - \sum_{i=1}^{5}\delta_i. \tag{14}$$

It should be noted that we are unable to show the asymptotic normality of the non-functional parameters $\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0$ because it is influenced by the functional parameter $\hat{f}_n - f_0$ as shown in (7). And it is difficult to derive an analogy of the central limit theorem for the functional parameter.

The (13) and (14) in corollary 1 are useful high-probability events, which can be used to obtain the confidence balls for $\boldsymbol{\alpha}_0$ and f_0 under the distance $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|^2$ and $\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\|^2$. They are also helpful for constructing testing statistics, and thus they conceive non-asymptotic hypothesis testing for functional regressions, see Yang *et al* (2020) for the case of non-parametric regressions.

Another corollary of the theorem 1 is the excess prediction risk $\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) = O_p(n^{-\frac{1}{1+\theta}})$. From the proof, we notice the convergence rate of the prediction risk contributed by the multivariate part of the predictor is $O_p(n^{-1})$, faster than the convergence rate corresponding to the functional part of the predictor, which is $O_p(n^{-\frac{1}{1+\theta}})$. Therefore, the convergence rate of the prediction risk of the partially functional linear model is the same as the optimal rate for the functional linear model Cai and Yuan (2012).

Remark 2. The assumption that the eigenvalues $\{\tau_k\}_{k\geqslant 1}$ decay as $\tau_k \leqslant c'k^{-2r}(r > \frac{1}{2})$ is equivalent to our assumption 6. For one direction, see the following derivation.

$$\begin{aligned} D(\lambda_n) &= \sum_{k=1}^{+\infty}\frac{\tau_k}{\tau_k + \lambda_n} \leqslant \sum_{k=1}^{+\infty}\frac{c'k^{-2r}}{c'k^{-2r} + \lambda_n} = \sum_{k=1}^{+\infty}\frac{c'}{c' + \lambda_n k^{2r}} \leqslant \int_0^{+\infty}\frac{c'}{c' + \lambda_n t^{2r}}\mathrm{d}t \\ &= \lambda_n^{-\frac{1}{2r}}\int_0^{+\infty}\frac{c'}{c' + s^{2r}}\mathrm{d}s \lesssim \lambda_n^{-\frac{1}{2r}} \asymp n^{\frac{1}{1+2r}}, \end{aligned}$$

where $s = \lambda_n^{\frac{1}{2r}}t$ and $\lambda_n = \omega n^{-\frac{2r}{1+2r}}$.

Corollary 2. *Suppose the Assumptions 1–5 are satisfied. Assume the eigenvalues τ_k decay as $\tau_k \leqslant c'k^{-2r}$ for some $c' > 0$ and $r > \frac{1}{2}$. For any δ_i that: $\delta_1, \delta_3, \delta_4, \delta_5 \in (0, 1)$ and $\delta_2 \in (0, 2e^{-1})$, by taking $\lambda_n = \omega n^{-\frac{2r}{1+2r}}$, there exists an integer n_0 such that for $n > n_0$, we have with probability at least $1 - \sum_{i=1}^{5}\delta_i$*

$$\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) \leqslant (2\lambda_{\max}(2c_4c_6 + c_5)^2 + 2c_9^2)n^{-1} + (4(c_7 + c_8)c_9\sqrt{\omega})n^{-\frac{1+4r}{2+4r}} + (2(c_7 + c_8)^2\omega)n^{-\frac{2r}{1+2r}},$$

where $c_i(4 \leqslant i \leqslant 9)$ and n_0 are the same as those of the theorem 1 except replacing θ by $\frac{1}{2r}$ and c by a constant relevant to c' and r.

A valuable and insightful application of the theorem 1 is that we can consider the situation where the number of multivariate covariates p increases as a function of n. We need the following assumptions in the increasing dimension background.

Assumption 7. The number of the multivariate covariates $p = p_n$ can increase as a function of n.

Assumption 8. The largest and smallest eigenvalues of D are bounded from below and above as n increases: there exist positive constants m' and M' such that $m' < \lambda_{\min}(D) < \lambda_{\max}(D) < M'$ for all n and p.

According to the definition of $c_i (4 \leqslant i \leqslant 9)$ and $N_i (i = 1, 2)$ in the theorem 1 and the Assumption 8 on the eigenvalues of D, we have the following estimation on the asymptotic order of each coefficients in the theorem 1.

$$c_4 = O(p\log(p)), \quad c_5 = O(p^{\frac{1}{2}}), \quad c_6 = O(p^{\frac{3}{2}}\log(p)), \quad c_7 = c_8 = O(1),$$
$$c_9 = O(p^{\frac{7}{2}}\log^3(p)), \quad N_1 = O(p^2\log(p)) \quad \text{and} \quad N_2 = O(p^{\frac{2(1+\theta)}{\theta}}\log^{\frac{4(1+\theta)}{\theta}}(p)).$$

From these orders, it implies

$$(2\lambda_{\max}(2c_4c_6 + c_5)^2 + 2c_9^2)n^{-1} = O(p^7\log^6(p)n^{-1}),$$

$$(4(c_7 + c_8)c_9\sqrt{\omega})n^{-\frac{2+\theta}{2+2\theta}} = O(p^{\frac{7}{2}}\log^3(p)n^{-\frac{2+\theta}{2+2\theta}})$$

and

$$(2(c_7 + c_8)^2\omega)n^{-\frac{1}{1+\theta}} = O(n^{-\frac{1}{1+\theta}}).$$

We write $n \gg f(p)$ when $f(p)n^{-1} \to 0$. Let $p^7\log^6(p)n^{-1} \to 0$, i.e. $n \gg O(p^7\log^6(p))$, we have as n, $p \to \infty$,

$$p^{\frac{7}{2}}\log^3(p)n^{-\frac{2+\theta}{2+2\theta}} \ll O(n^{\frac{1}{2}-\frac{2+\theta}{2+2\theta}}) \to 0,$$

from which we see the upper bound in the theorem 1 converges to 0.

To apply the theorem 1 in the increasing dimension background, except the convergence of the upper bound, we also need the condition $n > N_1$ and $n > N_2$ satisfied as $n, p \to \infty$. Notice $n \gg O(p^7\log^6(p)) \gg O(p^2\log(p)) = N_1$. If we let $\frac{2(1+\theta)}{\theta} < 7 \Leftrightarrow \theta > \frac{2}{5}$, we have $n \gg N_2$ under the condition $n \gg O(p^7\log^6(p))$ after noticing $p \gg \log^{\varepsilon}(p)$ for any $\varepsilon \in \mathbb{R}$ and the asymptotic order of $N_2 = O(p^{\frac{2(1+\theta)}{\theta}}\log^{\frac{4(1+\theta)}{\theta}}(p))$. Therefore we have the following prediction consistency for the increasing dimension situation of non-functional parameters.

Corollary 3. *Under the Assumptions 1- 8, if the constant $\frac{2}{5} < \theta \leqslant 1$ in the Assumption 6 and $p^7\log^6(p) = o(n)$ in the Assumption 7, we have the consistency for the excess prediction risk:*

$$\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) = o_p(1).$$

Remark 3. If we assume the eigenvalues τ_k decay as $\tau_k \leqslant c'k^{-2r} (r > \frac{1}{2})$ in the increasing-dimensional setting, by applying the corollary 3 and noticing $\theta = \frac{1}{2r}$, we need to further assume $r < \frac{5}{4}$ to obtain the prediction consistency, which means the convergence rate of eigenvalues can not be too fast. Intuitively, when p increases, r can not be too large or equivalently, the effective dimension $D(\lambda_n) \asymp n^{\frac{1}{1+2r}}$ can not be too small. It implies we need to find a trade-off between the number of non-functional predictors and the effective dimension to get the prediction consistency.

The prediction consistency theory has been well-established for non-parametric and high-dimensional statistics; see Zhuang and Lederer (2018) for the recent development of general regularized maximum likelihood estimators. However, their works mainly aim for non-parametric or high-dimensional models and do not cover the semi-parametric case as studied in our paper.

4. Minimax lower bound

In this section, we derive a minimax lower bound for the excess prediction risk in the following theorem 2 when $p \to \infty$. If p is fixed, the minimax lower bound is postponed at the end of section 6.8. To verify the optimality of the upper bound of the prediction risk for the proposed estimator, the result on the minimax lower bound below shows the prediction risk of our estimator achieves the theoretical lower bound caused by the intrinsic limitation of the PFLM. Let P_{α_0,β_0} be the probability taken over the space $(Z, \boldsymbol{X}, Y)$ where Z is generated by the true parameter $Z = \boldsymbol{X}^T\boldsymbol{\alpha}_0 + \langle Y, \beta_0\rangle + \varepsilon$. Before stating the main result, we need a regularity assumption relevant to the Kullback-Leibler distance of the random noises.

Assumption 9. For different $\beta_1, \beta_2 \in \mathcal{H}(K)$ and $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathbb{R}^p$. We assume that the Kullback-Leibler distance between P_{α_1,β_1} and P_{α_2,β_2} can be bounded by

$$K(P_{\boldsymbol{\alpha}_1,\beta_1}|P_{\boldsymbol{\alpha}_2,\beta_2}) = \mathbb{E}_{\boldsymbol{\alpha}_1,\beta_1}\left(\log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_1,\beta_1}}{\mathrm{d}P_{\boldsymbol{\alpha}_2,\beta_2}}\right)\right) \leqslant K_{\sigma^2}\mathbb{E}[\langle Y,\ \beta_1-\beta_2\rangle + \boldsymbol{X}^T(\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2)]^2,$$

where $K_{\sigma^2} > 0$ is a variance-dependent constant and $\mathbb{E}_{\boldsymbol{\alpha}_1,\beta_1}$ means the expectation is taken over $P_{\boldsymbol{\alpha}_1,\beta_1}$.

The examples of constant K_{σ^2} include noises of exponential families (see Du and Wang 2014, Abramovich and Grinshtein 2016) and noises with self-concordant log-density function (see Ostrovskii and Bach 2021). If we assume the random noise $\varepsilon \sim N(0,\sigma^2)$, thus the constant

$$K_{\sigma^2} = \frac{1}{2\sigma^2}. \tag{15}$$

The proof is left to section 6.7. Now we state the main theorem of this section, of which the proof is left to section 5.2.

Theorem 2. *Under the Assumptions 5 and 9 with* $\mathbb{E}\boldsymbol{X} = 0$, *suppose the eigenvalues* $\{\tau_k\}_{k\geqslant 1}$ *of the operator* T *decay as* $\tau_k = t_0 k^{-2r}$ *for some* $r,\ t_0 \in (0,\ \infty)$, *then for* $\rho \in \left(0,\ \frac{1}{8}\right)$, *there exists a sequence* $\{N_n\}_{n\geqslant 1}$ *satisfying*

$$\log N_n \geqslant \left[\left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{\frac{1}{1+2r}} n^{\frac{1}{1+2r}} + \left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{\frac{2r}{1+2r}} n^{\frac{2r}{1+2r}}\right]\frac{\log 2}{8}$$

such that when $n \geqslant \frac{\rho\log 2}{8t_0K_{\sigma^2}}$ *and* $p = O(n^{\frac{2r}{1+2r}}) < n$, *the excess prediction risk satisfies*

$$\inf_{\tilde{\eta}}\sup_{\eta_0\in\mathbb{R}^p\times\mathcal{H}(K)} P\left(\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant \left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{-\frac{2r}{1+2r}} \cdot \frac{2^{-2r}t_0+\lambda_{\min}(D)}{2^{2r+3}n^{\frac{2r}{1+2r}}}\right)$$
$$\geqslant \frac{\sqrt{N_n}}{1+\sqrt{N_n}}\left(1-4\rho-\sqrt{\frac{4\rho}{\log N_n}}\right),$$

where we identify the prediction rule $\tilde{\eta}$ *as a arbitrary estimator* $(\tilde{\boldsymbol{\alpha}},\ \tilde{\beta})$ *based on the training samples* $\{(Z_i,\ \boldsymbol{X}_i,\ Y_i)\}_{i=1}^n$, *and view* η_0 *as the true parameter* $(\boldsymbol{\alpha}_0,\ \beta_0) \in \mathbb{R}^p \times \mathcal{H}(K)$. *We emphasize the probability* P *is taken over the product space of training samples* $\{(Z_i,\ \boldsymbol{X}_i,\ Y_i)\}_{i=1}^n$ *generated by* $\eta_0 = (\boldsymbol{\alpha}_0,\ \beta_0)$.

In the existing literature, most results about the minimax bound are in the asymptotic sense, while the constants in our result are precise and specified. Letting $n \to \infty$ under the Assumption 8, the lower bound inequality in the theorem 2 implies as $N_n \to \infty$

$$\liminf_{n\to\infty}\inf_{\tilde{\eta}}\sup_{\eta_0\in\mathbb{R}^p\times\mathcal{H}(K)} P(\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant b'\rho^{\frac{2r}{1+2r}}n^{-\frac{2r}{1+2r}}) \geqslant 1-2\rho$$

for constant $b' > 0$, by which we get the asymptotic minimax lower bound:

$$\lim_{a\to 0}\liminf_{n\to\infty}\inf_{\tilde{\eta}}\sup_{\eta_0\in\mathbb{R}^p\times\mathcal{H}(K)} P(\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant an^{-\frac{2r}{1+2r}}) = 1.$$

5. Proof of theorems and key Lemmas

5.1. Proof of the theorem 1

To obtain the upper bound for the excess prediction risk, we resort to (9), and then use the representation (6) and (7). The lemmas 1 and 3- 2 are applied to study the error bounds between the penalized least squares $(\hat{\boldsymbol{\alpha}}_n,\ \hat{f}_n)$ and the true parameter $(\boldsymbol{\alpha}_0, f_0)$.

In lemma 1, when setting $\lambda_n = \omega n^{-\frac{1}{1+\theta}}$, we have $\frac{1}{n} < n^{-\frac{1}{1+\theta}} = \frac{\lambda_n}{\omega} = \frac{\sqrt{\lambda_n}}{\omega}\sqrt{\lambda_n}$ and

$$\frac{D(\lambda_n)}{n} \leqslant \frac{c(\omega n^{-\frac{1}{1+\theta}})^{-\theta}}{n} = c\omega^{-\theta}n^{-\frac{1}{1+\theta}} = c\omega^{-(1+\theta)}\lambda_n, \tag{16}$$

by which we have

$$B_n = \frac{2c_1}{n\sqrt{\lambda_n}} + \sqrt{\frac{2D(\lambda_n)}{n\log(2/\delta_1)}} \leqslant \left(2c_1\omega^{-1} + \omega^{-\frac{1+\theta}{2}}\sqrt{\frac{2c}{\log(2/\delta_1)}}\right)\sqrt{\lambda_n} \tag{17}$$

and $n\lambda_n = \omega n^{\frac{\theta}{1+\theta}} \geqslant \omega$.

Applying lemmas 2, 3 and 5 to (7), when

$$n > 24p\,\|D^{-1}\|_{\mathrm{op}}(48p\,\|D^{-1}\|_{\mathrm{op}}M_1^4 + M_1^2)\log\left(\frac{2p^2}{\delta_5}\right) \text{ in Lemma 5}$$

we have with probability at least $1-\delta_2-\delta_4-\delta_5$

$$\begin{aligned}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|_2 &\leqslant \|D_n^{-1}\|_{\mathrm{op}}\|G_n\|_{\mathrm{op}}\|\hat{f}_n - f_0\| + \|D_n^{-1}\|_{\mathrm{op}}\|a_n\| \\ &\leqslant \frac{3}{2}\,\|D^{-1}\|_{\mathrm{op}}\frac{p\kappa M_2[\nu + 8M_1\log^{1/2}(p/\delta_2)]}{\sqrt{n}}\|\hat{f}_n - f_0\| + \frac{3}{2}\,\|D^{-1}\|_{\mathrm{op}}\frac{c_3}{\sqrt{\delta_4}\sqrt{n}} = \frac{c_4}{\sqrt{n}}\|\hat{f}_n - f_0\| + \frac{c_5}{\sqrt{n}}, \quad (18)\end{aligned}$$

where we let $c_4 := p\kappa M_2[\nu + 8M_1\log^{1/2}(p/\delta_2)]\frac{3\,\|D^{-1}\|_{\mathrm{op}}}{2}$ and $c_5 := \frac{3c_3\,\|D^{-1}\|_{\mathrm{op}}}{2\sqrt{\delta_4}}$ with $c_3 := \sqrt{p}\,\sigma\nu$.

First, we give a crude bound for $\|\hat{f}_n - f_0\|$ in (18). According to (6), it gives

$$\begin{aligned}\|\hat{f}_n - f_0\| &\leqslant \lambda_n\,\|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}}\,\|f_0\| + \|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}}\|H_n\|_{\mathrm{op}}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| + \|(T_n + \lambda_n I)^{-1}g_n\| \\ &=: I_1 + I_2 + I_3.\end{aligned}$$

For the term I_1, we have

$$I_1 \leqslant \|f_0\|$$

because $\|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_n}$.

For the term I_2, using $\|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_n}$, $\|H_n\|_{\mathrm{op}} \leqslant \frac{c_2}{\sqrt{n}}$ in lemma 2 and (18) we have with probability at least $1-\delta_2-\delta_4-\delta_5$

$$\begin{aligned}I_2 =: \|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}}\|H_n\|_{\mathrm{op}}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| &\leqslant \frac{1}{\lambda_n}\frac{c_2}{\sqrt{n}}\frac{c_4}{\sqrt{n}}\|\hat{f}_n - f_0\| + \frac{1}{\lambda_n}\frac{c_2}{\sqrt{n}}\frac{c_5}{\sqrt{n}} \\ = \frac{c_2c_4}{n\lambda_n}\|\hat{f}_n - f_0\| + \frac{c_2c_5}{n\lambda_n} &\leqslant \frac{c_2c_4}{n\lambda_n}\|\hat{f}_n - f_0\| + \frac{c_2c_5}{\omega}, \text{ by using } n\lambda_n \geqslant \omega.\end{aligned}$$

For the term I_3, we obtain with probability at least $1-\delta_3$ by lemma 8 and $\|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_n}$

$$I_3 \leqslant \|(T_n + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}}\|(T_n + \lambda_n I)^{-\frac{1}{2}}g_n\| \leqslant \frac{1}{\sqrt{\lambda_n}}\frac{\sigma}{\sqrt{\delta_3}}B_n \leqslant \frac{\sigma}{\sqrt{\delta_3}}\left(2c_1\omega^{-1} + \omega^{-\frac{1+\theta}{2}}\sqrt{\frac{2c}{\log(2/\delta_1)}}\right),$$

where we use (17) in the last step.

Thus we can bound $\|\hat{f}_n - f_0\|$ by I_1, I_2 and I_3

$$\begin{aligned}\|\hat{f}_n - f_0\| &\leqslant \|f_0\| + \frac{c_2c_4}{n\lambda_n}\|\hat{f}_n - f_0\| + \frac{c_2c_5}{\omega} + \frac{\sigma}{\sqrt{\delta_3}}\left(2c_1\omega^{-1} + \omega^{-\frac{1+\theta}{2}}\sqrt{\frac{2c}{\log(2/\delta_1)}}\right) \\ &= \frac{c_2c_4}{n\lambda_n}\|\hat{f}_n - f_0\| + c_6,\end{aligned}$$

where $c_6 := \|f_0\| + \frac{c_2c_5}{\omega} + \frac{\sigma}{\sqrt{\delta_3}}\left(2c_1\omega^{-1} + \omega^{-\frac{1+\theta}{2}}\sqrt{\frac{2c}{\log(2/\delta_1)}}\right)$.

Notice when $\lambda_n = \omega n^{-\frac{1}{1+\theta}}$ and $n\lambda_n = \omega n^{\frac{\theta}{1+\theta}}$, and we have

$$\frac{c_2c_4}{n\lambda_n} \leqslant \frac{1}{2} \text{ for } n > \left(\frac{2c_2c_4}{\omega}\right)^{\frac{1+\theta}{\theta}} = \left(2\kappa^2M_2^2[\nu + 8M_1\log^{1/2}(p/\delta_2)]^2\,\frac{3\,\|D^{-1}\|_{\mathrm{op}}}{2\omega}\right)^{\frac{1+\theta}{\theta}}.$$

Therefore, when $n > n_0$, we obtain with probability at least $1-\delta_2-\delta_3-\delta_4-\delta_5$

$$\tfrac{1}{2}\|\hat{f}_n - f_0\| \leqslant \|\hat{f}_n - f_0\| - \tfrac{c_2c_4}{n\lambda_n}\|\hat{f}_n - f_0\| \leqslant c_6, \text{ which implies} \|\hat{f}_n - f_0\| \leqslant 2c_6.$$

Second, we turn to bound $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|$. By (18), we have with probability at least $1-\delta_2-\delta_3-\delta_4-\delta_5$,

$$\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| \leqslant \frac{2c_4c_6 + c_5}{\sqrt{n}}. \quad (19)$$

Then we find a way to bound $\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\|$. According to (6), we have

$$\begin{aligned}\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\| &\leqslant \lambda_n\,\|T^{\frac{1}{2}}(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}}\,\|f_0\| + \|T^{\frac{1}{2}}(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}}\|H_n\|_{\mathrm{op}}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| \\ &\quad + \|T^{\frac{1}{2}}(T_n + \lambda_n I)^{-1}g_n\| =: E_1 + E_2 + E_3.\end{aligned}$$

For the term E_1, we have with probability at least $1-\delta_1$ by Inequality 2 and lemma 1

$$\begin{aligned}
E_1 &\leqslant \lambda_n \|(T+\lambda_n I)^{\frac{1}{2}}(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|f_0\| \\
&\leqslant \lambda_n\left(\frac{1}{\sqrt{\lambda_n}}\|(T_n+\lambda_n I)^{-\frac{1}{2}}(T_n-T)\|_{\mathrm{op}}+1\right)\frac{1}{\sqrt{\lambda_n}}\|f_0\| \\
&\leqslant \sqrt{\lambda_n}\|f_0\|\left(c_1\log\left(\frac{2}{\delta_1}\right)\frac{B_n}{\sqrt{\lambda_n}}+1\right) \\
[\text{By (17)}] \quad &\leqslant \|f_0\|\left(c_1\left(2c_1\omega^{-1}+\omega^{-\frac{1+\theta}{2}}\sqrt{\frac{2c}{\log(2/\delta_1)}}\right)\log\left(\frac{2}{\delta_1}\right)+1\right)\sqrt{\lambda_n} \\
&= \|f_0\|\left(c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right)+1\right)\sqrt{\lambda_n} \\
&= c_7\sqrt{\lambda_n}, \qquad (20)
\end{aligned}$$

where we use the Inequality 2 in the second inequality and the (17) in the last inequality and define

$$c_7 := \|f_0\|\left[c_1\left(2c_1\omega^{-1}\log\left(\tfrac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\tfrac{2}{\delta_1}\right)}\right)+1\right].$$

For the term E_3, by applying Inequality 2 and lemma 8, we have with probability at least $1-\delta_1-\delta_3$

$$\begin{aligned}
E_3 &\leqslant \|(T+\lambda_n I)^{\frac{1}{2}}(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|(T_n+\lambda_n I)^{-\frac{1}{2}}(T+\lambda_n I)^{\frac{1}{2}}\|_{\mathrm{op}} \|(T+\lambda_n I)^{-\frac{1}{2}}g_n\| \\
[\text{Lemma(8)}] \quad &\leqslant \left(\frac{1}{\sqrt{\lambda_n}}\|(T_n+\lambda_n I)^{-\frac{1}{2}}(T_n-T)\|_{\mathrm{op}}+1\right)^2\frac{\sigma}{\sqrt{\delta_3}}\sqrt{\frac{D(\lambda_n)}{n}} \\
[\text{Lemma (1)}] \quad &\leqslant \left(c_1\log\left(\frac{2}{\delta_1}\right)\frac{B_n}{\sqrt{\lambda_n}}+1\right)^2\frac{\sigma}{\sqrt{\delta_3}}\sqrt{\frac{D(\lambda_n)}{n}} \\
[\text{By (16) } \textit{and} \text{ (17)}] \quad &\leqslant \left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right)+1\right]^2\frac{\sigma}{\sqrt{\delta_3}}\sqrt{c\omega^{-(1+\theta)}\lambda_n} = c_8\sqrt{\lambda_n},
\end{aligned}$$

where in the last inequality

$$c_8 := \left[c_1\left(2c_1\omega^{-1}\log\left(\tfrac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\tfrac{2}{\delta_1}\right)}\right)+1\right]^2\frac{\sigma(\omega^{-\frac{1+\theta}{2}}\sqrt{c})}{\sqrt{\delta_3}}.$$

Notice we have (19) with probability at least $1-\delta_2-\delta_3-\delta_4-\delta_5$. Therefore, for the term E_2 we obtain with probability at least $1-\sum_{i=1}^5\delta_i$, by using

$$\|(T+\lambda_n I)^{\frac{1}{2}}(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \leqslant c_1\left(2c_1\omega^{-1}\log\left(\tfrac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\tfrac{2}{\delta_1}\right)}\right)+1$$

in (20), $\|(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \leqslant \frac{1}{\sqrt{\lambda_n}}$, lemma 2 and (19),

$$\begin{aligned}
E_2 &\leqslant \|(T+\lambda_n I)^{\frac{1}{2}}(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|(T_n+\lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|H_n\|_{\mathrm{op}} \|\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\| \\
&\leqslant \left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right)+1\right]\frac{1}{\sqrt{\lambda_n}}\cdot\frac{\kappa M_2[\nu+8M_1\log^{1/2}(p/\delta_2)]}{\sqrt{n}}\cdot\frac{2c_4c_6+c_5}{\sqrt{n}} \\
&\leqslant \frac{\kappa M_2[\nu+8M_1\log^{1/2}(p/\delta_2)](2c_4c_6+c_5)}{\sqrt{\omega}}\left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right)+1\right]\frac{1}{\sqrt{n}} = \frac{c_9}{\sqrt{n}},
\end{aligned}$$

where in the last step we use $n\lambda_n \geqslant \omega$ and define

$$c_9 := \frac{\kappa M_2(2c_4c_6+c_5)}{\sqrt{\omega}}\left[c_1\left(2c_1\omega^{-1}\log\left(\frac{2}{\delta_1}\right)+\omega^{-\frac{1+\theta}{2}}\sqrt{2c\log\left(\frac{2}{\delta_1}\right)}\right)+1\right]\left(\nu+8M_1\log^{1/2}\left(\frac{p}{\delta_2}\right)\right).$$

Thus, we bound $\|T^{\frac{1}{2}}(\hat{f}_n-f_0)\|$ by

$$\|T^{\frac{1}{2}}(\hat{f}_n-f_0)\| \leqslant E_1+E_2+E_3 \leqslant (c_7+c_8)\sqrt{\lambda_n}+\frac{c_9}{\sqrt{n}}.$$

Recall the excess prediction risk can be bounded by

$$\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) \leqslant 2\lambda_{\max}\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|_2^2 + 2\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\|^2,$$

based on which we further have

$$\mathcal{E}(\hat{\eta}_n) - \mathcal{E}(\eta_0) \leqslant 2\lambda_{\max}\left(\frac{2c_4c_6 + c_5}{\sqrt{n}}\right)^2 + 2\left((c_7 + c_8)\sqrt{\lambda_n} + \frac{c_9}{\sqrt{n}}\right)^2 = C_1 n^{-1} + C_2\sqrt{\frac{\lambda_n}{n}} + C_3\lambda_n,$$

where $C_1 := 2\lambda_{\max}(2c_4c_6 + c_5)^2 + 2c_9^2$, $C_2 := 4(c_7 + c_8)c_9$ and $C_3 := 2(c_7 + c_8)^2$.

Finally we get the desired conclusion after we notice $\sqrt{\frac{\lambda_n}{n}} = \sqrt{\omega}\, n^{-\frac{2+\theta}{2+2\theta}}$ in the above proof. □

5.2. Proof of the theorem 2

The whole proof of theorem 2 is served for the condition where p has increasing number of dimension. Let M be the smallest integer greater than $b_0 n^{\frac{1}{1+2r}}$ and $L = M^{2r}$, where constant b_0 will be defined in later proof. For two binary sequences $\gamma = (\gamma_{L+1}, \cdots, \gamma_{2L}) \in \{0,1\}^L$ and $\theta = (\theta_{M+1}, \cdots, \theta_{2M}) \in \{0,1\}^M$, define

$$\beta_\theta = M^{-\frac{1}{2}} \sum_{k=M+1}^{2M} \theta_k L_K^{\frac{1}{2}}\varphi_k \quad \text{and} \quad \boldsymbol{\alpha}_\gamma = \frac{1}{L}(\gamma_{L+1},\cdots,\gamma_{2L}).$$

By applying $\langle L_K^{\frac{1}{2}}\varphi_j, \rangle L_K^{\frac{1}{2}}\varphi_{kK} = \langle \varphi_j, \rangle L_K\varphi_{kK} = \langle \varphi_j, \varphi_k\rangle = \delta_{jk}$, we have $\|\boldsymbol{\alpha}_\gamma\|_1 \leqslant 1$ and $\beta_\theta \in \mathcal{H}(K)$ noticing

$$\|\beta_\theta\|_K^2 = \left\| M^{-\frac{1}{2}} \sum_{k=M+1}^{2M} \theta_k L_K^{\frac{1}{2}}\varphi_k \right\|_K^2 = \sum_{k=M+1}^{2M} M^{-1}\theta_k^2 \|L_K^{\frac{1}{2}}\varphi_k\|_K^2 \leqslant M^{-1}\sum_{k=M+1}^{2M}\|L_K^{\frac{1}{2}}\varphi_k\|_K^2 = 1.$$

Using the lemma 10, there exist sets $\Gamma = \{\gamma^i\}_{i=0}^{N_\gamma} \subset \{0,1\}^L$ and $\Theta = \{\theta^i\}_{i=0}^{N_\theta} \subset \{0,1\}^M$ such that

$$\text{(ia)}\gamma^0 = (0,\cdots,0), \quad \text{(iia)}H(\gamma^i,\gamma^j) > \tfrac{L}{8} \text{ for all } i \neq j, \quad \text{(iiia)}\ N_\gamma \geqslant 2^{\frac{L}{8}},$$
$$\text{(ib)}\theta^0 = (0,\cdots,0), \quad \text{(iib)}\ H(\theta^i,\theta^j) > \tfrac{M}{8} \text{ for all } i \neq j, \quad \text{(iiib)}\ N_\theta \geqslant 2^{\frac{M}{8}},$$

where $H(\theta, \theta')$ is the Hamming distance between θ and θ'. Define the combined parameter

$$\vartheta = (\gamma_{L+1},\cdots,\gamma_{2L}, \theta_{M+1},\cdots,\theta_{2M}) \in \{0,1\}^{L+M}.$$

The construction of Γ and Θ implies that there exists a set $\Theta = \{\vartheta^i\}_{i=0}^{N_\vartheta} \subset \{0,1\}^{L+M}$ such that

$$\text{(ic)}\ \vartheta^0 = (0,\cdots,0), \quad \text{(iic)}\ H(\vartheta^i,\vartheta^j) > \frac{L+M}{8} \text{ for all } i \neq j, \quad \text{(iiic)}\ N_\vartheta \geqslant 2^{\frac{L+M}{8}}.$$

Let $P^n_{\boldsymbol{\alpha}_0,\beta_0}$ be the joint distribution on the product space of training samples $\{(Z_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^n$ generated by the true parameter $(\boldsymbol{\alpha}_0, \beta_0)$, where $Z_i = \boldsymbol{X}_i^T\boldsymbol{\alpha}_0 + \langle Y_i, \beta_0\rangle + \varepsilon_i$, and $P_{\boldsymbol{\alpha}_0,\beta_0}$ be the distribution on a single sample $(Z, \boldsymbol{X}, Y)$, where $Z = \boldsymbol{X}^T\boldsymbol{\alpha}_0 + \langle Y, \beta_0\rangle + \varepsilon$.

By the independence of the training samples, for different $\theta, \theta' \in \Theta$ and $\gamma, \gamma' \in \Gamma$, we have

$$\log\left(\frac{\mathrm{d}P^n_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}}{\mathrm{d}P^n_{\boldsymbol{\alpha}_\gamma,\beta_\theta}}(\{(Z_i,\boldsymbol{X}_i,Y_i)\}_{i=1}^n)\right) = \sum_{i=1}^n \log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}}{\mathrm{d}P_{\boldsymbol{\alpha}_\gamma,\beta_\theta}}(Z_i,\boldsymbol{X}_i,Y_i)\right).$$

Using the Assumption 9, we can bound the Kullback-Leibler distance between $P^n_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}$ and $P^n_{\boldsymbol{\alpha}_\gamma,\beta_\theta}$

$$K(P^n_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}|P^n_{\boldsymbol{\alpha}_\gamma,\beta_\theta}) = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}\left(\log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}}{\mathrm{d}P_{\boldsymbol{\alpha}_\gamma,\beta_\theta}}\right)\right) \leqslant nK_{\sigma^2}\mathbb{E}[\langle Y, \beta_{\theta'} - \beta_\theta\rangle + \boldsymbol{X}^T(\boldsymbol{\alpha}_{\gamma'} - \boldsymbol{\alpha}_\gamma)]^2$$
$$\leqslant 2nK_{\sigma^2}\mathbb{E}(\langle Y, \beta_{\theta'} - \beta_\theta\rangle)^2 + 2nK_{\sigma^2}\mathbb{E}[\boldsymbol{X}^T(\boldsymbol{\alpha}_{\gamma'} - \boldsymbol{\alpha}_\gamma)]^2.$$

Noticing $\langle L_K^{\frac{1}{2}}\varphi_j, L_C L_K^{\frac{1}{2}}\varphi_k\rangle = \langle \varphi_j, T\varphi_k\rangle = \tau_k\delta_{jk}$, we have

$$\begin{aligned}
\mathbb{E}(\langle Y, \beta_{\theta'} - \beta_\theta\rangle)^2 &= \langle \beta_{\theta'} - \beta_\theta, L_C(\beta_{\theta'} - \beta_\theta)\rangle \\
&= \left\langle M^{-\frac{1}{2}}\sum_{k=M+1}^{2M}(\theta'_k - \theta_k)L_K^{\frac{1}{2}}\varphi_k,\ M^{-\frac{1}{2}}\sum_{k=M+1}^{2M}(\theta'_k - \theta_k)L_C L_K^{\frac{1}{2}}\varphi_k \right\rangle \\
&= M^{-1}\sum_{k=M+1}^{2M}(\theta'_k - \theta_k)^2\tau_k \leqslant M^{-1}\tau_M\sum_{k=M+1}^{2M}(\theta'_k - \theta_k)^2 \\
&= M^{-1}\tau_M H(\theta', \theta) \leqslant \tau_M = t_0 M^{-2r},
\end{aligned}$$

where the last inequality is by the eigen-decay condition. For parametric part, we have

$$\mathbb{E}[\boldsymbol{X}^T(\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_{\gamma})]^2 \leqslant \lambda_{\max}(D)\|\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_{\gamma}\|_2^2 = \lambda_{\max}(D)H(\gamma',\gamma)/L^2 \leqslant \lambda_{\max}(D)/L = \lambda_{\max}(D)M^{-2r},$$

from which we obtain

$$K(P^n_{\boldsymbol{\alpha}_{\gamma'},\beta_{\theta'}}|P^n_{\boldsymbol{\alpha}_{\gamma},\beta_{\theta}}) \leqslant 2(t_0+\lambda_{\max}(D))nK_{\sigma^2}M^{-2r}.$$

If we put $b_0 := \left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{\frac{1}{1+2r}}$, then for any $\rho\in(0,\frac{1}{8})$, we have by $M \geqslant b_0 n^{\frac{1}{1+2r}}$

$$\frac{1}{N_\vartheta}\sum_{j=1}^{N_\vartheta} K(P_{\alpha_{\gamma_j},\beta_{\theta^j}}|P_{\alpha_{\gamma^0},\beta_{\theta^0}}) \leqslant 2(t_0+\lambda_{\max}(D))nK_{\sigma^2}M^{-2r} \leqslant 2\rho\log(2^{\frac{M}{8}}) \leqslant 2\rho\log(2^{\frac{L+M}{8}}) \leqslant 2\rho\log N_\vartheta.$$

For $\theta\in\Theta$ and and $\gamma\in\Gamma$, let the prediction rule $\eta_{\gamma,\theta}$ be $\eta_{\gamma,\theta}(\boldsymbol{X},Y) := \boldsymbol{X}^T\boldsymbol{\alpha}_\gamma + \langle Y,\beta_\theta\rangle$. For different $\theta,\ \theta'\in\Theta$ and $\gamma,\ \gamma'\in\Gamma$, when the true parameter is $(\boldsymbol{\alpha}_\gamma,\beta_\theta)$, the excess prediction risk for the prediction rule $\eta_{\gamma',\theta'}$ is

$$\begin{aligned}
\mathcal{E}(\eta_{\gamma',\theta'})-\mathcal{E}(\eta_{\gamma,\theta}) &= \mathbb{E}[\langle Y,\ \beta_{\theta'}-\beta_\theta\rangle + \boldsymbol{X}^T(\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_\gamma)]^2\\
&= \mathbb{E}(\langle Y,\ \beta_{\theta'}-\beta_\theta\rangle)^2 \ +\ 2\mathbb{E}(\langle Y,\ \beta_{\theta'}-\beta_\theta\rangle\cdot\boldsymbol{X}^T(\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_\gamma))\\
&\quad + \mathbb{E}[\boldsymbol{X}^T(\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_\gamma)]^2\\
[\text{By } \mathbb{E}\boldsymbol{X}=0 \text{ and } \boldsymbol{X}\perp Y] &= M^{-1}\sum_{k=M+1}^{2M}(\theta'_k-\theta_k)^2\tau_k + (\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_\gamma)^T\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T](\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_\gamma)\\
&\geqslant M^{-1}\tau_{2M}\sum_{k=M+1}^{2M}(\theta'_k-\theta_k)^2 + \lambda_{\min}(D)\|\boldsymbol{\alpha}_{\gamma'}-\boldsymbol{\alpha}_\gamma\|_2^2\\
&= M^{-1}\tau_{2M}H(\theta',\theta) + L^{-2}\lambda_{\min}(D)H(\gamma',\gamma)\\
&\geqslant t_0M^{-1}(2M)^{-2r}\frac{M}{8} + L^{-2}\lambda_{\min}(D)\frac{L}{8} = \frac{1}{8}(2^{-2r}t_0+\lambda_{\min}(D))M^{-2r}.
\end{aligned}$$

Notice M is the smallest integer greater than $b_0 n^{\frac{1}{1+2r}}$, and to control the lower bound of last expression we additionally set $M \leqslant 2b_0 n^{\frac{1}{1+2r}}$. Thus

$$b_0 n^{\frac{1}{1+2r}} \geqslant 1 \Leftrightarrow n \geqslant \frac{\rho\log 2}{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}.$$

Therefore, we obtain the lower bound for $\mathcal{E}(\eta_{\gamma',\theta'})-\mathcal{E}(\eta_{\gamma,\theta})$

$$\begin{aligned}
\mathcal{E}(\eta_{\gamma',\theta'})-\mathcal{E}(\eta_{\gamma,\theta}) &\geqslant 2^{-3}(2^{-2r}t_0+\lambda_{\min}(D))(2b_0n^{\frac{1}{1+2r}})^{-2r}\\
&= 2^{-(2r+3)}(2^{-2r}t_0+\lambda_{\min}(D))\left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{-\frac{2r}{1+2r}} n^{-\frac{2r}{1+2r}}.
\end{aligned}$$

Consider the set $\Xi := \{(\boldsymbol{\alpha}_\gamma,\beta_\theta)\colon \gamma\in\Gamma,\theta\in\Theta\}$. By the lemma 9, we have

$$\begin{aligned}
&\inf_{\tilde\eta}\sup_{\eta_0\in\Xi} P\left(\mathcal{E}(\tilde\eta)-\mathcal{E}(\eta_0) \geqslant \left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{-\frac{2r}{1+2r}}\cdot\frac{2^{-2r}t_0+\lambda_{\min}(D)}{2^{2r+3}n^{\frac{2r}{1+2r}}}\right)\\
&\geqslant \frac{\sqrt{N_\vartheta}}{1+\sqrt{N_\vartheta}}\left(1-4\rho-\sqrt{\frac{4\rho}{\log N_\vartheta}}\right).
\end{aligned}$$

From the construction of $\boldsymbol{\alpha}_\gamma\in\mathbb{R}^p$, we see $p = L = M^{2r} \asymp n^{\frac{2r}{1+2r}}$. Note that for fixed $\tilde\eta$, we have $\sup_{\eta_0\in\mathbb{R}^p\times\mathcal{H}(K)} P\{\mathcal{E}(\tilde\eta)-\mathcal{E}(\eta_0)\geqslant\cdots\} \geqslant \sup_{\eta_0\in\Xi} P\{\mathcal{E}(\tilde\eta)-\mathcal{E}(\eta_0)\geqslant\cdots\}$ and by (iiic) we get

$$\log N_\vartheta \geqslant \frac{M+L}{8}\log 2 = \left\{\left(\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right)^{\frac{1}{1+2r}} n^{\frac{1}{1+2r}} + \left[\frac{8(t_0+\lambda_{\max}(D))K_{\sigma^2}}{\rho\log 2}\right]^{\frac{2r}{1+2r}} n^{\frac{2r}{1+2r}}\right\}\frac{\log 2}{8}.$$

We obtain the desired conclusion. □

5.3. Proofs of the key lemmas

5.3.1. The derivation of (6) and (7)

To obtain the minimizer $(\hat{\boldsymbol{\alpha}}_n,\hat f_n)$, we take derivative of the following $F_n(\boldsymbol{\alpha},f)$ with respect to $\boldsymbol{\alpha}$ and use variation calculus with respect to the functional parameter f.

Recall that $Z_i = \boldsymbol{X}_i^T\boldsymbol{\alpha}_0 + \langle Y_i, L_K^{\frac{1}{2}}f_0\rangle + \varepsilon_i$, where $\{(\boldsymbol{X}_i,Y_i,\varepsilon_i)\}_{i=1}^n$ are independent copies of $(\boldsymbol{X},Y,\varepsilon)$ in (1). Thus the right side of (5) can be written as

$$F_n(\boldsymbol{\alpha}, f) \coloneqq \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i^T(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)+\langle Y_i, L_K^{\frac{1}{2}}(f-f_0)\rangle-\varepsilon_i)^2+\lambda_n\|f\|^2.$$

Notice $(\hat{\boldsymbol{\alpha}}_n, \hat{f}_n)$ is the minimum of $F_n(\boldsymbol{\alpha}, f)$, therefore $\frac{\partial F_n(\hat{\boldsymbol{\alpha}}_n, \hat{f}_n)}{\partial \boldsymbol{\alpha}} = 0$, from which we have

$$\begin{aligned}0 &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i(\boldsymbol{X}_i^T(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+\langle Y_i, L_K^{\frac{1}{2}}(\hat{f}_n-f_0)\rangle-\varepsilon_i)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+\frac{1}{n}\sum_{i=1}^{n}\langle Y_i, L_K^{\frac{1}{2}}(\hat{f}_n-f_0)\rangle\boldsymbol{X}_i-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i \coloneqq D_n(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\\ &\quad + G_n(\hat{f}_n-f_0)-a_n,\end{aligned}$$

thus $\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0 = -D_n^{-1}G_n(\hat{f}_n-f_0)+D_n^{-1}a_n$.

Next define the function $\varphi_n(t;\boldsymbol{\alpha}, f, g) \coloneqq F_n(\boldsymbol{\alpha}, f+tg)$, and the fact that $(\hat{\boldsymbol{\alpha}}_n, \hat{f}_n)$ minimizes $F_n(\boldsymbol{\alpha}, f)$ implies

$$\left.\frac{\mathrm{d}\varphi(t;\hat{\boldsymbol{\alpha}}_n, \hat{f}_n, g)}{\mathrm{d}t}\right|_{t=0}, \quad \forall g\in L^2(\mathcal{T}),$$

from which we have

$$0 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i^T(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+\langle Y_i, L_K^{\frac{1}{2}}(\hat{f}_n-f_0)\rangle-\varepsilon_i)\langle Y_i, L_K^{\frac{1}{2}}g\rangle+\lambda_n\langle\hat{f}_n, g\rangle. \tag{21}$$

From (21), we have

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)L_K^{\frac{1}{2}}Y_i+\frac{1}{n}\sum_{i=1}^{n}\langle Y_i, L_K^{\frac{1}{2}}(\hat{f}_n-f_0)\rangle L_K^{\frac{1}{2}}Y_i-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i L_K^{\frac{1}{2}}Y_i+\lambda_n\hat{f}_n = 0. \tag{22}$$

Notice $L_{C_n}f = \left\langle \frac{1}{n}\sum_{i=1}^{n}Y_i(s)Y_i(t), f(t)\right\rangle = \frac{1}{n}\sum_{i=1}^{n}\langle Y_i, f\rangle Y_i$ and recall that $a_n \coloneqq \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\boldsymbol{X}_i$, $H_n(\boldsymbol{\alpha}) \coloneqq \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i^T\boldsymbol{\alpha})L_K^{\frac{1}{2}}Y_i$ and $T_n = L_K^{\frac{1}{2}}\circ L_{C_n}\circ L_K^{\frac{1}{2}}$. Thus (22) can be reformulated as

$$(T_n+\lambda_n I)\hat{f}_n - T_n f_0 + H_n(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)-g_n = 0,$$

which yields

$$\begin{aligned}\hat{f}_n - f_0 &= (T_n+\lambda_n I)^{-1}T_n f_0 - f_0 - (T_n+\lambda_n I)^{-1}H_n(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+(T_n+\lambda_n I)^{-1}g_n\\ &= -\lambda_n(T_n+\lambda_n I)^{-1}f_0 - (T_n+\lambda_n I)^{-1}H_n(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+(T_n+\lambda_n I)^{-1}g_n.\end{aligned}$$

□

5.3.2. Proof of lemma 2

To show $\|G_n\|_{\mathrm{op}} = \|H_n\|_{\mathrm{op}}$, we first prove H_n is the conjugate operator of G_n. To bound the operator norm $\|G_n\|_{\mathrm{op}}$, we need to bound the norm $\|G_{n,j}\|_{\mathrm{op}}$ defined below. Notice the operator $G_{n,j}$ can be viewed as the average of independent sum of operators, so the concentration inequality of random variables in Hilbert space (Corollary 5) can be used.

We first prove $H_n = G_n^*$, which shows $\|G_n\|_{\mathrm{op}} = \|H_n\|_{\mathrm{op}}$. For any $\boldsymbol{\gamma}\in\mathbb{R}^p$ and $f\in L^2(\mathcal{T})$, one has

$$\boldsymbol{\gamma}^T G_n(f) = \frac{1}{n}\sum_{i=1}^{n}\langle Y_i, L_K^{\frac{1}{2}}f\rangle(\boldsymbol{\gamma}^T\boldsymbol{X}_i) = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i^T\boldsymbol{\gamma})\langle L_K^{\frac{1}{2}}Y_i, f\rangle = \langle H_n(\boldsymbol{\gamma}), f\rangle.$$

Now we turn to bound $\|G_n\|_{\mathrm{op}}$, for $1\leqslant j\leqslant p$, define the operator $G_{n,j}$: $L^2(\mathcal{T})\mapsto\mathbb{R}$ by

$$G_{n,j}(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\langle Y_i, L_K^{\frac{1}{2}}f\rangle X_{i,j},$$

which can also be viewed as a random variable taking values in a Hilbert space $L^2(\mathcal{T})^*$=$L^2(\mathcal{T})$.

Notice $G_n = (G_{n,1}, \cdots, G_{n,p})$, thus we have

$$\|G_n(f)\|^2 = \sum_{j=1}^{p}|G_{n,j}(f)|^2 \leqslant \sum_{j=1}^{p}\|G_{n,j}\|_{\mathrm{op}}^2\|f\|^2 \leqslant \left(\sum_{j=1}^{p}\|G_{n,j}\|_{\mathrm{op}}\right)^2\|f\|^2,$$

which means $\|G_n\|_{\mathrm{op}} \leqslant \sum_{j=1}^{p}\|G_{n,j}\|_{\mathrm{op}}$. Define the operator $\xi_{i,j}$: $L^2(\mathcal{T})\mapsto\mathbb{R}$ and ξ_j: $L^2(\mathcal{T})\mapsto\mathbb{R}$ by

$$\xi_{i,j}(f) \coloneqq \langle Y_i, L_K^{\frac{1}{2}}f\rangle X_{i,j} \quad \text{and} \quad \xi_j(f) \coloneqq \langle Y, L_K^{\frac{1}{2}}f\rangle X_j,$$

from which we rewrite $G_{n,j} = \frac{1}{n}\sum_{i=1}^{n}\xi_{i,j}$, where $\{\xi_{i,j}\}_{i=1}^n$ are independent copies of ξ_j.

By the definition of ξ_j, after noticing the isomorphism between $L^2(\mathcal{T})^*$ and $L^2(\mathcal{T})$, we have

$$\|\xi_j\|_{\text{op}} = \|X_j L_K^{\frac{1}{2}} Y\| \leqslant |X_j| \cdot \|L_K^{\frac{1}{2}}\|_{\text{op}} \cdot \|Y\| \leqslant \vartheta M_2 |X_j|$$

by $\|L_K^{\frac{1}{2}}\|_{\text{op}} = \kappa$ in Assumption 1. After taking expectation and using the sub-Gaussian growth of moments condition for X_j, we have

$$\|\|\xi_j\|_{\text{op}}\|_G = \sup_{k\geqslant 1}\left[\frac{\mathbb{E}\|\xi_j\|_{\text{op}}^{2k}}{(2k-1)!!}\right]^{1/(2k)} \leqslant \kappa M_2 \sup_{k\geqslant 1}\left[\frac{\mathbb{E}|X_j|^{2k}}{(2k-1)!!}\right]^{1/(2k)} = \kappa M_2 \|\|X_j|\|_G \leqslant \kappa M_2 M_1.$$

Because $\boldsymbol{X}$ and Y are independent with zero mean, we have

$$(\mathbb{E}\xi_j)(f) = \langle \mathbb{E}Y_i, L_K^{\frac{1}{2}} f\rangle \cdot \mathbb{E}X_j = 0, \quad \text{which means } \mathbb{E}\xi_j = 0.$$

By using corollary 5, we have for any $\delta_2 \in (0,1)$, with probability at least $1 - \frac{\delta_2}{p}$,

$$\|G_{n,j}\|_{\text{op}} = \left\| \frac{1}{n}\sum_{i=1}^n \xi_{i,j} \right\|_{\text{op}} \leqslant \frac{1}{\sqrt{n}}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|\xi_{i,j}\|_{\text{op}}^2} + 8\sqrt{\frac{1}{n}\sum_{i=1}^n \|\|\xi_{i,j}\|_{\text{op}}\|_G^2 \log(\frac{p}{\delta_2})}\right\}$$

$$[\text{By Assumptions3}] \quad \leqslant \frac{\kappa M_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}} =: \frac{c_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}}.$$

Notice $\|G_n\|_{\text{op}} \leqslant \sum_{j=1}^p \|G_{n,j}\|_{\text{op}}$, thus we have the following relation for the events

$$\left\{\|G_n\|_{\text{op}} \geqslant \frac{p\kappa M_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}}\right\} \subset \bigcup_{1\leqslant j\leqslant p}\left\{\|G_{n,j}\|_{\text{op}} \geqslant \frac{\kappa M_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}}\right\},$$

from which we have

$$P\left(\|G_n\|_{\text{op}} \geqslant \frac{\kappa M_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}}\right) \leqslant \sum_{j=1}^p P\left(\|G_{n,j}\|_{\text{op}} \geqslant \frac{\kappa M_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}}\right)$$

$$\leqslant \sum_{j=1}^p \frac{\delta_2}{p} = \delta_2.$$

Therefore, $P\left(\|G_n\|_{\text{op}} \leqslant \frac{p\kappa M_2[v + 8M_1 \log^{1/2}(p/\delta_2)]}{\sqrt{n}}\right) \geqslant 1 - \delta_2$.

5.3.3. Lemma 3

The lemma 3 shows the fact $\|a_n\| = O_p(n^{-\frac{1}{2}})$, and its proof is based on the Markov's inequality.

Lemma 3. *Under the Assumptions 3 and 5, for any $\delta_4 \in (0,1)$, with probability at least $1 - \delta_4$, we have*

$$\|a_n\| \leqslant \frac{c_3}{\sqrt{\delta_4}\sqrt{n}} \quad \text{with} \quad c_3 := \sqrt{p}\,\sigma v.$$

Proof. Since a_n is the average of independent sum of random vectors, the weak law of large numbers based on the Markov's inequality can be used to bound $\|a_n\|$.

Define $\eta := \varepsilon \boldsymbol{X}$ and $\eta_i := \varepsilon_i \boldsymbol{X}_i (1 \leqslant i \leqslant n)$, by which we rewrite $a_n = \frac{1}{n}\sum_{i=1}^n \xi_i$. Notice for $i \neq j$, we have $\mathbb{E}(\eta_i^T \eta_j) = \mathbb{E}(\varepsilon_i \varepsilon_j \boldsymbol{X}_i^T \boldsymbol{X}_j) = \mathbb{E}(\varepsilon_j \boldsymbol{X}_i^T \boldsymbol{X}_j \mathbb{E}(\varepsilon_i|\boldsymbol{X}_i)) = 0$. And

$$\mathbb{E}(\|\eta\|^2) = \mathbb{E}(\varepsilon^2\|\boldsymbol{X}\|^2) = \mathbb{E}(\|\boldsymbol{X}\|^2 \mathrm{E}(\varepsilon^2|\boldsymbol{X})) \leqslant \sigma^2 \sum_{j=1}^p \mathbb{E}|X_j|^2 \leqslant p\sigma^2 v^2,$$

where we use the second moment condition of $\boldsymbol{X}$ in Assumption 3. Therefore, we have

$$\mathbb{E}(\|a_n\|^2) = \mathbb{E}\left(\left\| \frac{1}{n}\sum_{i=1}^n \eta_i \right\|^2\right) = \frac{\mathbb{E}(\|\eta_1\|^2)}{n} \leqslant \frac{p\sigma^2 v^2}{n},$$

from which we obtain the Markov's inequality for a_n: $P(\|a_n\| \geqslant t) \leqslant \frac{p\sigma^2 v^2}{nt^2}$. Therefore, we can conclude for $\delta_4 \in (0,1)$, we have with probability at least $1 - \delta_4$

$$\|a_n\| \leqslant \frac{c_3}{\sqrt{\delta_4}\sqrt{n}} \quad \text{with} \quad c_3 := \sqrt{p}\,\sigma v.$$

□

5.3.4. Lemma 4

The lemma 4 is the concentration inequality for the empirical covariance matrix D_n.

Lemma 4. *Under the Assumption 3, for $t > 0$, we have*

$$P(\|D_n - D\|_\infty \geqslant t) \leqslant 2p^2 \exp\left(-\frac{nt^2}{8(64M_1^4 + M_1^2 t)}\right).$$

Proof. The L-infinity norm of $D_n - D$ is bounded by each element of the difference. Notice each entry of $D_n - D$ is the average of i.i.d. sum of random variables, which means concentration inequality can be used.

For $1 \leqslant j, k \leqslant p$, put

$$d_{jk} \coloneqq (D)_{jk} \quad \text{and} \quad d_{jk}^n \coloneqq (D_n)_{jk} = \frac{1}{n}\sum_{i=1}^n X_{i,j}X_{i,k} = \frac{1}{n}\sum_{i=1}^n d_{jk,i}^n \quad \text{with} \quad d_{jk,i}^n \coloneqq X_{i,j}X_{i,k}.$$

Let $e_{jk,i}^n$ be the centralization of $d_{jk,i}^n$, i.e. $e_{jk,i}^n \coloneqq d_{jk,i}^n - \mathbb{E}d_{jk,i}^n = d_{jk,i}^n - d_{jk}$. Thus we have

$$\{\|D_n - D\|_\infty \geqslant t\} = \bigcup_{1\leqslant j,k\leqslant p} \{|d_{jk}^n - d_{jk}| \geqslant t\} = \bigcup_{1\leqslant j,k\leqslant p} \left\{\left|\frac{1}{n}\sum_{i=1}^n e_{jk,i}^n\right| \geqslant t\right\}.$$

By the Cauchy-Schwarz inequality and $\mathbb{E}Y^{2k} \leqslant \frac{(2k)!}{2^k k!}\|Y\|_G^{2k}$ for $k \geqslant 1$, we have

$$\begin{aligned}\mathbb{E}(|d_{jk,i}^n|^l) = \mathbb{E}(|X_{i,j}|^l|X_{i,k}|^l) &\leqslant (\mathbb{E}|X_{i,j}|^{2l})^{\frac{1}{2}}(\mathbb{E}|X_{i,k}|^{2l})^{\frac{1}{2}} = (\mathbb{E}|X_j|^{2l})^{\frac{1}{2}}(\mathbb{E}|X_k|^{2l})^{\frac{1}{2}} \leqslant \frac{(2l)!}{2^l l!}(\|X_j\|_G^{2l}\|X_k\|_G^{2l})^{1/2} \\ \text{[By } \max_{1\leqslant j\leqslant p}\|X_j\|_G \leqslant M_1 \text{ in Assumption3]} \quad &\leqslant \frac{(2l)!}{2^l(l!)^2}M_1^{2l}l! \leqslant \frac{(2^l l!)^2}{2^l(l!)^2}M_1^{2l}l! \\ &= 2^l M_1^{2l} l! = \frac{8M_1^4}{2}(2M_1^2)^{l-2}\cdot l!,\end{aligned}$$

where we use the growth of sub-Gaussian moments condition of $\{X_j\}_{j=1}^p$ in the second last inequality, and the last inequality stems from:

$$\begin{aligned}(2k)! = (2k)(2k-1)(2k-2)\cdots 4\cdot 3\cdot 2\cdot 1 &= 2^k\cdot k!(2k-1)(2k-3)\cdots 3\cdot 1 \\ &\leqslant 2^k\cdot k!(2k)(2k-2)\cdots 4\cdot 2 = (2^k k!)^2 \text{ for } k \geqslant 1.\end{aligned}$$

Using $|a-b|^l \leqslant 2^l(|a|^l + |b|^l)$, we have

$$|e_{jk,i}^n|^l = |d_{jk,i}^n - \mathbb{E}d_{jk,i}^n|^l \leqslant 2^l(|d_{jk,i}^n|^l + |\mathbb{E}d_{jk,i}^n|^l).$$

Take expectation and use the Jensen's inequality $|\mathbb{E}d_{jk,i}^n|^l \leqslant \mathbb{E}|d_{jk,i}^n|^l$, we have

$$\mathbb{E}(|e_{jk,i}^n|^l) \leqslant 2^{l+1}\mathbb{E}|d_{jk,i}^n|^l \leqslant 2^{l+1}\frac{8M_1^4}{2}(2M_1^2)^{l-2}\cdot l! = \frac{64M_1^4}{2}(4M_1^2)^{l-2}\cdot l!. \tag{23}$$

For the independent random variables $\{e_{jk,i}^n\}_{i=1}^n$, by the Bernstein's inequality with the growth of moments condition (23) (see corollary 4.6 in Zhang and Chen 2021), we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n e_{jk,i}^n\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{n^2t^2}{128M_1^4 n + 8M_1^2 nt}\right) = 2\exp\left(-\frac{nt^2}{8(16M_1^4 + M_1^2 t)}\right).$$

Thus we conclude

$$P(\|D_n - D\|_\infty \geqslant t) \leqslant \sum_{1\leqslant j,k\leqslant p} P\left(\left|\frac{1}{n}\sum_{i=1}^n e_{jk,i}^n\right| \geqslant t\right) \leqslant 2p^2\exp\left(-\frac{nt^2}{8(16M_1^4 + M_1^2 t)}\right).$$

□

5.3.5. Lemma 5

The lemma 5 shows we can use $\frac{3}{2}\|D^{-1}\|_{\text{op}}$ to bound $\|D_n^{-1}\|_{\text{op}}$ from above.

Lemma 5. *Under the Assumption 3, for any $\delta_5 \in (0, 1)$, let*

$$N_1 \coloneqq 24p\|D^{-1}\|_{\text{op}}(48p\|D^{-1}\|_{\text{op}}M_1^4 + M_1^2)\log\left(\frac{2p^2}{\delta_5}\right), \tag{24}$$

we have when $n > N_1$

$$P\left(\|D_n^{-1}\|_{\text{op}} \leqslant \frac{3}{2}\|D^{-1}\|_{\text{op}}\right) \geqslant 1 - \delta_5.$$

Proof. The crucial point of the proof is the usage of the inequality on the norm of inverse of matrices, by which we notice if D_n is close to D, then D_n^{-1} is also close to D^{-1}.

Notice the fact that if $A, B \in \mathbb{R}^{p\times p}$ are invertible and $\|A^{-1}\|_{\text{op}}\|A - B\|_{\text{op}} < 1$, then

$$\|A^{-1} - B^{-1}\|_{\text{op}} \leqslant \frac{\|A^{-1}\|_{\text{op}}^2\|A - B\|_{\text{op}}}{1 - \|A^{-1}\|_{\text{op}}\|A - B\|_{\text{op}}}.$$

(see lemma E.4 in Sun *et al* 2017). Let $A = D$ and $B = D_n$, when $\|D^{-1}\|_{\text{op}}\|D - D_n\|_{\text{op}} \leqslant \frac{1}{3}$, we have

$$\|D^{-1} - D_n^{-1}\|_{\text{op}} \leqslant \frac{\frac{1}{3}\|D^{-1}\|_{\text{op}}}{1 - \frac{1}{3}} = \frac{1}{2}\|D^{-1}\|_{\text{op}},$$

from which we have $\|D_n^{-1}\|_{\text{op}} \leqslant \|D^{-1}\|_{\text{op}} + \|D^{-1} - D_n^{-1}\|_{\text{op}} \leqslant \frac{3}{2}\|D^{-1}\|_{\text{op}}$. Therefore, it gives

$$P\left(\|D_n^{-1}\|_{\text{op}} \leqslant \frac{3}{2}\|D^{-1}\|_{\text{op}}\right) \geqslant P\left(\|D^{-1}\|_{\text{op}}\|D - D_n\|_{\text{op}} \leqslant \frac{1}{3}\right).$$

Recall that we have $\|A\|_{\text{op}} \leqslant p\|A\|_\infty$ for $A \in \mathbb{R}^{p\times p}$, by the lemma 4 with $t = 1/(3p\|D^{-1}\|_{\text{op}})$, we have

$$\begin{aligned} P\left(\|D_n^{-1}\|_{\text{op}} \geqslant \frac{3}{2}\|D^{-1}\|_{\text{op}}\right) &\leqslant P\left(\|D^{-1}\|_{\text{op}}\|D - D_n\|_{\text{op}} \geqslant \frac{1}{3}\right) \leqslant P\left(p\|D^{-1}\|_{\text{op}}\|D - D_n\|_\infty \geqslant \frac{1}{3}\right) \\ &\leqslant 2p^2\exp\left(-\frac{n}{24p\|D^{-1}\|_{\text{op}}(48p\|D^{-1}\|_{\text{op}}M_1^4 + M_1^2)}\right). \end{aligned}$$

Let N_1 be defined in (24), as $n > N_1$, we get $P\left(\|D_n^{-1}\|_{\text{op}} \geqslant \frac{3}{2}\|D^{-1}\|_{\text{op}}\right) \leqslant \delta_5$. □

6. The auxiliary lemmas and results

The lemma 1, lemma 8, Inequalities (1) and (2) are from the proof of Tong and Ng (2018). We provide all the complete proofs in this section for integrity.

6.1. Proof of lemma 1

Define the random variable

$$\xi(f) := (T + \lambda_n I)^{-\frac{1}{2}}\langle L_K^{\frac{1}{2}}Y, f\rangle L_K^{\frac{1}{2}}Y \quad \text{and} \quad \xi_i(f) := (T + \lambda_n I)^{-\frac{1}{2}}\langle L_K^{\frac{1}{2}}Y_i, f\rangle L_K^{\frac{1}{2}}Y_i \quad (1 \leqslant i \leqslant n).$$

Then ξ_i are independent copies of ξ, which takes values in $\text{HS}(\mathcal{T})$.

Recall we define $\{(\tau_k, \varphi_k)\}_{k\geqslant 1}$ to be the set of eigenvalue-eigenfunction pairs of the operator T,

$$\begin{aligned} \|\xi\|_{\text{HS}}^2 &= \sum_{k=1}^{+\infty}\|(T + \lambda_n I)^{-\frac{1}{2}}\langle L_K^{\frac{1}{2}}Y, \varphi_k\rangle L_K^{\frac{1}{2}}Y\|^2 \leqslant \|(T + \lambda_n I)^{-\frac{1}{2}}\|_{\text{op}}^2\|L_K^{\frac{1}{2}}Y\|^2\sum_{k=1}^{+\infty}|\langle L_K^{\frac{1}{2}}Y, \varphi_k\rangle|^2 \\ &= \|(T + \lambda_n I)^{-\frac{1}{2}}\|_{\text{op}}^2\|L_K^{\frac{1}{2}}Y\|^4 \leqslant \frac{\kappa^4 M_2^4}{\lambda_n}, \end{aligned}$$

where we use the fact $\sum_{k=1}^{+\infty}|\langle L_K^{\frac{1}{2}}Y, \varphi_k\rangle|^2 = \|L_K^{\frac{1}{2}}Y\|^2, \|(T + \lambda_n I)^{-\frac{1}{2}}\|_{\text{op}} \leqslant \frac{1}{\sqrt{\lambda_n}}$ and $\|L_K^{\frac{1}{2}}Y\| \leqslant \|L_K^{\frac{1}{2}}\|_{\text{op}}\|Y\| \leqslant \kappa M_2$ from (3).

Notice $L_K^{\frac{1}{2}}Y$ can be expended to $\sum_{l=1}^{+\infty}\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle\varphi_l$, we have

$$\begin{aligned}\|\xi\|_{\mathrm{HS}}^2 &= \sum_{k=1}^{+\infty}\left\|\sum_{l=1}^{+\infty}\langle L_K^{\frac{1}{2}}Y,\ \varphi_k\rangle\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle(T+\lambda_n I)^{-\frac{1}{2}}\varphi_l\right\|^2\\ &=\left(\sum_{k=1}^{+\infty}|\langle L_K^{\frac{1}{2}}Y,\ \varphi_k\rangle|^2\right)\left\|\sum_{l=1}^{+\infty}\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle(T+\lambda_n I)^{-\frac{1}{2}}\varphi_l\right\|^2\\ &=\|L_K^{\frac{1}{2}}Y\|^2\left\|\sum_{l=1}^{+\infty}\frac{1}{\sqrt{\tau_l+\lambda_n}}\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle\varphi_l\right\|^2 = \|L_K^{\frac{1}{2}}Y\|^2\sum_{l=1}^{+\infty}\frac{1}{\tau_l+\lambda_n}|\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle|^2\\ &\leqslant\kappa^2M_2^2\sum_{l=1}^{+\infty}\frac{1}{\tau_l+\lambda_n}|\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle|^2.\end{aligned}$$

Using (8) we have

$$\mathbb{E}[|\langle L_K^{\frac{1}{2}}Y,\ \varphi_l\rangle|^2] = \mathbb{E}[|\langle Y,\ L_K^{\frac{1}{2}}\varphi_l\rangle|^2] = \langle L_K^{\frac{1}{2}}\varphi_l,\ L_C L_K^{\frac{1}{2}}\varphi_l\rangle = \langle T\varphi_l,\ \varphi_l\rangle = \tau_l,$$

from which we get

$$\mathbb{E}(\|\xi\|_{\mathrm{HS}}^2) \leqslant \kappa^2M_2^2\sum_{l=1}^{+\infty}\frac{\tau_l}{\tau_l+\lambda_n} = \kappa^2M_2^2D(\lambda_n).$$

Notice

$$\begin{aligned}\mathbb{E}(\langle L_K^{\frac{1}{2}}Y,f\rangle\langle L_K^{\frac{1}{2}}Y,g\rangle) &= \mathbb{E}(\langle Y,\ L_K^{\frac{1}{2}}f\rangle\langle Y,\ L_K^{\frac{1}{2}}g\rangle) = \mathbb{E}\iint_{\mathcal{T}\times\mathcal{T}}Y(s)Y(t)(L_K^{\frac{1}{2}}f)(s)(L_K^{\frac{1}{2}}g)(t)\mathrm{d}s\mathrm{d}t\\ &= \int_{\mathcal{T}}\left(\int_{\mathcal{T}}C(s,\ t)(L_K^{\frac{1}{2}}f)(s)\mathrm{d}s\right)(L_K^{\frac{1}{2}}g)(t)\mathrm{d}t = \langle L_C L_K^{\frac{1}{2}}f,\ L_K^{\frac{1}{2}}g\rangle = \langle Tf,\ g\rangle,\end{aligned}$$

from which we have $\mathbb{E}(\langle L_K^{\frac{1}{2}}Y,f\rangle L_K^{\frac{1}{2}}Y) = T(f)$. Therefore $(\mathbb{E}\xi)(f) = (T+\lambda_n I)^{-\frac{1}{2}}T(f)$.

Taking $\|\xi\|_{\mathrm{H}} \leqslant \frac{\kappa^2M_2^2}{\sqrt{\lambda_n}}$ and $\mathbb{E}(\|\xi\|_{\mathrm{H}}^2) \leqslant \kappa^2M_2^2D(\lambda_n)$ in the lemma 7, we have for any $\delta_1\in(0,2e^{-1})$, with probability at least $1-\delta_1$

$$\begin{aligned}\|(T+\lambda_n)^{-\frac{1}{2}}(T_n-T)\|_{\mathrm{op}} &\leqslant \left\|\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\mathbb{E}\xi_i)\right\|_{\mathrm{HS}}\\ &\leqslant \frac{2\kappa^2M_2^2\log\left(\frac{2}{\delta_1}\right)}{n\sqrt{\lambda_n}} + \sqrt{\frac{2\kappa^2M_2^2D(\lambda_n)\log\left(\frac{2}{\delta_1}\right)}{n}} \leqslant c_1\log\left(\frac{2}{\delta_1}\right)B_n,\end{aligned}$$

where we let $c_1 := \kappa M_2$ and $B_n := \frac{2c_1}{n\sqrt{\lambda_n}} + \sqrt{\frac{2D(\lambda_n)}{n\log(2/\delta_1)}}$. □

6.2. Lemma 7, lemma 6 and corollary 5

The lemma 7 and lemma 6 can be seen as the Bernstein-type concentration inequalities for the random variables taking values in a Hilbert space. In the lemma 7, we assume the random variables to be bounded with regard to the norm in the Hilbert space, while we assume that the random variables satisfy the Bernstein's growth of moments condition in the lemma 6. The first lemma is based on theorem 3.3.4a in Yurinsky (2006).

Lemma 6. *Let $\mathcal{H}$ be a Hilbert space endowed with norm $\|\cdot\|_{\mathrm{H}}$. Let $\{\xi_i\}_{i=1}^n$ be a sequence of* n *independent random variables in $\mathcal{H}$ with zero mean. Assume there exist $B,\ M>0$ such that for all integers $l\geqslant 2$:*

$$\mathbb{E}(\|\xi_i\|_{\mathrm{H}}^l) \leqslant \frac{B^2}{2}l!M^{l-2},\quad i = 1,2,\cdots,n,$$

then for any $\delta\in(0,\ 1)$, we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right\|_{\mathrm{H}} \geqslant \frac{2M\log\left(\frac{2}{\delta}\right)}{n} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}}B\right) \leqslant \delta.$$

Proof. We refer readers to lemma 2 in Lv and Feng (2012) for the proof of this lemma. □

The growth of moments condition with $l=2$ gives $B = \sqrt{\max_{1\leqslant i\leqslant n}\mathbb{E}(\|\xi_i\|_{\mathrm{H}}^2)}$. Then we have the following lemma.

Lemma 7. *Let $\mathcal{H}$ be a Hilbert space endowed with norm $\|\cdot\|_{\mathrm{H}}$ and $\{\xi_i\}_{i=1}^n$ be a sequence of* n *independent zero-mean random variables taking values in $\mathcal{H}$. Assume that $\|\xi\|_{\mathrm{H}} \leqslant M$ (a.s.), then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$.*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i \right\|_{\mathrm{H}} \leqslant \frac{2M \log\left(\frac{2}{\delta}\right)}{n} + \sqrt{\frac{2 \max\limits_{1\leqslant i\leqslant n} \mathbb{E}(\|\xi_i\|_{\mathrm{H}}^2) \log\left(\frac{2}{\delta}\right)}{n}}$$

Proof. For $l > 2$, one has $\mathbb{E}(\|\xi_i\|_{\mathrm{H}}^l) \leqslant M^{l-2}\mathbb{E}(\|\xi_i\|_{\mathrm{H}}^2) < \frac{B^2}{2} l! M^{l-2}$, $i = 1, 2, \cdots, n$ with $B = \sqrt{\max_{1\leqslant i\leqslant n} \mathbb{E}(\|\xi_i\|_{\mathrm{H}}^2)}$. The result follows from lemma 6. □

Next, we consider the sub-Gaussian concentration for norm of sum of random vectors in the Hilbert space. In order to derive concentration for the norm of sum of sub-Gaussian vectors in Hilbert space, we consider following framework in Maurer and Pontil (2021).

Let $Z = (Z_1, \ldots, Z_n)$ be a vector of independent data with values in a space H, define Z' as an independent copy of Z. Given a function $f\colon \mathrm{H}^n \to \mathbb{R}$, it is of interest to study the concentration inequality for $f(Z) - \mathbb{E}f(Z)$. A special case is $f(Z)$ the norm function of data. For $w \in \mathrm{H}$ and $k \in \{1, \ldots, n\}$ define the substitution operator $S_w^k\colon \mathrm{H}^n \to \mathrm{H}^n$ by

$$S_w^k(z) = (z_1, \ldots, z_{k-1}, w, z_{k+1}, \ldots, z_n)$$

and the centered conditional version of f as the random variable is given by

$$\begin{aligned} D_{f,Z_k}(z) &\equiv f(z_1, \ldots, z_{k-1}, Z_k, z_{k+1}, \ldots, z_n) - \mathbb{E}[f(z_1, \ldots, z_{k-1}, Z_k', z_{k+1}, \ldots, z_n)] \\ &= f(S_{Z_k}^k(z)) - \mathbb{E}[f(S_{Z_k'}^k(z))] = \mathbb{E}[f(S_{Z_k}^k(z)) - f(S_{Z_k'}^k(z))|Z_k]. \end{aligned} \tag{25}$$

Here $\{D_{f,Z_k}(z)\}_{k=1}^n$ can be viewed as random-valued functions $z \in \mathrm{H}^n \mapsto D_{f,Z_k}(z)$. If $f(z) = \sum_{i=1}^n z_i$ then $D_{f,Z_k}(Z) = Z_k - \mathbb{E}Z_k$ is independent of z.

Based on the norm $\|\cdot\|_G$, we aim to obtain a tighter and extended McDiarmid's inequality for $f(X) - \mathbb{E}f(X)$ with stochastic bounded difference conditions concerning the structure of f. The following corollary is a tighter sub-Gaussian concentration for the norm of sum of sub-Gaussian vectors in Hilbert space, comparing to theorem 3 in Maurer and Pontil (2021).

Corollary 4. *Suppose that $\{D_{f,Z_i}(z)\}_{i=1}^n$ have zero mean defined by (25). If $\{D_{f,Z_i}(z)\}_{i=1}^n$ have finite $\|\cdot\|_G$-norm for all $z \in \mathrm{H}$, we have, $\forall\, t \geqslant 0$*

$$f(Z) - \mathbb{E}f(Z) \sim \mathrm{subG}\left(8 \sup_{z\in\mathrm{H}} \sum_{i=1}^{n} \|D_{f,Z_i}(z)\|_G^2\right) \text{ and } P\{f(Z) - \mathbb{E}f(Z) > t\}$$

$$\leqslant \exp\left(\frac{-t^2}{16 \sup\limits_{z\in\mathrm{H}} \sum_{k=1}^n \| D_{f,Z_k}(z) \|_G^2}\right).$$

Proof. Let the *tilted expectation* $\mathbb{E}_Y$ be $\mathbb{E}_Y[Z] = \mathbb{E}\left[Z \cdot \frac{e^Y}{\mathbb{E}e^Y}\right]$ with the exponential weighted $\frac{e^Y}{\mathbb{E}e^Y}$. The proof is based on the entropy of a random variable Y defined by

$$S(Y) \coloneqq \mathbb{E}_Y[Y] - \log \mathbb{E}[e^Y] = \mathbb{E}\left[Y \cdot \frac{e^Y}{\mathbb{E}e^Y}\right] - \log \mathbb{E}[e^Y],$$

which is free of centering, i.e. $S(Y - \mathbb{E}Y) = S(Y)$.

Suppose that $\mathbb{E}Y = 0$, by Jensen's inequality we have

$$S(Y) = \mathbb{E}_Y\left[\log\left(\frac{e^Y}{\mathbb{E}e^Y}\right)\right] \leqslant \log \mathbb{E}_Y\left[\frac{e^Y}{\mathbb{E}e^Y}\right] = \log \mathbb{E}\left[\frac{e^Y}{\mathbb{E}e^Y} \cdot \frac{e^Y}{\mathbb{E}e^Y}\right] = \log \mathbb{E}e^{2Y} - 2\log \mathbb{E}e^Y \leqslant \log \mathbb{E}e^{2Y}, \tag{26}$$

where the last inequality is also derived by Jensen's inequality $\mathbb{E}e^Y \geqslant e^{\mathbb{E}Y} = 1$.

The concentration inequality is by Cramer-Chernoff method with the logarithm of the MGF represented as the integral of entropy (theorems 1 in Maurer 2012).

$$\log \mathbb{E}[e^{t(Y-\mathbb{E}Y)}] = \log \mathbb{E}[e^{tY}] = t\int_0^t \frac{S(\gamma Y)\, d\gamma}{\gamma^2}, \quad \forall \quad t > 0 \tag{27}$$

and the *subadditivity of entropy* (see theorems 6 and section 3.1 in Maurer 2012)

$$S(f(Z)) \leqslant \mathbb{E}_{f(Z)}\left[\sum_{i=1}^{n} S(D_{f,Z_i}(Z))\right], \quad \text{where we denote } S(D_{f,Z_i}(Z)) \coloneqq S(D_{f,Z_k}(z))|_{z=Z}. \tag{28}$$

If Y has zero mean, then $\mathbb{E}e^{tY} = 1 + \sum_{k=2}^{\infty} \frac{t^k \mathbb{E}Y^k}{k!}$. Note that by Cauchy's inequality and arithmetic-geometric mean inequality

$$\mathbb{E}|tY|^{2k+1} \leqslant (\mathbb{E}|tY|^{2k}\mathbb{E}|tY|^{2k+2})^{1/2} \leqslant \frac{1}{2}(t^{2k}\mathbb{E}Y^{2k} + t^{2k+2}\mathbb{E}Y^{2k+2}).$$

Then, $\frac{\mathbb{E}\,|\,tY\,|^3}{3!} \leqslant \frac{1}{2\cdot 3!}(t^2\mathbb{E}Y^2 + t^4\mathbb{E}Y^4)$ implies

$$\begin{aligned}\mathbb{E}e^{tY} &\leqslant 1 + \left(\frac{1}{2} + \frac{1}{2\cdot 3!}\right)t^2\mathbb{E}Y^2 + \sum_{k=2}^{\infty}\left(\frac{1}{(2k)!} + \frac{1}{2}\left[\frac{1}{(2k-1)!} + \frac{1}{(2k+1)!}\right]\right)t^{2k}\mathbb{E}Y^{2k} \\ &\leqslant \sum_{k=0}^{\infty} 2^k \frac{t^{2k}\mathbb{E}Y^{2k}}{(2k)!} \leqslant \exp\{t^2 \,\|Y\|_G^2\},\end{aligned} \tag{29}$$

where the last inequality is by the definition of $\|Y\|_G < \infty$ and $\mathbb{E}Y^{2k} \leqslant \frac{(2k)!}{2^k k!}\,\|Y\|_G^{2k}$.

Thus (29) shows $\log \mathbb{E}e^{2tY} \leqslant 4t^2\,\|Y\|_G^2$ and (26) gives

$$S(D_{f,Z_i}(z)) \leqslant \log \mathbb{E}e^{2D_{f,Z_i}(z)} \leqslant 4\,\|D_{f,Z_i}(z)\|_G^2, \quad z \in \mathrm{H}. \tag{30}$$

Denote $\|D_{f,Z_i}(Z)\|_G^2 \coloneqq \|D_{f,Z_i}(z)\|_G^2|_{z=Z}$. Combing (27) and (28), we obtain

$$\begin{aligned}\log \mathbb{E}[e^{t(f(Z)-\mathbb{E}f(Z))}] &= t\int_0^t \frac{S(\gamma f(Z))\,d\gamma}{\gamma^2} \leqslant t\int_0^t \frac{1}{\gamma^2}\mathbb{E}_{\gamma f(Z)}\left[\sum_{i=1}^{n} S(D_{\gamma f,Z_i}(Z))\right]d\gamma \\ &= t\int_0^t \frac{1}{\gamma^2}\mathbb{E}_{\gamma f(Z)}\left[\sum_{i=1}^{n} S(\gamma D_{f,Z_i}(z))|_{z=Z}\right]d\gamma \\ [\text{By (30)}] &\leqslant t\int_0^t \frac{4\gamma^2}{\gamma^2}\mathbb{E}_{\gamma f(Z)}\left[\sum_{i=1}^{n} \|D_{f,Z_i}(z)\|_G^2|_{z=Z}\right]d\gamma \\ &\leqslant t\int_0^t \frac{4\gamma^2}{\gamma^2}\mathbb{E}_{\gamma f(Z)}\left[\sup_{z\in \mathrm{H}}\sum_{i=1}^{n} \|D_{f,Z_i}(z)\|_G^2\right]d\gamma = 8\sup_{z\in Z}\sum_{i=1}^{n} \|D_{f,Z_i}(z)\|_G^2 \cdot \frac{t^2}{2},\end{aligned}$$

which shows $f(Z) - \mathbb{E}f(Z) \sim \mathrm{subG}\,(8\sup_{z\in \mathrm{H}}\sum_{i=1}^{n}\|D_{f,Z_i}(z)\|_G^2)$.

□

Corollary 5 Hoeffding-type inequality for norm of vector. *Suppose* $\{\xi_i\}_{i=1}^n$ *are independent random elements with values in a Hilbert space* H *s.t.* $\max_{i\in[n]}\|\|\xi_i\|_{\mathrm{H}}\|_G < \infty$. *Then, with probability at least* $1-\delta$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)\right\|_{\mathrm{H}} \leqslant \frac{1}{\sqrt{n}}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\xi_i - \mathbb{E}\xi_i\|^2} + 8\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\|\xi_i\|_{\mathrm{H}}\|_G^2 \log\left(\frac{1}{\delta}\right)}\right\}. \tag{31}$$

Proof. Consider the function $f(\boldsymbol{z}) \coloneqq \|\sum_{i=1}^{n}\boldsymbol{z}_i - \boldsymbol{v}\|$ for any $\boldsymbol{v}, \boldsymbol{z}_i \in \mathrm{H}$. From (25), we have

$$\mathbb{E}[f(S_{\xi_k}^k(\boldsymbol{z})) - f(S_{\xi_k'}^k(\boldsymbol{z}))|\xi_k] = \mathbb{E}\left[\left|\left\|\sum_{i\neq k}\boldsymbol{z}_i + \xi_k - \boldsymbol{v}\right\| - \left\|\sum_{i\neq k}\boldsymbol{z}_i + \xi_k' - \boldsymbol{v}\right\|\right| \mid \xi_k\right] \leqslant \mathbb{E}[\|\xi_k - \xi_k'\|_{\mathrm{H}} \mid \xi_k], \tag{32}$$

for $k = 1, 2, \cdots, n$, which directly implies $\|D_{f,\xi_k}(\boldsymbol{z})\|_G \leqslant \|\|\xi_k - \xi_k'\|_{\mathrm{H}}\|_G$ by following derivation:

$$\begin{aligned}\|D_{f,\xi_k}(\boldsymbol{z})\|_G &= \|f(z_1,\ldots,z_{k-1},\,\xi_k,\,z_{k+1},\ldots,x_n) - \mathbb{E}[f(z_1,\ldots,z_{k-1},\,\xi_k',\,z_{k+1},\ldots,z_n)]\|_G \\ &= \|\mathbb{E}[f(S_{\xi_k}^k(\boldsymbol{z})) - f(S_{\xi_k'}^k(\boldsymbol{z}))|\xi_k]\|_G \\ [\text{By (32)}] &\leqslant \|\mathbb{E}[\|\xi_k - \xi_k'\|_{\mathrm{H}}|\xi_k]\|_G.\end{aligned} \tag{33}$$

The conditional Jensen's inequality and Hölders inequality show for any $k \geqslant 1$

$$\mathbb{E}[|\mathbb{E}[\|\xi_k - \xi_k'\|_{\mathrm{H}}|\xi_k]|^{2k}] \leqslant \mathbb{E}[\mathbb{E}[\|\xi_k - \xi_k'\|_{\mathrm{H}}|\xi_k]^{2k}] \leqslant \mathbb{E}\{\mathbb{E}[\|\xi_k - \xi_k'\|_{\mathrm{H}}^{2k}|\xi_k]\} = \mathbb{E}\,|\|\xi_k - \xi_k'\|_{\mathrm{H}}|^{2k}.$$

Then the definition $\|Y\|_G = \sup_{k\geqslant 1}\left[\frac{2^k k!}{(2k)!}\mathbb{E}Y^{2k}\right]^{1/(2k)}$ and (33) show

$$\|D_{f,\xi_k}(\boldsymbol{z})\|_G \leqslant \|\|\xi_k - \xi_k'\|_{\mathrm{H}}\|_G \leqslant 2\|\|\xi_k\|_{\mathrm{H}}\|_G. \tag{34}$$

Now we take $\boldsymbol{\nu} = \sum_{i=1}^{n} \mathbb{E}\xi_i$ and apply corollary 4 with the first inequality in (34). Let $e^t = \delta$ for solving t, we have with probability at least $1 - \delta$, for $Z \coloneqq (\xi_1, \ldots, \xi_n)$

$$f(Z) - \mathbb{E}f(Z) = \left\| \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \right\|_{\mathrm{H}} - \mathbb{E} \left\| \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \right\|_{\mathrm{H}} \leqslant 4 \sqrt{\sum_{k=1}^{n} \| \|\xi_k - \xi_k'\|_{\mathrm{H}} \|_G^2 \log(1/\delta)}$$

$$\text{[By the second inequality in (34)]} \quad \leqslant 8 \sqrt{\sum_{k=1}^{n} \| \|\xi_k\|_{\mathrm{H}} \|_G^2 \log(1/\delta)}.$$

Since H is a Hilbert space, $\{\xi_i\}_{i=1}^n$ are independent, and Jensen's inequality implies

$$\mathbb{E} \left\| \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \right\|_{\mathrm{H}} \leqslant \sqrt{\mathbb{E} \left\| \sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \right\|_{\mathrm{H}}^2} = \sqrt{\sum_{i=1}^{n} \mathbb{E}\|\xi_i - \mathbb{E}\xi_i\|^2}.$$

Then, with probability at least $1 - \delta$, we have (31). □

6.3. Lemma 8

The lemma 8 shows the concentration property of $(T + \lambda_n I)^{-\frac{1}{2}} g_n$.

Lemma 8. *Under the Assumption 5, for any $\delta_3 \in (0, 1)$, with probability at least $1 - \delta_3$, there exists*

$$\|(T + \lambda_n I)^{-\frac{1}{2}} g_n\| \leqslant \frac{\sigma}{\sqrt{\delta_3}} \sqrt{\frac{D(\lambda_n)}{n}}.$$

Proof. Define random variables ξ and $\{\xi_i\}_{i=1}^n$) taking values in the Hilbert space $L^2(\mathcal{T})$ by

$$\xi \coloneqq \varepsilon (T + \lambda_n I)^{-\frac{1}{2}} L_K^{\frac{1}{2}} Y \quad \text{and} \quad \xi_i \coloneqq \varepsilon_i (T + \lambda_n I)^{-\frac{1}{2}} L_K^{\frac{1}{2}} Y_i,$$

where $\{\xi_i\}_{i=1}^n$ are independent copies of ξ and $(T + \lambda_n I)^{-\frac{1}{2}} g_n = \frac{1}{n} \sum_{i=1}^{n} \xi_i$.

Noticing the random noise ε has conditional zero mean given Y, we have

$$\mathbb{E}\xi = \mathbb{E}((T + \lambda_n I)^{-\frac{1}{2}} L_K^{\frac{1}{2}} Y \cdot \mathbb{E}(\varepsilon | Y)) = 0.$$

Expanding ξ by the basis $\{\varphi_k : k \geqslant 1\}$ of the operator T, and we have

$$\mathbb{E}(\|\xi\|^2) = \mathbb{E}\left(\left\| \sum_{k=1}^{+\infty} \langle \varepsilon (T + \lambda_n I)^{-\frac{1}{2}} L_K^{\frac{1}{2}} Y, \varphi_k \rangle \varphi_k \right\|^2 \right) = \mathbb{E}\left(\varepsilon^2 \left\| \sum_{k=1}^{+\infty} \langle L_K^{\frac{1}{2}} Y, (T + \lambda_n I)^{-\frac{1}{2}} \varphi_k \rangle \varphi_k \right\|^2 \right)$$

$$= \mathbb{E}\left(\varepsilon^2 \left\| \sum_{k=1}^{+\infty} \frac{1}{\sqrt{\tau_k + \lambda_n}} \langle L_K^{\frac{1}{2}} Y, \varphi_k \rangle \varphi_k \right\|^2 \right) = \mathbb{E}\left(\varepsilon^2 \sum_{k=1}^{+\infty} \frac{|\langle L_K^{\frac{1}{2}} Y, \varphi_k \rangle|^2}{\tau_k + \lambda_n} \right)$$

$$= \mathbb{E}\left(\sum_{k=1}^{+\infty} \frac{|\langle L_K^{\frac{1}{2}} Y, \varphi_k \rangle|^2}{\tau_k + \lambda_n} \cdot \mathbb{E}(\varepsilon^2 | Y) \right) \leqslant \sigma^2 \sum_{k=1}^{+\infty} \frac{\mathbb{E}(|\langle L_K^{\frac{1}{2}} Y, \varphi_k \rangle|^2)}{\tau_k + \lambda_n} = \sigma^2 \sum_{k=1}^{+\infty} \frac{\langle T\varphi_k, \varphi_k \rangle}{\tau_k + \lambda_n} = \sigma^2 D(\lambda_n).$$

Hence we have

$$\mathbb{E}(\|(T + \lambda_n I)^{-\frac{1}{2}} g_n\|^2) = \mathbb{E}(\|\frac{1}{n} \sum_{i=1}^{n} \xi_i\|^2) = \frac{\mathbb{E}(\|\xi\|^2)}{n} \leqslant \frac{\sigma^2 D(\lambda_n)}{n}.$$

Using Markov inequality, we get $P(\|(T + \lambda_n I)^{-\frac{1}{2}} g_n\| \geqslant t) \leqslant \frac{\sigma^2 D(\lambda_n)}{n t^2}$ from which we conclude with probability at least $1 - \delta_3$, we have $\|(T + \lambda_n I)^{-\frac{1}{2}} g_n\| \leqslant \frac{\sigma}{\sqrt{\delta_3}} \sqrt{\frac{D(\lambda_n)}{n}}$. □

6.4. Two crucial inequalities

The following two inequalities play an important role in the proof of the theorem 1, which shows $\|(T + \lambda_n I)(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}}$ and $\|(T + \lambda_n I)^{\frac{1}{2}} (T_n + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}}$ can be bounded by $\|(T + \lambda_n I)^{-\frac{1}{2}} (T_n - T)\|_{\mathrm{op}}$.

Inequality 1. $\|(T + \lambda_n I)(T_n + \lambda_n T)^{-1}\|_{\mathrm{op}} \leqslant \left(\frac{1}{\sqrt{\lambda_n}} \|(T + \lambda_n I)^{-\frac{1}{2}} (T_n - T)\|_{\mathrm{op}} + 1 \right)^2$.

Proof. Using the following decomposition of the operator product

$$BA^{-1} = (B - A)B^{-1}(B - A)A^{-1} + (B - A)B^{-1} + I$$

with $A = T_n + \lambda_n I$ and $B = T + \lambda_n I$, we have

$$(T + \lambda_n I)(T_n + \lambda_n T)^{-1} = (T - T_n)(T + \lambda_n I)^{-1}(T - T_n)(T_n + \lambda_n I)^{-1} + (T - T_n)(T + \lambda_n I)^{-1} + I =: F_1 + F_2 + I.$$

For the operator F_1, we have

$$\|F_1\|_{\mathrm{op}} \leqslant \|(T - T_n)(T + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|(T + \lambda_n I)^{-\frac{1}{2}}(T - T_n)\|_{\mathrm{op}} \cdot \frac{1}{\lambda_n} = \frac{1}{\lambda_n} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}}^2,$$

where we use the fact $\|AB\|_{\mathrm{op}} = \|(AB)^*\|_{\mathrm{op}} = \|B^*A^*\|_{\mathrm{op}} = \|BA\|_{\mathrm{op}}$ for any self-adjoint operators A and B, and the bound $\|(T_n + \lambda_n I)^{-1}\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_n}$.

For the operator F_2, applying $\|(T + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \leqslant \frac{1}{\sqrt{\lambda_n}}$, we have

$$\|F_2\|_{\mathrm{op}} \leqslant \|(T - T_n)(T + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \|(T + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \leqslant \frac{1}{\sqrt{\lambda_n}} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}}.$$

Thus we obtain

$$\|(T + \lambda_n I)(T_n + \lambda_n T)^{-1}\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_n} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}}^2 + \frac{1}{\sqrt{\lambda_n}} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}} + 1 \leqslant \left(\frac{1}{\sqrt{\lambda_n}} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}} + 1\right)^2.$$

□

Inequality

2. $\|(T_n + \lambda_n I)^{-\frac{1}{2}}(T + \lambda_n I)^{\frac{1}{2}}\|_{\mathrm{op}} = \|(T + \lambda_n I)^{\frac{1}{2}}(T_n + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \leqslant \frac{1}{\sqrt{\lambda_n}} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}} + 1.$

Proof. Applying the fact that

$$\|A^\gamma B^\gamma\|_{\mathrm{op}} \leqslant \|AB\|_{\mathrm{op}}^\gamma, \quad \gamma \in (0, 1)$$

for positive operators A and B defined on Hilbert space (see lemma A.7 in Blanchard and Krämer 2010), we have

$$\|(T_n + \lambda_n I)^{-\frac{1}{2}}(T + \lambda_n I)^{\frac{1}{2}}\|_{\mathrm{op}} = \|(T + \lambda_n I)^{\frac{1}{2}}(T_n + \lambda_n I)^{-\frac{1}{2}}\|_{\mathrm{op}} \leqslant \|(T + \lambda_n I)(T_n + \lambda_n T)^{-1}\|_{\mathrm{op}}^{\frac{1}{2}} \leqslant \frac{1}{\sqrt{\lambda_n}} \|(T + \lambda_n I)^{-\frac{1}{2}}(T_n - T)\|_{\mathrm{op}} + 1,$$

where in the last step we use the inequality (1). □

6.5. Lemma 9

The lemma 9 is helpful in constructing the lower bound which is based on the testing multiple hypothesis.

Lemma 9. *Assume $N \geqslant 2$ and suppose there exists $\Theta = \{\theta_i\}_{i=0}^N$ such that the conditions are satisfied:*

1. *$2r$-separated condition: $d(\theta_j, \theta_k) \geqslant 2r > 0, \quad \forall\, 0 \leqslant j < k \leqslant N$,*
2. *Kullback-Leibler average condition: if $P_j \ll P_0$ for $1 \leqslant j \leqslant N$ and*

$$\frac{1}{N}\sum_{j=1}^N K(P_j|P_0) \leqslant \rho \log N \text{ for some } 0 < \rho < \frac{1}{8} \text{ and } P_j = P_{\theta_j} (0 \leqslant j \leqslant N).$$

Then for all possible random variables $\tilde{\theta}$, we have

$$\inf_{\tilde{\theta}} \sup_{\theta \in \Theta} P_\theta(d(\tilde{\theta}, \theta) \geqslant r) \geqslant \frac{\sqrt{N}}{1 + \sqrt{N}} \left(1 - 2\rho - \sqrt{\frac{2\rho}{\log N}}\right) > 0.$$

Proof. We refer readers to theorem 2.5 in Tsybakov (2008) for the proof of this lemma. □

6.6. Varhsamov-Gilbert lemma

We need the Varhsamov-Gilbert lemma in the proof of the theorem 2 to construct the analogy of Θ in lemma 9.

Lemma 10. *Let $H(\theta, \theta') = \sum_{k=1}^M 1(\theta_k \neq \theta'_k)$ be the Hamming distance between elements θ, θ' in $\{0, 1\}^M$. For any integer $M \geqslant 8$, there exist vectors $\{\theta^i\}_{i=0}^N \subset \{0, 1\}^M$ such that*

$$\text{(i) } \theta^0 = (0,\cdots,0), \quad \text{(ii)} \quad H(\theta^i, \theta^j) > \frac{M}{8} \quad \text{for all} \quad i \neq j, \quad \text{(iii)} \quad N \geqslant 2^{\frac{M}{8}}.$$

Proof. We refer readers to page 104 in Tsybakov (2008) for the proof of this lemma. □

6.7. Derivation of the equality (15)

Let $f_1(\boldsymbol{X}, Y)$ be the density of $(\boldsymbol{X}, Y)$ and $f_2(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}}$ be the density of ε since we assume $\varepsilon \sim N(0, \sigma^2)$, then the density of $P_{\boldsymbol{\alpha}_1,\beta_1}$ can be written as

$$\frac{\mathrm{d}P_{\boldsymbol{\alpha}_1,\beta_1}}{\mathrm{d}\mu}(Z, \boldsymbol{X}, Y) = f_1(\boldsymbol{X}, Y) f_2(Z - \boldsymbol{X}^T\boldsymbol{\alpha}_1 - \langle Y, \beta_1\rangle),$$

where μ is a dominant measure on the space $\mathbb{R} \times \mathbb{R}^p \times L^2(\mathcal{T})$.

Therefore, under the assumption of $f_2(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}}$, we have

$$\begin{aligned}
\log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_1,\beta_1}}{\mathrm{d}P_{\boldsymbol{\alpha}_2,\beta_2}}\right) &= \log\left(\frac{f_2(Z - \boldsymbol{X}^T\boldsymbol{\alpha}_1 - \langle Y, \beta_1\rangle)}{f_2(Z - \boldsymbol{X}^T\boldsymbol{\alpha}_2 - \langle Y, \beta_2\rangle)}\right) \\
&= \frac{1}{2\sigma^2}((Z - \boldsymbol{X}^T\boldsymbol{\alpha}_2 - \langle Y, \beta_2\rangle)^2 - (Z - \boldsymbol{X}^T\boldsymbol{\alpha}_1 - \langle Y, \beta_1\rangle)^2) \\
&= \frac{1}{\sigma^2}(Z - \boldsymbol{X}^T\boldsymbol{\alpha}_1 - \langle Y, \beta_1\rangle)(\boldsymbol{X}^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) + \langle Y, \beta_1 - \beta_2\rangle) \\
&\quad + \frac{1}{2\sigma^2}(\boldsymbol{X}^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) + \langle Y, \beta_1 - \beta_2\rangle)^2.
\end{aligned}$$

Notice when the true parameter is $(\boldsymbol{\alpha}_1, \beta_1)$, we have $Z = \boldsymbol{X}^T\boldsymbol{\alpha}_1 + \langle Y, \beta_1\rangle + \varepsilon$, based on which we obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\alpha}_1,\beta_1}(Z - \boldsymbol{X}^T\boldsymbol{\alpha}_1 - \langle Y, \beta_1\rangle)(\boldsymbol{X}^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) + \langle Y, \beta_1 - \beta_2\rangle) \\
=&\mathbb{E}\varepsilon(\boldsymbol{X}^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) + \langle Y, \beta_1 - \beta_2\rangle) = \mathbb{E}[(\boldsymbol{X}^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) + \langle Y, \beta_1 - \beta_2\rangle) \cdot \mathbb{E}(\varepsilon|\boldsymbol{X}, Y)] = 0.
\end{aligned}$$

Thus the Kullback-Leibler distance

$$K(P_{\boldsymbol{\alpha}_1,\beta_1}|P_{\boldsymbol{\alpha}_2,\beta_2}) = \mathbb{E}_{\boldsymbol{\alpha}_1,\beta_1}\left(\log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_1,\beta_1}}{\mathrm{d}P_{\boldsymbol{\alpha}_2,\beta_2}}\right)\right) = \frac{1}{2\sigma^2}\mathbb{E}(\boldsymbol{X}^T(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2) + \langle Y, \beta_1 - \beta_2\rangle)^2.$$

□

6.8. Minimax rate with given Euclidean predictors

Assumption 10. For a fixed $\boldsymbol{\alpha}_0{}^* \in \mathbb{R}^p$ and different $\beta_1, \beta_2 \in \mathcal{H}(K)$. We assume that the Kullback-Leibler distance between $P_{\boldsymbol{\alpha}_0*,\beta_1}$ and $P_{\boldsymbol{\alpha}_0*,\beta_2}$ can be bounded by

$$K(P_{\boldsymbol{\alpha}_0^*,\beta_1}|P_{\boldsymbol{\alpha}_0^*,\beta_2}) \coloneqq \mathbb{E}_{\boldsymbol{\alpha}_0^*,\beta_1}\log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_0^*,\beta_1}}{\mathrm{d}P_{\boldsymbol{\alpha}_0^*,\beta_2}}\right) \leqslant K_{\sigma^2}\mathbb{E}(\langle Y, \beta_1 - \beta_2\rangle)^2,$$

where $K_{\sigma^2} > 0$ is a variance-dependent constant and $\mathbb{E}_{\boldsymbol{\alpha}_0*,\beta_1}$ is the expectation taken over $P_{\boldsymbol{\alpha}_0*,\beta_1}$.

Corollary 6 Minimax rate with given Euclidean predictors. *Under the Assumptions 5 and 10, suppose the eigenvalues* $\{\tau_k\colon k \geqslant 1\}$ *of the operator* T *decay as* $\tau_k = t_0 k^{-2r}$ *for some* r, $t_0 \in (0, \infty)$, *then for* $\rho \in (0, \frac{1}{8})$, *there exists a sequence* $\{N_n\}_{n\geqslant 1}$ *satisfying*

$$\log(N_n) \geqslant \left(\frac{8}{\log 2}\right)^{-\frac{2r}{1+2r}}\left(\frac{t_0 K_{\sigma^2}}{\rho}\right)^{\frac{1}{1+2r}} n^{\frac{1}{1+2r}}, \tag{35}$$

such that when $n \geqslant \frac{\rho\log 2}{8t_0 K_{\sigma^2}}$, *the excess prediction risk satisfies*

$$\inf_{\tilde{\eta}} \sup_{\eta_0 \in \mathbb{R}^p \times \mathcal{H}(K)} P\left(\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant \frac{t_0}{2^{4(1+r)}}\left(\frac{8t_0 K_{\sigma^2}}{\rho\log 2}\right)^{-\frac{2r}{1+2r}} n^{-\frac{2r}{1+2r}}\right) \geqslant \frac{\sqrt{N_n}}{1+\sqrt{N_n}}\left(1 - 2\rho - \sqrt{\frac{2\rho}{\log N_n}}\right),$$

where we identify the prediction rule $\tilde{\eta}$ *as the arbitrary estimator* $(\tilde{\boldsymbol{\alpha}}, \tilde{\beta})$ *based on the training samples* $\{(Z_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^n$, *and view* η_0 *as the true parameter* $(\boldsymbol{\alpha}_0, \beta_0) \in \mathbb{R}^p \times \mathcal{H}(K)$. *We emphasize the probability* P *is taken over the product space of training samples* $\{(Z_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^n$ *generated by* $\eta_0 = (\boldsymbol{\alpha}_0, \beta_0)$.

In Cai and Yuan (2012), a asymptotically minimax lower bound is derived for the FLM in the theorem 1, which is also a corollary of our result on the non-asymptotic and constant-specified minimax lower bound when $N_n \to \infty$.

Proof. Let M be the smallest integer greater than $b_0 n^{\frac{1}{1+2r}}$, where b_0 will be defined in later proof. For a binary sequence $\theta = (\theta_{M+1},\cdots,\theta_{2M}) \in \{0,1\}^M$, define

$$\beta_\theta = M^{-\frac{1}{2}} \sum_{k=M+1}^{2M} \theta_k L_K^{\frac{1}{2}} \varphi_k.$$

By applying $\langle L_K^{\frac{1}{2}}\varphi_j, \rangle L_K^{\frac{1}{2}}\varphi_{kK} = \langle \varphi_j, \rangle L_K \varphi_{kK} = \langle \varphi_j, \varphi_k\rangle = \delta_{jk}$, we can show $\beta_\theta \in \mathcal{H}(K)$, because

$$\|\beta_\theta\|_K^2 = \|M^{-\frac{1}{2}} \sum_{k=M+1}^{2M} \theta_k L_K^{\frac{1}{2}}\varphi_k\|_K^2 = \sum_{k=M+1}^{2M} M^{-1}\theta_k^2 \|L_K^{\frac{1}{2}}\varphi_k\|_K^2 \leqslant M^{-1} \sum_{k=M+1}^{2M} \|L_K^{\frac{1}{2}}\varphi_k\|_K^2 = 1.$$

Using the lemma 10, there exist a set $\Theta = \{\theta^i\}_{i=0}^N \subset \{0,1\}^M$ such that

$$\text{(i)} \quad \theta^0 = (0,\cdots,0), \quad \text{(ii)} \quad H(\theta^i, \theta^j) > \frac{M}{8} \quad \text{for all} \quad i \neq j, \quad \text{(iii)} \quad N \geqslant 2^{\frac{M}{8}}.$$

For $\eta_0 = (\boldsymbol{\alpha}_0, \beta_0)$, let $P^n_{\boldsymbol{\alpha}_0,\beta_0}$ be the joint distribution on the product space of training samples $\{(Z_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^n$ generated by the true parameter $(\boldsymbol{\alpha}_0, \beta_0)$, where $Z_i = \boldsymbol{X}_i^T\boldsymbol{\alpha}_0 + \langle Y_i, \beta_0\rangle + \varepsilon_i$, and $P_{\boldsymbol{\alpha}_0,\beta_0}$ be the distribution on a single sample $(Z, \boldsymbol{X}, Y)$, where $Z = \boldsymbol{X}^T\boldsymbol{\alpha}_0 + \langle Y, \beta_0\rangle + \varepsilon$. By the independence of the training samples, for fixed $\boldsymbol{\alpha}_0{}^* \in \mathbb{R}^p$ and different $\theta, \theta' \in \Theta$, we have

$$\log\left(\frac{\mathrm{d}P^n_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}}}{\mathrm{d}P^n_{\boldsymbol{\alpha}_0{}^*,\beta_\theta}}(\{(Z_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^n)\right) = \sum_{i=1}^n \log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}}}{\mathrm{d}P_{\boldsymbol{\alpha}_0{}^*,\beta_\theta}}(Z_i, \boldsymbol{X}_i, Y_i)\right).$$

Using the Assumption 9, we can bound the Kullback-Leibler distance between $P^n_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}}$ and $P^n_{\boldsymbol{\alpha}_0{}^*,\beta_\theta}$

$$K(P^n_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}} | P^n_{\boldsymbol{\alpha}_0{}^*,\beta_\theta}) = \sum_{i=1}^n \mathbb{E}_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}} \log\left(\frac{\mathrm{d}P_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}}}{\mathrm{d}P_{\boldsymbol{\alpha}_0{}^*,\beta_\theta}}\right) \leqslant nK_{\sigma^2}\mathbb{E}(\langle Y, \beta_{\theta'} - \beta_\theta\rangle)^2.$$

Noticing $\langle L_K^{\frac{1}{2}}\varphi_j, L_C L_K^{\frac{1}{2}}\varphi_k\rangle = \langle \varphi_j, T\varphi_k\rangle = \tau_k\delta_{jk}$, we have

$$\begin{aligned}
\mathbb{E}(\langle Y, \beta_{\theta'} - \beta_\theta\rangle)^2 &= \langle \beta_{\theta'} - \beta_\theta, L_C(\beta_{\theta'} - \beta_\theta)\rangle \\
&= \left\langle M^{-\frac{1}{2}} \sum_{k=M+1}^{2M} (\theta'_k - \theta_k) L_K^{\frac{1}{2}}\varphi_k, M^{-\frac{1}{2}} \sum_{k=M+1}^{2M} (\theta'_k - \theta_k) L_C L_K^{\frac{1}{2}}\varphi_k \right\rangle \\
&= M^{-1} \sum_{k=M+1}^{2M} (\theta'_k - \theta_k)^2\tau_k \leqslant M^{-1}\tau_M \sum_{k=M+1}^{2M} (\theta'_k - \theta_k)^2 \\
&= M^{-1}\tau_M H(\theta', \theta) \leqslant \tau_M = t_0 M^{-2r},
\end{aligned}$$

from which we have $K(P^n_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta'}} | P^n_{\boldsymbol{\alpha}_0{}^*,\beta_\theta}) \leqslant t_0 nK_{\sigma^2}M^{-2r}$.

If we let $b_0 := \left(\frac{8t_0K_{\sigma^2}}{\log 2}\right)^{\frac{1}{1+2r}} \rho^{-\frac{1}{1+2r}}$, then for any $\rho \in (0, \frac{1}{8})$, we have

$$\frac{1}{N}\sum_{j=1}^N K(P_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta^j}} | P_{\boldsymbol{\alpha}_0{}^*,\beta_{\theta^0}}) \leqslant t_0 nK_{\sigma^2}M^{-2r} \leqslant \rho\log(2^{\frac{M}{8}}) \leqslant \rho\log(N).$$

For $\theta \in \Theta$ and a fixed $\boldsymbol{\alpha}_0{}^* \in \mathbb{R}^p$, we consider the prediction rule $\eta_\theta(\boldsymbol{X}, Y) := \boldsymbol{X}^T\boldsymbol{\alpha}_0{}^* + \langle Y, \beta_\theta\rangle$. For different $\theta, \theta' \in \Theta$, when the true parameter is $(\boldsymbol{\alpha}_0{}^*, \beta_\theta)$, the excess prediction risk for $\eta_{\theta'}$ is

$$\begin{aligned}
\mathcal{E}(\eta_{\theta'}) - \mathcal{E}(\eta_\theta) &= \mathbb{E}[\boldsymbol{X}^{*T}(\boldsymbol{\alpha}_0{}^* - \boldsymbol{\alpha}_0{}^*) + \langle Y^*, \beta_{\theta'} - \beta_\theta\rangle]^2 = \mathbb{E}(\langle Y^*, \beta_{\theta'} - \beta_\theta\rangle)^2 \\
&= M^{-1} \sum_{k=M+1}^{2M} (\theta'_k - \theta_k)^2\tau_k \geqslant M^{-1}\tau_{2M} \sum_{k=M+1}^{2M} (\theta'_k - \theta_k)^2 = M^{-1}\tau_{2M} H(\theta', \theta) \\
&\geqslant M^{-1} t_0 (2M)^{-2r} \frac{M}{8} = t_0 2^{-(2r+3)} M^{-2r}.
\end{aligned}$$

Since M is the smallest integer greater than $b_0 n^{\frac{1}{1+2r}}$, so when $b_0 n^{\frac{1}{1+2r}} \geqslant 1 \Leftrightarrow n \geqslant \frac{\rho\log 2}{8t_0K_{\sigma^2}}$, $M \leqslant 2b_0 n^{\frac{1}{1+2r}}$. Thus we obtain the lower bound for $\mathcal{E}(\eta_{\theta'}) - \mathcal{E}(\eta_\theta)$

$$\mathcal{E}(\eta_{\theta'}) - \mathcal{E}(\eta_\theta) \geqslant t_0 2^{-(2r+3)}(2b_0 n^{\frac{1}{1+2r}})^{-2r} = t_0 2^{-(3+4r)} \left(\frac{8t_0K_{\sigma^2}}{\log 2}\right)^{-\frac{2r}{1+2r}} \rho^{\frac{2r}{1+2r}} n^{-\frac{2r}{1+2r}}.$$

For fixed $\boldsymbol{\alpha}_0{}^* \in \mathbb{R}^p$, consider the set $\Xi := \{(\boldsymbol{\alpha}_0{}^*, \beta_\theta) : \theta \in \Theta\}$. By the lemma 9, we have

$$\inf_{\tilde{\eta}} \sup_{\eta_0 \in \Xi} P\left(\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant t_0 2^{-4(1+r)} \left(\frac{8t_0K_{\sigma^2}}{\log 2}\right)^{-\frac{2r}{1+2r}} \rho^{\frac{2r}{1+2r}} n^{-\frac{2r}{1+2r}}\right) \geqslant \frac{\sqrt{N}}{1+\sqrt{N}}\left(1 - 2\rho - \sqrt{\frac{2\rho}{\log N}}\right).$$

Notice $\sup_{\eta_0\in\Xi} P\{\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant \cdots\} \leqslant \sup_{\eta_0\in\mathbb{R}^p\times\mathcal{H}(K)} P\{\mathcal{E}(\tilde{\eta}) - \mathcal{E}(\eta_0) \geqslant \cdots\}$ and $\log(N) \geqslant \frac{\log 2}{8}M$, we have the desired conclusion. □

7. Conclusions and future studies

Recently, the PFLM has raised a sizable amount of challenging problems in functional data analysis. Numerous studies focus on the asymptotic convergence rate. However, we analyze the kernel ridge estimator for the RKHS-based PFLM and obtain the non-asymptotic upper bound for the corresponding excess prediction risk. Our work to drive the optimal upper bound weakens the common assumptions in the existing literature on (partially) functional linear regressions. The optimal bound reveals that the prediction consistency holds under the setting where the number of non-functional parameters p slightly increases with the sample size n. For fixed p, the convergence rate of the excess prediction risk attains the optimal minimax convergence rate under the eigenvalue decay assumption of the covariance operator.

More works could be done to study the non-asymptotic upper bound for the double penalized partially functional regressions. The penalization for the non-functional parameters could be Lasso, Elastic-net, or their generalizations. The proposed non-asymptotic upper bound is novel and substantially beneficial. It is also of interest to do non-asymptotic testing based on large deviation bounds for $\|\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\|^2$ and $\|T^{\frac{1}{2}}(\hat{f}_n - f_0)\|^2$. As a further study, PFLM analysis could also be considered for the analysis of variability of multiple trajectories (Contreras-Reyes *et al* (2018)).

Acknowledgments

H Zhang is supported in part by NSFC Grant No. 12101630, and the 'double first-class' construction projects of Chinese universities (ZG216S2348). The authors would like to thank the anonymous referees for their valuable comments.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Xiaoyu Lei https://orcid.org/0000-0002-9756-8959

References

Abramovich F and Grinshtein V 2016 Model selection and minimax estimation in generalized linear models *IEEE Trans. Inf. Theory* **62** 3721–30

Aneiros G, Ferraty F and Vieu P 2015 Variable selection in partial linear regression with functional covariate *Statistics* **49** 1322–47

Baíllo A and Grané A 2009 Local linear regression for functional predictor and scalar response *J. Multivariate Anal.* **100** 102–11

Blanchard G and Krämer N 2010 Optimal learning rates for kernel conjugate gradient regression *In Advances in Neural Information Processing Systems* **20** 226–34

Brunel E, Mas A and Roche A 2016 Non-asymptotic adaptive prediction in functional linear models *J. Multivariate Anal.* **143** 208–32

Buldygin V V and Kozachenko I V 2000 *Metric Characterization of Random Variables and Random Processes* (Providence: American Mathematical Soc.) Vol. 188

Cai T T, Hall P *et al* 2006 Prediction in functional linear regression *The Annals of Statistics* **34** 2159–79

Cai T T and Yuan M 2012 Minimax and adaptive prediction for functional linear regression *J. Am. Stat. Assoc.* **107** 1201–16

Caponnetto A and De Vito E 2007 Optimal rates for the regularized least-squares algorithm *Foundations of Computational Mathematics* **7** 331–68

Cardot H, Ferraty F and Sarda P 2003 Spline estimators for the functional linear model *Statistica Sinica* **13** 571–91

Contreras-Reyes J E, Quintero F O L and Wiff R 2018 Bayesian modeling of individual growth variability using back-calculation: application to pink cusk-eel (genypterus blacodes) off chile *Ecol. Modell.* **385** 145–53

Cucker F and Smale S 2001 On the mathematical foundations of learning *Bull. Am. Math. Soc.* **39** 1–49

Cui X, Lin H and Lian H 2020 Partially functional linear regression in reproducing kernel hilbert spaces *Comput. Stat. & Data Analysis* 106978

Du P and Wang X 2014 Penalized likelihood functional regression *Statistica Sinica* **24** 1017–41

Grenander U 1950 Stochastic processes and statistical inference *Arkiv för Matematik* **1** 195–277

Hall P, Horowitz J L *et al* 2007 Methodology and convergence rates for functional linear regression *The Annals of Statistics* **35** 70–91

Horn R A and Johnson C R 2012 *Matrix Analysis* (Cambridge: Cambridge University Press)

Hsing T and Eubank R 2015 *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators* (United Kingdom: Wiley) Vol. 997

Jolliffe I T 1982 A note on the use of principal components in regression *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **31** 300–3
Kokoszka P and Reimherr M 2017 *Introduction to Functional Data Analysis* (Boca Raton, FL: CRC Press)
Kong D, Xue K, Yao F and Zhang H H 2016 Partially functional linear regression in high dimensions *Biometrika* **103** 147–59
Liu Z and Li M 2020 Non-asymptotic analysis in kernel ridge regression arXiv:2006.01350
Lv S-G and Feng Y-L 2012 Integral operator approach to learning theory with unbounded sampling *Complex Analysis and Operator Theory* **6** 533–48
Maurer A 2012 Thermodynamics and concentration *Bernoulli* **18** 434–54
Maurer A and Pontil M 2021 Concentration inequalities under sub-gaussian and sub-exponential conditions *Advances in Neural Information Processing Systems* **34** 7588–97
Okamoto M 1973 Distinctness of the eigenvalues of a quadratic form in a multivariate sample *The Annals of Statistics* 763–5
Ostrovskii D M and Bach F 2021 Finite-sample analysis of m-estimators using self-concordance *Electron. J. Stat.* **15** 326–91
Preda C 2007 Regression models for functional data by reproducing kernel hilbert spaces methods *J. Stat. Plan. Inference* **137** 829–40
Ramsay J 1982 When the data are functions *Psychometrika* **47** 379–96
Ramsay J O and Silverman B W 2007 *Applied Functional Data Analysis: Methods and Case Studies* (Berlin: Springer)
Reimherr M L, Sriperumbudur B K and Taoufik B 2018 Optimal prediction for additive function-on-function regression *Electron. J. Stat.* **12** 4571–601
Schechter M 2001 *Principles of functional analysis* (Providence: American Mathematical Soc.) Number 36
Shin H 2009 Partial functional linear regression *J. Stat. Plan. Inference* **139** 3405–18
Sun H 2005 Mercer theorem for rkhs on noncompact sets *J. Complexity* **21** 337–49
Sun Q, Tan K M, Liu H and Zhang T 2017 Graphical nonconvex optimization for optimal estimation in gaussian graphical models *ICML 2018* **80** 4810–4817
Tong H and Ng M 2018 Analysis of regularized least squares for functional linear regression model *J. Complexity* **49** 85–94
Tsybakov A B 2008 *Introduction to Nonparametric Estimation* (New York: Springer Science & Business Media)
Wahba G 1990 *Spline Models for Observational Data* (Philadelphia, PA: SIAM)
Wahl M 2018 A note on the prediction error of principal component regression arXiv:1811.02998
Wang J-L, Chiou J-M and Müller H-G 2016 Functional data analysis *Annual Review of Statistics and Its Application* **3** 257–95
Yang Y, Shang Z and Cheng G 2020 Non-asymptotic analysis for nonparametric testing *Conference on Learning Theory* 3709–55
Yao F, Müller H-G and Wang J-L 2005 Functional linear regression analysis for longitudinal data *The Annals of Statistics* **33** 2873–903
Yurinsky V 2006 *Sums and Gaussian Vectors* (Berlin: Springer)
Zhang F and Lian H 2019 Partially functional linear regression with quadratic regularization *Inverse Prob.* **35** 105002
Zhang H and Chen S X 2021 Concentration inequalities for statistical inference *Communications in Mathematical Research* **37** 1–85
Zhang T 2005 Learning bounds for kernel regression using effective data dimensionality *Neural Comput.* **17** 2077–98
Zhou H, Yao F and Zhang H 2023 Functional linear regression for discretely observed data: from ideal to reality *Biometrika* **110** 381–93
Zhu H, Zhang R, Yu Z, Lian H and Liu Y 2019 Estimation and testing for partially functional linear errors-in-variables models *J. Multivariate Anal.* **170** 296–314
Zhuang R and Lederer J 2018 Maximum regularized likelihood estimators: a general prediction theory and applications *Stat* **7** e186