Functional linear and single-index models: A unified approach via Gaussian Stein identity

Krishnakumar Balasubramanian[1], Hans-Georg Müller[1], and Bharath K. Sriperumbudur[2]

[1]Department of Statistics, University of California, Davis
[2]Department of Statistics, Pennsylvania State University
[1]{kbala,hgmueller}@ucdavis.edu
[2]bks18@psu.edu

Abstract

Functional linear and single-index models are core regression methods in functional data analysis and are widely used for performing regression in a wide range of applications when the covariates are random functions coupled with scalar responses. In the existing literature, however, the construction of associated estimators and the study of their theoretical properties is invariably carried out on a case-by-case basis for specific models under consideration. In this work, assuming the predictors are Gaussian processes, we provide a unified methodological and theoretical framework for estimating the index in functional linear, and its direction in single-index models. In the latter case, the proposed approach does not require the specification of the link function. In terms of methodology, we show that the reproducing kernel Hilbert space (RKHS) based functional linear least-squares estimator, when viewed through the lens of an *infinite-dimensional Gaussian Stein's identity*, also provides an estimator of the index of the single-index model. Theoretically, we characterize the convergence rates of the proposed estimators for both linear and single-index models. Our analysis has several key advantages: (i) it does not require restrictive commutativity assumptions for the covariance operator of the random covariates and the integral operator associated with the reproducing kernel; and (ii) the true index parameter can lie outside of the chosen RKHS, thereby allowing for index misspecification as well as for quantifying the degree of such index misspecification. Several existing results emerge as special cases of our analysis.

1 Introduction

Functional regression with observed random functions as predictors coupled with scalar responses is one of the core tools of functional data analysis (Ramsay and Dalzell, 1991; Morris, 2015; Wang et al., 2016). The classical model of functional regression is the functional linear model, which emerges for example, when one assumes a joint Gaussian distribution between the predictor process $X(t)$ and response $Y \in \mathbb{R}$ and is given by

$$Y = \int_S X(t)\beta^*(t)\,dt + \epsilon = \langle X, \beta^* \rangle_{L^2(S)} + \epsilon, \tag{1.1}$$

where ϵ is an exogenous additive noise such that $\mathbb{E}[\epsilon|X] = 0$, $\mathbb{E}[\epsilon^2] = \sigma^2$. In the following, we set $S = [0, 1]$. A semi-parametric extension of the above model is the functional single-index model,

$$Y = g\left(\int_S X(t)\beta^*(t)\,dt\right) + \epsilon = g\left(\langle X, \beta^* \rangle_{L^2(S)}\right) + \epsilon, \tag{1.2}$$

for some function $g : \mathbb{R} \to \mathbb{R}$. Following standard terminology, the functional parameter β^* is referred to as the index parameter, and the function g as the link function. Note that when g is the identity function, the single-index model in (1.2) becomes the functional linear model.

Given n observations $\{(X_i, Y_i)\}_{1\le i\le n}$ that are independent and identically distributed copies of (X, Y), a fundamental problem is to estimate the index parameter β^* in (1.2). It is worth emphasizing here that for the case of single-index models, an efficient estimator for β^* is crucial for subsequently obtaining an estimate of the link function g; see, for example, Chen et al. (2011). Hence, our focus is on constructing an estimator of β^* that does not require information about the link function. To do so, the interaction between the allowed class of link functions and the distribution of the covariate X becomes crucial, as is also the case in the finite-dimensional case (Yang et al., 2017a). Indeed this has been well-explored in the case of multivariate single-index models (i.e., when $X \in \mathbb{R}^d$). As will be seen below, this is true in the functional setting as well.

In this work, under a Gaussian process assumption on the covariate X, we provide a unified reproducing kernel Hilbert space (RKHS) based framework for estimating the index in functional linear, the direction of the index in the single-index models, for a class of *unknown* link functions. Specifically, we illustrate that the standard functional linear least-squares estimator also provides an efficient estimator of the index parameter in the single-index model under the Gaussian process assumption. While it might come across as a rather surprising observation at first, it has an elementary justification when the functional linear least-squares estimator is viewed through the lens of *infinite-dimensional* analogs of Gaussian Stein's identity. Similar observations have been made in the multivariate setting (see, for example, Brillinger, 1983; Li and Duan, 1989; Plan and Vershynin, 2016 and Yang et al., 2017a), based on the *finite-dimensional* Gaussian Stein's identity. As our index parameter estimator is agnostic to the choice of the link function, it also naturally handles misspecification with respect to the link function. Furthermore, compared to existing theoretical results, our analysis handles the case when the true index β^* is not necessarily contained in the RKHS that is used for estimation.

1.1 Main contributions

We now elaborate more on our main contributions in this work.

Methodology: The RKHS-based functional penalized linear least-squares estimator (for a penalty parameter $\lambda > 0$), given by

$$\hat{\beta} \coloneqq \operatorname*{arg\,min}_{\beta\in\mathcal{H}} \ \frac{1}{n}\sum_{i=1}^{n} [Y_i - \langle \beta, X_i\rangle]^2 + \lambda\|\beta\|_{\mathcal{H}}^2, \tag{1.3}$$

was proposed and analyzed in Yuan and Cai (2010) for the linear setting, where $\mathcal{H}$ corresponds to an RKHS with the associated norm $\|\cdot\|_{\mathcal{H}}$; see Steinwart and Christmann (2008, Chapter 4) for an introduction to RKHS. While the minimization of the regularized objective in (1.3) is over a possibly infinite-dimensional RKHS $\mathcal{H}$, using the classical ideas of the *representer theorem*, it has been shown in Yuan and Cai (2010) that the minimizer of (1.3) can be computed by solving a finite-dimensional regularized linear inverse problem. In the current work, we illustrate that the above estimator that was designed for linear models, rather surprisingly also serves as a good estimator for the direction of the index in functional single-index models for various link functions g, when the random covariates X_i follow a Gaussian process. This provides a unified methodology for estimating the index parameter in functional linear and single-index settings, without regard to the specific nature of the link function g, thereby allowing for its misspecification.

Our proposed estimator is based on infinite-dimensional extensions of Gaussian Stein's identity. This goes informally as follows: For a zero-mean Gaussian random element X in a separable Hilbert space with covariance operator C, i.e., the linear integral operator with kernel $\mathrm{cov}(X(t), X(s))$, and for smooth enough real-valued functions f, it holds that

$$\mathbb{E}[Xf(X)] = C\mathbb{E}[\nabla f(X)], \tag{1.4}$$

where ∇ is the Fréchet derivative. Replacing $f(X)$ in (1.4) by $g(\langle X, \beta^* \rangle)$, we obtain $\mathbb{E}[Xg(\langle X, \beta^* \rangle)] = C\mathbb{E}[\nabla g(\langle X, \beta^* \rangle)]$, with the left hand side being equivalent to $\mathbb{E}[YX]$ since $\mathbb{E}[YX] = \mathbb{E}[Xg(\langle X, \beta^* \rangle) + X\varepsilon] = \mathbb{E}[Xg(\langle X, \beta^* \rangle)]$ and the right hand side being equivalent to $\vartheta_{g,\beta^*} C\beta^*$ with the constant defined as $\vartheta_{g,\beta^*} := \mathbb{E}[g'(\langle X, \beta^* \rangle)]$ since $C\mathbb{E}[\nabla g(\langle X, \beta^* \rangle)] = C\mathbb{E}[g'(\langle X, \beta^* \rangle)\beta^*]$, where g' is the derivative of g. Therefore, for the single-index model, the Gaussian Stein identity reduces to

$$\mathbb{E}[YX] = \vartheta_{g,\beta^*} C\beta^*, \tag{1.5}$$

which in turn reduces to the classical "functional normal equation" $C\beta^* = \mathbb{E}\,[YX]$ when the model is linear, i.e., $\vartheta_{g,\beta^*} = 1$ when g is linear; see, for example, He et al. (2000). This provides a justification for using the estimator in (1.3) for estimating the direction of the index in the context of single-index models, as long as $\vartheta_{g,\beta^*} \neq 0$. To the best of our knowledge, applying this viewpoint and Stein's identity is novel, even in the face of the extensive literature on functional linear and single-index regression models. We also emphasize that the use of Stein's identity in this context enables one to work only with unconditional covariance operators, which is in stark contrast with sufficient dimensionality reduction techniques (for example, sliced inverse regression) that require conditional covariance operators to be estimated. The constraint $\vartheta_{g,\beta^*} \neq 0$ places restrictions on the class of link functions g for which the proposed approach could be used. Examples of link functions for which the constraint holds include logistic functions, odd-powered polynomials, and exponential functions. However, an important function for which it does not hold is the quadratic (or even-powered polynomials). This is due to the fact that the odd moments of Gaussians are zero.

We note that in the following developments, it is assumed throughout that the predictor process X is a zero-mean Gaussian process that is fully observed. However, the assumption that the process is fully observed may be too strict for some relevant applications, where the functional predictors X_i are observed not continuously but rather intermittently on a dense grid of equidistant design points $(t_1, \ldots, t_m)$ on the domain S. Also, the measurements taken at these gridpoints may be contaminated with i.i.d. measurement errors ϵ_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, i.e., one has m (or more) measurement times t_{ij} that form a dense grid on the domain of the predictor functions X. In this situation, the data is observed as $X_{ij} = X_i(t_{ij}) + \epsilon_{ij}$ instead of complete trajectories X_i. One can then utilize a uniform convergence result that states that when passing the data (t_{ij}, X_{ij}) through a local linear smoother that utilizes appropriate bandwidth choices, one may obtain for the resulting curve estimates $\hat{X}_{i}$,: for any $\varepsilon > 0$, $\sup_{t \in S} |\hat{X}_i(t) - X_i(t)| = O_p(m^{-1/(3+\varepsilon)})$. The bound on the right-hand side does not depend on i and can be made arbitrarily small by assuming very dense sampling of individual trajectories and thus the error induced by the pre-smoothing step becomes negligible if m is large enough relative to the sample size n. In order not to detract from the main theme of this paper, we refer for further details about the necessary regularity conditions to Corollary 2 in Chen and Müller (2023); see also Müller et al. (2006) and Hall and Van Keilegom (2007) for earlier studies on pre-smoothing of functional data.

Theory: While the true index parameter β^* could lie either inside or outside the RHKS under consideration (characterized via interpolation spaces determined by a parameter α), the estimator

$\hat{\beta}$ in (1.3) always lies in the RKHS by definition. Previous works (Yuan and Cai, 2010; Cai and Yuan, 2012; Tong and Ng, 2018) relied on the assumption that $\beta^* \in \mathcal{H}$. In this work, we relax this assumption and obtain convergence rates for estimating β^* using $\hat{\beta}$ by capturing the interaction between the integral operator T associated with the RKHS and the covariance operator C of the Gaussian process, through the decay behavior of the eigenvalues of the operator $\Lambda := T^{1/2}CT^{1/2}$ and the alignment of its eigenfunctions to those of T. Our main result (Theorem 3.1), stated informally below, captures the interaction between T and C. Specialized versions (Theorems 3.2, 3.3, and 3.4) provide the rate of convergence for $\hat{\beta}$ for both linear and single-index models under appropriate eigenvalue decay assumptions. For the case of a linear model, using a variation of Theorem 3.1 (see Theorem 5.1), we also provide prediction error results without assuming $\beta^* \in \mathcal{H}$ in Theorem 5.3 and recover the results of Cai and Yuan (2012) in Theorem 5.2.

Main result (Informal)**.** *Define $\tilde{\beta}^* := \vartheta_{g,\beta^*}\beta^*$, where ϑ_{g,β^*} is a model constant that depends on g and β^*. Suppose $\tilde{\beta}^* \in \mathscr{R}(T^\alpha)$ for $\alpha \in (0, 1/2]$,*

$$\mathbb{E}\left[\left(g(\langle X, \tilde{\beta}^*\rangle) - \langle X, \tilde{\beta}^*\rangle\right)^4\right] < \infty,$$

and $\operatorname{trace}(C^{1/2}) < \infty$*. Then, defining $\Lambda := T^{1/2}CT^{1/2}$ and $\Lambda_\lambda := \Lambda + \lambda I$, up to constants, with high probability, we have*

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim \textsc{bias}(\lambda) + \|\Xi\|^{\frac{1}{4}}\left[\sqrt{\frac{N(\lambda)}{n}} + \lambda\sqrt{\frac{\|\Theta\|\operatorname{trace}(\Theta)}{n}}\right],$$

where $\Xi := T\Lambda_\lambda^{-2}T$, $N(\lambda) := \operatorname{trace}\left(\Lambda_\lambda^{-1}\Lambda\right)$, $\Theta := T^{\alpha-\frac{1}{2}}\Lambda_\lambda^{-1}\Lambda\Lambda_\lambda^{-1}T^{\alpha-\frac{1}{2}}$, and the bias factor is given by $\textsc{bias}(\lambda) := \|T^{1/2}\Lambda_\lambda^{-1}T^{1/2}C\tilde{\beta}^* - \tilde{\beta}^*\|$.

In the above result, the definition of $\tilde{\beta}^*$ allows to treat both single-index and linear models in a unified manner with $\vartheta_{g,\beta^*} = 1$ when the model is linear. The case $\alpha = 1/2$ corresponds to $\tilde{\beta}^* \in \mathcal{H}$ and $\alpha < 1/2$ corresponds to $\tilde{\beta}^* \in L^2(S)\backslash\mathcal{H}$. The smoothness of the target function $\tilde{\beta}^*$ is captured by α, i.e., larger values of α are associated with smoother values of $\tilde{\beta}^*$, where α controls the behavior of $\textsc{bias}(\lambda)$ and Θ. The behavior of $\|\hat{\beta} - \tilde{\beta}^*\|$ is also controlled by the decay rate of the eigenvalue of Λ, which in turn is related to the smoothness of $\mathcal{H}$ and that of the covariance function of the Gaussian process. Finally, the degree of alignment between the eigenfunctions of Λ and T controls the behavior of $\textsc{bias}(\lambda)$ and Ξ.

We now highlight the differences and benefits of our results compared to the directly related works of Yuan and Cai (2010), Cai and Yuan (2012) and Tong and Ng (2018) that consider only the functional *linear* regression in the RKHS setup. All these works consider the estimator in (1.3) and predominantly provide convergence results for the easier case of prediction error in the linear setup. While Yuan and Cai (2010) consider the problem of estimation in the linear setting, they make the restrictive assumption that the operators T and C commute. Cai and Yuan (2012) and Tong and Ng (2018) do not make the commutativity assumption, however, they do not provide any results for estimation. Furthermore, all these works assume the true index parameter β^* to reside inside the RKHS under consideration. In comparison, our results provide a complete characterization of the rates of estimation without the above assumptions, for both linear and single-index models.

1.2 Related works

The functional linear model (i.e., the case when g is the identity function) was derived for the Gaussian case by Grenander (1950) and in various statistical settings was considered by many authors, with early work by Engle et al. (1986), motivated by analyzing the relation between weather and electricity sales. Subsequent work includes Ramsay and Dalzell (1991); Brumback and Rice (1998); Cardot et al. (1999); Cuevas et al. (2002); Cardot et al. (2003) and Zhu et al. (2014), to mention a few; reviews include Morris (2015) and Wang et al. (2016).

Regarding single-index models, James (2002) and Müller and Stadtmüller (2005) studied functional versions of generalized linear models that feature a single-index where the latter work included nonparametric estimation of the link function. In subsequent work, Chen et al. (2011) studied estimators for general single and multiple index models and provided consistency results, while Shang and Cheng (2015) developed predictive inferential results for generalized linear models when the true index β^* lies in Sobolev RKHS spaces (which is a special case of the well-specified setting). Their approach follows that of Yuan and Cai (2010) and consequently suffers from similar shortcomings as discussed above. Furthermore, the estimators in the above works depend on the specification of the link function g. Several works, for example, Hsing and Ren (2009); Li and Song (2017); Jiang and Wang (2011); Li and Hsing (2010) and Jiang et al. (2014), also considered extension of sufficient dimension reduction methods to the functional data setting, however, only consistency of the estimator is established with no deeper study of convergence rates.

Finally, in the setting of multivariate data, the use of Stein's identity in developing estimation methodology for single and multiple-index models has been well explored. Specifically, we refer to Li and Duan (1989); Plan and Vershynin (2016); Yang et al. (2017a); Goldstein et al. (2018) and Goldstein and Wei (2019) for the case of single-index models. Similarly, we refer to Li (1991, 1992); Yang et al. (2017b) and Babichev and Bach (2018) for multiple-index models.

The techniques used in our analysis have a connection to the statistical learning theory literature on analyzing kernel ridge regression methods and linear inverse problems. A comprehensive operator theoretic analysis of kernel ridge regression was provided in Caponnetto and De Vito (2007) building on the seminal works of Cucker and Smale (2002); De Vito et al. (2005) and Smale and Zhou (2005). We also refer the interested reader to the works of Wu et al. (2006); Smale and Zhou (2007); Wang and Zhou (2011); Hsu et al. (2014); Dicker et al. (2017) and Lin et al. (2020) for other related works. Compared to our work, the above works are predominantly focused on excess error bounds for nonlinear regression in the learning theory framework. Another key difference of these works from ours is that they do not involve the covariance operator C and all results are determined by the integral operator T in contrast to ours which depends on the behavior of $T^{1/2}CT^{1/2}$ and its interaction with T. In the context of linear inverse problems, Blanchard and Mücke (2018) recently used operator-theoretic analysis to also provide estimation error bounds. However, their setting is not directly comparable to our setting of functional linear and single-index model regression.

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we elaborate our unified methodology highlighting the viewpoint obtained by the infinite-dimensional Gaussian Stein's identity. In Section 3, we present our unified theoretical results for estimating the index parameter. In Section 4, we provide examples to interpret and illustrate the assumptions required to derive the main results. In Section 5, the consequences of our results for prediction in the linear setting are highlighted. In Section 6, numerical simulations for both the Gaussian and non-Gaussian settings are provided. The proofs of all the results are provided in Section 7 and the auxiliary results are provided in Sections A and B.

1.4 Notations

Unless mentioned explicitly, $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ refer to $\langle \cdot, \cdot \rangle_{L^2(S)}$ and $\| \cdot \|_{L^2(S)}$, respectively. We also require the RKHS inner-product and the associated norm sometimes, which we refer to by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ respectively. For an operator A, we denote by $\mathscr{R}(A)$ and $\mathscr{N}(A)$, its range space and the null space respectively. We denote the operator norm of A by $\|A\|$. For $x \in H$, $x \otimes x : H \to H$ is defined as $(x \otimes x)z = x\langle x, z\rangle_H$ for any $z \in H$, where H is a Hilbert space. We also use $a \lesssim b$ to represent that $a \le Kb$ for a large enough constant K. Furthermore, for a random variable $\chi \in \mathbb{R}$ with distribution P and a constant x, we use $\chi \lesssim_p x$ to denote the fact that for any $\delta > 0$, there exists a positive constant $K_\delta < \infty$ such that $P(\chi \le K_\delta x) \ge \delta$.

2 Methodology

Let $\mathcal{H}$ be an RKHS with the associated kernel $k : S \times S \to \mathbb{R}$. Define $\mathfrak{I} : \mathcal{H} \to L^2(S)$, $f \mapsto f$, to be the inclusion operator mapping functions in the RKHS $\mathcal{H}$ to $L^2(S)$ and $\mathfrak{I}^* : L^2(S) \to \mathcal{H}$ to denote the adjoint of $\mathfrak{I}$. We also define the following two important operators that arise in our analysis:

$$T := \mathfrak{I}\mathfrak{I}^* : L^2(S) \to L^2(S) \quad \text{and} \quad C := \mathbb{E}[X \otimes X] : L^2(S) \to L^2(S), \tag{2.1}$$

where $\otimes$ represents the $L^2(S)$ tensor product. Throughout the paper, we assume that X is a centered random element, i.e., $\mathbb{E}[X] = 0_{L_2(S)}$.

Our goal is to estimate β^* in the presence of an unknown link function g. First note that one may view the Gaussian processes X as random elements taking values in Hilbert space following a Gaussian measure (Rajput, 1972; Rajput and Cambanis, 1972; Grenander, 2008; Hsing and Eubank, 2015). As discussed in Section 1.1, leveraging the version of Stein's identity for Hilbert-valued random elements yields (1.5), i.e., $\mathbb{E}[YX] = \vartheta_{g,\beta^*} C\beta^*$. We refer to Shih (2011) and Kuo and Lee (2011) for the details of the infinite-dimensional Gaussian Stein's identity (see (1.4)) and the associated integration by parts formula and provide a formal statement in the supplementary material for completeness. Here ϑ_{g,β^*} is a constant depending on the link function g and the index β^*. The exact value of the constant could be calculated for a given fixed link function g, which also fixes the true index parameter β^*. However, the exact value is irrelevant for our purpose as we focus on estimating the direction of the index parameter (which is the best one could hope for without the knowledge of g, due to the lack of identifiability in the model). Hence, we assume throughout that g is such that $\vartheta_{g,\beta^*} \neq 0$, where $\vartheta_{g,\beta^*} := \mathbb{E}[g'(\langle X, \beta^* \rangle)]$ is recalled from Section 1.1. In particular, when g is the identity function, it is easy to see that $\vartheta_{g,\beta^*} = 1$. We define $\tilde{\beta}^* := \vartheta_{g,\beta^*}\beta^*$ to handle the single-index and linear model in a unified manner and note that

$$\tilde{\beta}^* := \arg \min_{\beta \in L^2(S)} \mathbb{E}\left[Y - \langle X, \beta \rangle\right]^2.$$

This variational formulation is the key step in constructing a regularized estimator as in (1.3), whose details are provided below.

Let $(X_1, Y_1), \ldots (X_n, Y_n)$ be n i.i.d. copies of random variables (X, Y). Recalling the definitions in and above (2.1), for some $\lambda > 0$, our estimator is based on minimizing the penalized least-squares criterion over the RKHS $\mathcal{H}$,

$$\begin{aligned} \hat{\beta}_{n,\lambda} &= \arg\min_{\beta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \langle \beta, X_i \rangle]^2 + \lambda \|\beta\|_{\mathcal{H}}^2 \\ &= \arg\min_{\beta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \langle \mathfrak{I}\beta, X_i \rangle]^2 + \lambda \|\beta\|_{\mathcal{H}}^2 \end{aligned}$$

$$= \operatorname*{arg\,min}_{\beta\in\mathcal{H}} \ \frac{1}{n}\sum_{i=1}^{n}[Y_i - \langle \beta, \mathfrak{I}^* X_i\rangle_{\mathcal{H}}]^2 + \lambda\|\beta\|^2_{\mathcal{H}} \tag{2.2}$$

$$= \operatorname*{arg\,min}_{\beta\in\mathcal{H}} \ \frac{1}{n}\sum_{i=1}^{n}\left[Y_i^2 + \langle \beta, (\mathfrak{I}^* X_i \otimes \mathfrak{I}^* X_i)\,\beta\rangle_{\mathcal{H}} - 2\langle Y_i \mathfrak{I}^* X_i, \beta\rangle_{\mathcal{H}}\right] + \lambda\|\beta\|^2_{\mathcal{H}}. \tag{2.3}$$

This estimator does not require any knowledge of the link function g. Indeed, for the case where g is the identity, this estimator was studied in Yuan and Cai (2010) for estimation in the functional linear model. As we will demonstrate, this same estimator continues to be applicable for a general single-index model, under the assumption that X is a Gaussian process, where we focus on the estimation of the direction of the true index parameter β^*.

By completing the squares w.r.t. β in (2.3), it is easy to verify that

$$\hat{\beta}_{n,\lambda} = \left[\mathfrak{I}^*\left(\frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i\right)\mathfrak{I} + \lambda I\right]^{-1} \mathfrak{I}^*\left[\frac{1}{n}\sum_{i=1}^{n} Y_i X_i\right].$$

Defining $\hat{C} := \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i$ and $\hat{R} := \frac{1}{n}\sum_{i=1}^{n} Y_i X_i$, we obtain

$$\hat{\beta}_{n,\lambda} = \left[\mathfrak{I}^*\hat{C}\mathfrak{I} + \lambda I\right]^{-1}\mathfrak{I}^*\hat{R}. \tag{2.4}$$

In what follows, we denote $\hat{\beta}_{n,\lambda}$ by $\hat{\beta}$ for simplicity. We emphasize that the above form of the estimator is useful for our analysis, while an alternate form that is more useful for implementation purposes is as follows. By applying the representer theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001) to (2.2), $\hat{\beta} \in \operatorname{span}\left\{\int_S k(\cdot,t)X_i(t)\,dt : i = 1,\dots,n\right\}$, i.e., there exists a $\boldsymbol{\alpha} := (\alpha_1,\dots,\alpha_n)^\top \in \mathbb{R}^n$ such that $\hat{\beta} = \sum_{i=1}^{n}\alpha_i \int_S k(\cdot,t)X_i(t)\,dt$. Using this in (2.2) and solving for $\boldsymbol{\alpha}$ yields $\boldsymbol{\alpha} = (\boldsymbol{K} + n\lambda I)^{-1}\boldsymbol{y}$, where $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ with $[\boldsymbol{K}]_{ij} := \int_S\int_S k(t,s)X_i(t)X_j(t)\,dt\,ds$ and $\boldsymbol{y} = (Y_1,\dots,Y_n)^\top \in \mathbb{R}^n$. Therefore, $\hat{\beta}$ can be computed by solving a finite-dimensional linear system of size n, which is not obvious from the expression in (2.4).

3 Main results

Here we present our main results concerning the rate of convergence of $\hat{\beta}$ to $\tilde{\beta}^*$ in $L^2(S)$. Theorem 3.1 (proved in Section 7.1) is a general result about the behavior of $\|\hat{\beta} - \tilde{\beta}^*\|$ in terms of certain key operators involving T and C. More specialized results are presented in Theorems 3.2, 3.3 and 3.4, depending on whether T and C commute or not.

Theorem 3.1 (Master theorem for estimation)**.** *Let $\|T^{-\alpha}\tilde{\beta}^*\| < \infty$, i.e., $\tilde{\beta}^* \in \mathscr{R}(T^\alpha)$ for $\alpha \in (0, 1/2]$. Define*

$$\varkappa := \mathbb{E}\left[\left(g(\langle X, \tilde{\beta}^*\rangle) - \langle X, \tilde{\beta}^*\rangle\right)^4\right]. \tag{3.1}$$

Suppose one of the following conditions hold: (a) $\operatorname{trace}(C^{1/2}) < \infty$ *and $\varkappa \in (0,\infty)$, (b) $\varkappa = 0$ and* $\operatorname{trace}(C) < \infty$*. Define*

$$\Theta := T^\alpha(CT + \lambda I)^{-1}C(TC + \lambda I)^{-1}T^\alpha, \qquad d(\lambda) := \frac{\operatorname{trace}(\Theta)}{\|\Theta\|}, \tag{3.2}$$

$$\Xi := T(T^{1/2}CT^{1/2} + \lambda I)^{-2}T, \quad \textit{and}$$

$$N(\lambda) := \text{trace}\left[(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^{1/2}\right]. \tag{3.3}$$

Then, for

$$\delta \in (0,1/e], \qquad n \gtrsim (d(\lambda) \vee \log(1/\delta)), \quad \textit{and}$$
$$\frac{\text{trace}(T^{1/2}CT^{1/2})}{n} \lesssim \lambda \lesssim \|T^{1/2}CT^{1/2}\|, \tag{3.4}$$

with probability at least $1-3\delta$*, we have*

$$\begin{aligned}\|\hat{\beta}-\tilde{\beta}^*\| &\lesssim \text{BIAS}(\lambda) + \|\Xi\|^{\frac{1}{4}}\sqrt{\frac{(\sigma^2+\sqrt{\varkappa})N(\lambda)}{n\delta}} \\ &+ \lambda\|\Xi\|^{\frac{1}{4}}\left(\left\|T^{1/2}CT^{1/2}\right\|^{1/2}+\sqrt{\lambda}\right)\|T\|^{\frac{1}{2}-\alpha}\|T^{-\alpha}\tilde{\beta}^*\|\sqrt{\frac{\|\Theta\|\text{trace}(\Theta)}{n}},\end{aligned} \tag{3.5}$$

where $\text{BIAS}(\lambda) := \|T(CT+\lambda I)^{-1}C\tilde{\beta}^*-\tilde{\beta}^*\|$.

Remark 3.1. (i) The assumption $\tilde{\beta}^* \in \mathscr{R}(T^\alpha)$ imposes certain smoothness condition on $\tilde{\beta}^*$. For example, it is well-known (Steinwart and Christmann, 2008, Theorem 4.51) that $\tilde{\beta}^* \in \mathcal{H}$ when $\alpha = \frac{1}{2}$, which we refer to as the *well-specified setting*. This assumption is equivalent to the condition that $\tilde{\beta}^*$ lies in an interpolation space between $L^2(S)$ and $\mathcal{H}$ with α being the interpolating index.

(ii) While $\text{trace}(C) < \infty$ is guaranteed by the well-definedness of the Gaussian process, Theorem 3.1 requires a slightly stronger condition, namely $\text{trace}(C^{1/2}) < \infty$, when $\varkappa \neq 0$.

(iii) The parameter $\varkappa$ captures the degree of non-linearity of the model. Indeed, $\varkappa = 0$ implies $g(\langle X, \tilde{\beta}^*\rangle) = \langle X, \tilde{\beta}^*\rangle$ with probability 1. Conversely, when the model is linear, $\varkappa = 0$. For the non-linear case, the condition of $\varkappa < \infty$ is rather mild since it is satisfied by any g that satisfies $g(x) = o(e^{x^{2+\epsilon}})$ as $x \to \infty$ for any $\epsilon > 0$. Since $\langle X, \tilde{\beta}^*\rangle$ is a zero mean Gaussian random variable, clearly, $\varkappa < \infty$ if $\mathbb{E}[g^4(Z)] < \infty$ which is true if the above condition holds.

The following result (proved in Section 7.2) provides a concrete convergence rate when the operators T and C commute.

Theorem 3.2 (Commutative operators)**.** *Let* $\|T^{-\alpha}\tilde{\beta}^*\| < \infty$ *for* $\alpha \in (0,1/2]$. *Assume that the operators* T *and* C *commute and have simple eigenvalues (i.e., of multiplicity one) denoted by* μ_i *and* ξ_i *for* $i \in \mathbb{N}$*, such that*

$$i^{-t} \lesssim \mu_i \lesssim i^{-t} \quad \textit{and} \quad i^{-c} \lesssim \xi_i \lesssim i^{-c}, \tag{3.6}$$

where $t > 1$ *and* $c > 1$. *Suppose one of the following conditions hold: (a)* $\varkappa \in (0,\infty)$ *and* $c > 2$*, (b)* $\varkappa = 0$ *and* $c > 1$. *Then*

$$\|\hat{\beta}-\tilde{\beta}^*\| \lesssim_p n^{-\frac{\alpha t}{1+c+2t(1-\alpha)}} \quad \textit{for} \quad \lambda = n^{-\frac{t+c}{1+c+2t(1-\alpha)}}. \tag{3.7}$$

Remark 3.2. (i) When $\alpha = 1/2$, i.e., $\tilde{\beta}^* \in \mathcal{H}$ (well-specified case), we obtain

$$\|\hat{\beta}-\tilde{\beta}^*\| \lesssim_p n^{-\frac{t}{2(1+t+c)}},$$

which exactly matches the minimax optimal rate obtained in Yuan and Cai (2010) for the functional linear model. Remarkably, this same rate applies in the much more general framework of a single index functional regression model when $c > 2$ and $\varkappa < \infty$.

(ii) Even for the special case of the functional linear model, Theorem 3.2 extends the results of Yuan and Cai (2010) to the misspecified setting, i.e., $\tilde{\beta}^* \in L^2(S)\backslash\mathcal{H}$, since Yuan and Cai (2010) only investigated the well-specified setting (i.e., $\tilde{\beta}^* \in \mathcal{H}$).

(iii) The term α controls the smoothness of $\tilde{\beta}^*$ with large values of α corresponding to smooth $\tilde{\beta}^*$. Therefore, we should expect the convergence rates to get faster with increasing α, which is confirmed by Theorem 3.2. However, based on our current proof technique, Theorem 3.2 handles the range of smoothness corresponding to $\alpha \in (0, 1/2]$. The case of $\alpha > 1/2$ remains open and is an artifact of our proof technique.

(iv) The requirement $c > 2$ ensures that $\text{trace}(C^{1/2}) < \infty$.

In the following, we relax the assumption of commutativity of C and T and investigate the convergence rates for $\|\hat{\beta} - \tilde{\beta}^*\|$ by directly exploiting the eigenvalue decay of $T^{1/2}CT^{1/2}$ in Theorem 3.3 (proved in Section 7.3). Under additional assumptions about the alignment between the eigenfunctions of $T^{1/2}CT^{1/2}$ and T faster convergence rates can be obtained. This is the result stated in Theorem 3.4 (proved in Section 7.4) and the rates there are seen to be faster than those in Theorem 3.3, while both these convergence rates are slower than those obtained in Theorem 3.2 because the commutativity assumption is stronger than these relaxed assumptions.

Theorem 3.3 (Noncommutative operators)**.** *Let $(\zeta_i)_{i\in\mathbb{N}}$ denote the eigenvalues of $T^{1/2}CT^{1/2}$ with $i^{-b} \lesssim \zeta_i \lesssim i^{-b}$, for some $b > 1$. Suppose $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ for $\nu \in (0, 1]$ and $\varkappa < \infty$. Then, for*

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim_p n^{-\frac{b\nu}{1+b+2b\nu}} \quad \textit{for} \quad \lambda = n^{-\frac{b}{1+b+2b\nu}}. \tag{3.8}$$

Remark 3.3. (i) Unlike in the commutative case, the results are presented in terms of the eigen decay behavior of $T^{1/2}CT^{1/2}$. When T and C commute, we obtain $b = t + c$.

(ii) To the best of our knowledge, to date there is no result available in the literature for the estimation error $\|\hat{\beta} - \tilde{\beta}^*\|$ in the noncommutative setting, even for the special case of functional linear models.

(iii) The assumption $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ implies there is a function $h \in L^2(S)$ such that

$$T^{1/2}(T^{1/2}CT^{1/2})^{\nu}h = \tilde{\beta}^*,$$

whence $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}) = \mathcal{H}$ (i.e., $\alpha = 1/2$ in Theorems 3.1 and 3.2). Therefore, the assumption $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ is stronger than assuming $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2})$. The key reason for this assumption is to obtain sharper bounds of $\textsc{bias}(\lambda)$, thus obtaining non-trivial convergence rates. Indeed, simply assuming $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2})$ ensures $\textsc{bias}(\lambda) \to 0$ as $\lambda \to 0$, and consistency of $\hat{\beta}$ can be established, but with no handle on the convergence rate.

(iv) While it is difficult to grasp the smoothness properties of $\tilde{\beta}^*$ entailed by the condition $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ in the noncommutative setting, an understanding of this condition can be gained for the special case where T and C do commute. In this setting, when the eigenvalues of T and C satisfy the conditions of Theorem 3.2, the assumption $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ is equivalent to $\tilde{\beta}^* \in \mathscr{R}\,(T^{\frac{1}{2}+\nu+\frac{c\nu}{t}}) \subset \mathscr{R}\,(T^{1/2})$, which implies that $\tilde{\beta}^*$ is restricted to a smaller subspace of $\mathcal{H}$. The larger the values of ν or $\frac{c}{t}$ are, the smaller is this subspace of $\mathcal{H}$. This means that $\tilde{\beta}^*$ is smoother when ν increases and when $\nu > 0$ as compared to $\nu = 0$ (where only $\nu = 0$ is actually needed in the commutative case).

(v) Denoting the eigenfunctions of $T^{1/2}CT^{1/2}$ by $(\phi_i)_{i\in\mathbb{N}}$, in the commutative case, the assumption that $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ implies that the bias term behaves as

$$\begin{aligned}
\textsc{bias}(\lambda) &= \|T(CT+\lambda I)^{-1}C\tilde{\beta}^* - \tilde{\beta}^*\| \\
&= \|T(CT+\lambda I)^{-1}CT^{1/2}(T^{1/2}CT^{1/2})^{\nu}h - T^{1/2}(T^{1/2}CT^{1/2})^{\nu}h\| \\
&\le \left[\sum_i \left[\frac{i^{-t-t/2-c-\nu(t+c)}}{i^{-(t+c)}+\lambda} - i^{-t/2-\nu(t+c)}\right]^2 \langle h, \phi_i\rangle^2\right]^{1/2} \\
&\le \lambda\left[\sup_i \frac{i^{-t/2-\nu(t+c)}}{i^{-(t+c)}+\lambda}\right]\|h\| \le \; \lambda\left[\lambda^{\frac{t/2+\nu(t+c)-(t+c)}{t+c}}\right] = \; \lambda^{\nu+\frac{t}{2(t+c)}},
\end{aligned}$$

where the last inequality follows from Lemma A.6 when $(t+c)(1-\nu) \ge t/2$, i.e., $\nu \le \frac{t+2c}{2t+2c}$. Note that this upper bound is better than $\lambda^{\frac{t}{2(t+c)}}$ when $\alpha = 1/2$. Hence, we obtain

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim_p \frac{\lambda^{-\frac{1+c}{2(t+c)}}}{\sqrt{n}} + \lambda^{\nu+\frac{t}{2(t+c)}},$$

where the first term is directly taken from the proof of Theorem 3.2 under $\alpha = 1/2$. Therefore,

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim_p n^{-\frac{\nu(t+c)+\frac{t}{2}}{2\nu(t+c)+1+c+t}} \quad \text{for} \quad \lambda = n^{-\frac{t+c}{2\nu(t+c)+1+t+c}}.$$

On the other hand, the bound in Theorem 3.3 under the commutativity assumption, i.e., $b = t+c$ yields

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim_p n^{-\frac{\nu(t+c)}{2\nu(t+c)+1+c+t}} \quad \text{for} \quad \lambda = n^{-\frac{t+c}{2\nu(t+c)+1+t+c}}.$$

Thus the bound in Theorem 3.2 is better than the one in Theorem 3.3, as expected.

As can be seen from the proof of Theorem 3.3, the terms $\|\Xi\|$ and $\textsc{bias}(\lambda)$ with the assumption $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ involve interaction terms between T and $T^{1/2}CT^{1/2}$. In contrast, the terms $N(\lambda)$, $\|\Theta\|$ and $\operatorname{trace}(\Theta)$ are entirely determined by $T^{1/2}CT^{1/2}$. Theorem 3.3 ignores the interaction between T and $T^{1/2}CT^{1/2}$. It is of interest to investigate if more refined bounds than those in Theorem 3.3 can be obtained by additionally capturing interaction terms. To this end, let (ζ_i, ϕ_i) and (μ_i, ψ_i) for $i \in \mathbb{N}$ denote the eigensystems of $T^{1/2}CT^{1/2}$ and T, respectively. Then we have

$$\Xi = T(T^{1/2}CT^{1/2}+\lambda I)^{-2}T = T\left[\sum_i (\zeta_i+\lambda)^{-2}\phi_i \otimes \phi_i + \sum_i \lambda^{-2}\tilde{\phi}_i \otimes \tilde{\phi}_i\right]T,$$

where the $(\tilde{\phi}_i)_i$ span the null space $\mathscr{N}\,(T^{1/2}CT^{1/2})$ of $T^{1/2}CT^{1/2}$. Therefore,

$$\begin{aligned}
\|\Xi\| &= \left\|\sum_i (\zeta_i+\lambda)^{-2}T\phi_i \otimes T\phi_i + \sum_i \frac{1}{\lambda^2}T\tilde{\phi}_i \otimes T\tilde{\phi}_i\right\| \\
&\le \sum_i \frac{\|T\phi_i \otimes T\phi_i\|}{(\zeta_i+\lambda)^2} + \frac{1}{\lambda^2}\left\|\sum_i T\tilde{\phi}_i \otimes T\tilde{\phi}_i\right\| = \sum_i \frac{\|T\phi_i\|^2}{(\zeta_i+\lambda)^2} + \frac{1}{\lambda^2}\left\|\sum_i T\tilde{\phi}_i \otimes T\tilde{\phi}_i\right\| \\
&= \sum_i \frac{\left\|\sum_j \mu_j\langle\phi_i, \psi_j\rangle\psi_j\right\|^2}{(\zeta_i+\lambda)^2} + \frac{1}{\lambda^2}\left\|\sum_i T\tilde{\phi}_i \otimes T\tilde{\phi}_i\right\|.
\end{aligned}$$

Note that the first term in the above inequality can be further bounded as follows,

$$\sum_i \frac{\left\|\sum_j \mu_j \langle \phi_i, \psi_j\rangle \psi_j\right\|^2}{(\zeta_i+\lambda)^2} = \sum_i \frac{\mu_i^2}{(\zeta_i+\lambda)^2} \sum_j \frac{\mu_j^2}{\mu_i^2} \langle \phi_i, \psi_j\rangle^2$$
$$\leq \sum_i \frac{\mu_i^2}{(\zeta_i+\lambda)^2} \sup_i \frac{1}{\mu_i^2} \sum_j \mu_j^2 \langle \phi_i, \psi_j\rangle^2.$$

Under the assumption that $\sup_i \frac{1}{\mu_i^2}\sum_j \mu_j^2\langle \phi_i, \psi_j\rangle^2 < \infty$ (this condition captures the interaction between T and $T^{1/2}CT^{1/2}$ and is naturally satisfied when T and C commute), we obtain

$$\sum_i \frac{\left\|\sum_j \mu_j \langle \phi_i, \psi_j\rangle \psi_j\right\|^2}{(\zeta_i+\lambda)^2} \lesssim \sum_i \frac{\mu_i^2}{(\zeta_i+\lambda)^2} \lesssim \sum_i \frac{i^{-2t}}{(i^{-b}+\lambda)^2} \lesssim \lambda^{-\frac{1+2b-2t}{b}},$$

where the last inequality follows from Lemma A.5 when $b \geq 2t$ and $b \geq t$, i.e., $b \geq 2t$. Therefore,

$$\|\Xi\| \lesssim \lambda^{-\frac{1+2b-2t}{b}} + \lambda^{-2} \lesssim \lambda^{-2}$$

since the first term is of smaller order than λ^{-2} as $\lambda \to 0$.

This shows that because of the interaction between T and $\mathscr{N}(T^{1/2}CT^{1/2})$, it appears that a better bound is not possible for $\|\Xi\|^{1/4}$, as we showed in the proof of Theorem 3.3 (see (7.21)) that $\|\Xi\|^{1/4} \leq \lambda^{-1/2}$ without capturing any interaction between T and $\mathscr{N}(T^{1/2}CT^{1/2})$. On the other hand, the bound on $\textsc{bias}(\lambda)$ seems to be improvable. Indeed, note that as $\tilde{\beta}^* = T^{1/2}(T^{1/2}CT^{1/2})^{\nu}h$ for some $h \in L^2(S)$, we obtain $\textsc{bias}(\lambda) = \|T^{1/2}(\Lambda+\lambda I)^{-1}\Lambda^{1+\nu}h - T^{1/2}\Lambda^{\nu}h\|$ with $\Lambda := T^{1/2}CT^{1/2}$, where the interaction between T and $\mathscr{N}(T^{1/2}CT^{1/2})$ does not play a role. These observations lead to the following result (proved in Section 7.4), which is an improvement over Theorem 3.3.

Theorem 3.4 (Noncommutative operators with alignment of eigenfunctions)**.** *Let (ζ_i, ϕ_i) and (μ_i, ψ_i) for $i \in \mathbb{N}$, denote the eigensystems of $T^{1/2}CT^{1/2}$ and T respectively. Suppose*

$$i^{-b} \lesssim \zeta_i \lesssim i^{-b} \text{ and } i^{-t} \lesssim \mu_i \lesssim i^{-t}$$

for some $b, t > 1$ and that the eigenfunctions of $T^{1/2}CT^{1/2}$ and T satisfy

$$\sup_{i,l} \frac{1}{\mu_i\mu_l}\left|\sum_j \mu_j \langle \phi_i, \psi_j\rangle \langle \phi_l, \psi_j\rangle\right|^2 < \infty. \tag{3.9}$$

Assuming $\varkappa < \infty$ and $\tilde{\beta}^* \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})^{\nu})$ for some $\nu \in \left(0, \frac{1}{2} - \frac{t}{2b}\right]$, we have

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim_p n^{-\frac{b\nu + (t-1)/2}{t+b+2b\nu}} \quad \text{for} \quad \lambda = n^{-\frac{b}{t+b+2b\nu}}.$$

Remark 3.4. (i) For $\nu \in (0, \frac{1}{2} - \frac{t}{2b}]$, the rate in Theorem 3.4 is clearly faster than that in Theorem 3.3.

(ii) When T and C commute, the condition in (3.9) is satisfied as

$$\sup_{i,l} \frac{1}{\mu_i\mu_l}\left|\sum_j \mu_j \langle \phi_i, \psi_j\rangle \langle \phi_l, \psi_j\rangle\right|^2 = \sup_{i,l} \frac{1}{\mu_i\mu_l}\left|\sum_j \mu_j \langle \phi_i, \phi_j\rangle \langle \phi_l, \phi_j\rangle\right|^2$$

$$= \sup_i \frac{1}{\mu_i^2} \left| \sum_j \mu_j \langle \phi_i, \phi_j \rangle^2 \right|^2 = 1.$$

Since $b = t + c$ in the commutative setting, by setting $\lambda = n^{-\frac{t+c}{2t+c+2(t+c)\nu}}$, we obtain

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim_p n^{-\frac{(t+c)\nu+(t-1)/2}{2t+c+2(t+c)\nu}}.$$

This rate is still slower than the rate provided by Theorem 3.2, which is obtained directly under the commutativity assumption since the interaction between $T^{1/2}CT^{1/2}$ and T is not captured in Ξ.

4 Interpreting range space conditions on $\tilde{\beta}^*$

In this section, we provide an interpretation of the range space condition $\tilde{\beta}^* \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})^\nu)$, for $\nu \in (0, 1]$, for specific choices of covariance operator C and the kernel k that induces the integral operator T. The following result (proved in Section 7.5) provides a generic characterization of the range space condition, which is elaborated through examples. We consider the case $\nu = 1$ for simplicity.

Proposition 4.1. *For $x, y \in [0, 1]$, suppose that the reproducing kernel k and the covariance function c are given respectively by*

$$k(x, y) = \sum_{i \geq 1} a_i \phi_i(x)\phi_i(y), \quad c(x, y) = \sum_{m \geq 1} b_m \psi_m(x)\psi_m(x),$$

where $a_i \geq 0$ for all i, $b_m \geq 0$ for all m, $\sum_{i\geq 1} a_i \leq \infty$, $\sum_{m \geq 1} b_m \leq \infty$ and $(\phi_i)_i$ and $(\psi_m)_m$ form an orthonormal basis of $L^2([0, 1])$. Define $\tau_j := \sum_i a_i \eta_{ij}^2$ where $\eta_{ij} := \sum_{m\geq 1} b_m \theta_{mi}\theta_{mj}$ and $\theta_{mj} := \langle \psi_m, \phi_i \rangle$, and assume $\sup_j \tau_j < \infty$. Then it holds that

(i) The RKHS induced by the kernel k is given by

$$\mathcal{H} = \left\{ f(x) = \sum_{i \geq 1} f_i \phi_i(x), x \in [0, 1] : \sum_i \frac{f_i^2}{a_i} < \infty \right\},$$

with the associated inner product defined by $\langle f, g \rangle_{\mathcal{H}} = \sum_i a_i^{-1} f_i g_i$.

(ii) The space $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2}))$ satisfies the inclusion

$$\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})) \subset \tilde{\mathcal{H}} \subset \mathcal{H},$$

where

$$\tilde{\mathcal{H}} = \left\{ f(x) = \sum_i f_i \phi_i(x), x \in [0, 1] : \sum_i \frac{f_i^2}{a_i \tau_i} < \infty \right\},$$

is an RKHS induced by the kernel $\tilde{k}(x, y) = \sum_{i \geq 1} a_i \tau_i \phi_i(x)\phi_i(y)$ with inner product $\langle f, g \rangle_{\tilde{\mathcal{H}}} = \sum_{i\geq 1} f_i g_i (\tau_i a_i)^{-1}$.

Remark 4.1. While $T^{1/2}CT^{1/2}$ is a positive self-adjoint operator, its eigenvalues and eigenfunctions are unknown, see (7.28). If $\theta_{mi} = \delta_{mi}$, which happens when $\psi_m = \phi_i$ (i.e., in the commutative setting), then we obtain $\eta_{ij} = b_i\delta_{ij}$ and so $T^{1/2}CT^{1/2} = \sum_i a_ib_i\phi_i \otimes \phi_i$, yielding $(a_ib_i, \phi_i)_i$ as the eigensystem of $T^{1/2}CT^{1/2}$, which then implies that $(a_i^{3/2}b_i, \phi_i)_i$ is the eigensystem of $T^{1/2}(T^{1/2}CT^{1/2})$. Therefore, for any $f \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})^\nu)$, there is a $h \in L^2([0,1])$ such that we have $f = T^{1/2}(T^{1/2}CT^{1/2})^\nu h$. This implies $f = \sum_i a_i^{\nu+1/2}b_i^\nu h_i\phi_i$ where $h_i = \langle h, \phi_i\rangle$. It is easy to verify that $f \in \mathcal{H}'$, where

$$\mathcal{H}' = \left\{ f(x) = \sum_i f_i\phi_i(x), x \in [0,1] : \sum_i \frac{f_i^2}{a_i^{2\nu+1}b_i^{2\nu}} < \infty \right\},$$

which is an RKHS induced by the kernel $k'(x,y) = \sum_i a_i^{2\nu+1}b_i^{2\nu}\phi_i(x)\phi_i(y)$ since, we have that $f = T^{1/2}(T^{1/2}CT^{1/2})^\nu h$,

$$\|f\|^2_{\mathcal{H}'} = \sum_i \frac{a_i^{2\nu+1}b_i^{2\nu}h_i^2}{a_i^{2\nu+1}b_i^{2\nu}} = \sum_i h_i^2 = \|h\|^2 < \infty.$$

Remark 4.2. Suppose $\phi_i = \cos(i\pi\cdot)$ and $a_i \propto i^{-2\alpha}$, for some $\alpha \in \mathbb{N}$. Then, for $f \in \mathcal{H}$, we have $f(x) = \sum_i f_i\phi_i(x) = \sum_i f_i \cos(i\pi x)$, for $x \in [0,1]$ where $\sum_i i^{2\alpha} f_i^2 < \infty$. Note that we have $f^{(\alpha)}(x) = \sum_i \pi^\alpha i^\alpha f_i \cos(i\pi x)$, which implies $\|f^{(\alpha)}\|^2 = \pi^{2\alpha}\sum_i i^{2\alpha}f_i^2 = c_1\|f\|^2_{\mathcal{H}}$, for some constant $c_1 > 0$. That is, $\mathcal{H}$ consists of α-times differentiable functions that are square integrable. Suppose $b_i \propto i^{-2\lambda}$ for some $\lambda \in \mathbb{N}$. Then under the conditions of Proposition 4.1, we obtain that $\tilde{\mathcal{H}}$ consists of functions that are $(\alpha+\lambda)$-times differentiable and square-integrable, i.e., the degree of smoothness of $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2}))$ is at least λ more than that of $\mathcal{H}$.

Note that Mercer's theorem allows expansion of the kernel as in Proposition 4.1, wherein $(a_i, \phi_i)_i$ forms the eigensystem of the integral operator, $\mathcal{T}$. Since $S = [0,1]$ and the measure is Lebesgue, the choice of $\phi_i(t) = \cos(i\pi t)$ yields a translation-invariant kernel (see Remark 4.4) but other choices are possible that yield kernels that are not translation invariant, e.g., $a_i = \frac{1}{i!}$, $\phi_i(x) = x^i$ and $k(x,y) = e^{xy}$. We now consider concrete examples of covariance kernels and provide interpretations of the result in Proposition 4.1. We let $\phi_i(x) = \cos(i\pi x)$, $x \in [0,1]$.

Example (Fourier basis). Suppose $\psi_m = \cos(\omega_m\pi\cdot)$ where $\omega_m = am + b$ for some $a, b \in \mathbb{R}$ such that $\omega_m \notin \mathbb{Z}$ and $m \in \mathbb{N}$. Let $b_m \lesssim m^{-(1+\delta)}$, for some $\delta > 0$. In fact, one can assume without loss of generality that $\omega_m > 0$ for all $m \in \{1,2,3,\ldots\}$ or equivalently $a > 0$. Then by Lemma A.7, we have

$$\begin{aligned}\theta_{mi} &= \langle\psi_m, \phi_i\rangle \\ &= \int \cos(\omega_m\pi x)\cos(i\pi x)dx \\ &= \frac{i\pi}{(i\pi)^2 - \omega_m^2\pi^2}\cos(\pi\omega_m)\sin(i\pi) - \frac{\omega_m\pi}{(i\pi)^2 - \omega_m^2\pi^2}\sin(\pi\omega_m)\cos(i\pi) \\ &= \frac{\pi\omega_m}{\pi^2\omega_m^2 - (i\pi)^2}\sin(\pi\omega_m)(-1)^i.\end{aligned}$$

Furthermore,

$$\begin{aligned}\eta_{ij} &= \sum_m b_m\theta_{mi}\theta_{mj} = \frac{1}{\pi^2}\sum_m b_m \frac{\omega_m}{\omega_m^2 - i^2}\frac{\omega_m}{\omega_m^2 - j^2}\sin^2(\pi\omega_m)(-1)^{i+j} \\ &\overset{(*)}{\lesssim} (ij)^{-\min\left(1, \frac{\delta+1}{2}\right)},\end{aligned} \tag{4.1}$$

where $(*)$ is proved in Appendix B. This implies that $\tau_j \lesssim j^{-\min(\delta+1,2)}$ and $\sup_j |\tau_j| < \infty$. Hence, the inclusion $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})) \subset \tilde{\mathcal{H}} \subset \mathcal{H}$, follows from Proposition 4.1, where $\tilde{\mathcal{H}}$ consists of functions with a degree of smoothness that is by an amount $\min\left(1, \frac{1+\delta}{2}\right)$ higher than that of the functions in $\mathcal{H}$.

Remark 4.3. The covariance function of a standard Brownian motion is $c(x,y) = \min(x,y)$. It is well known that the eigenvalues and eigenfunctions are $b_m = \frac{1}{\pi^2(m-\frac{1}{2})^2}$ and $\psi_m(x) = \sqrt{2}\sin\left(\pi(m-\frac{1}{2})x\right)$. From the above example, it then follows that the space $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2}))$ consists of functions that are at least one degree smoother than the functions in $\mathcal{H}$.

Example (Haar wavelet basis). Let ψ_m be the Haar-wavelet basis function given by

$$\psi_{2^j+\ell-1}(x) = \begin{cases} +2^{\frac{j}{2}}, & \text{for } x \in \left[\frac{\ell-1}{2^j}, \frac{\ell-1/2}{2^j}\right] \\ -2^{\frac{j}{2}}, & \text{for } x \in \left[\frac{\ell-1/2}{2^j}, \frac{\ell}{2^j}\right] \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

In this case, $|\eta_{ij}| = |\sum_m b_m\theta_{mi}\theta_{mj}| \le \sum_m b_m|\theta_{mi}||\theta_{mj}|$ with

$$\theta_{mi} = \int_0^1 \psi_m(x)\cos(i\pi x)dx \le \frac{4}{i\pi}2^{\lfloor \log_2 m\rfloor/2},$$

which follows from Lemma A.8. Therefore,

$$|\eta_{ij}| \le \sum_m b_m \frac{16}{ij\pi^2}2^{\lfloor \log_2 m\rfloor} \le \frac{16}{ij\pi^2}\sum_m b_m 2^{\log_2 m} = \frac{16}{ij\pi^4}\sum_m b_m m.$$

Hence, we have

$$\eta_{ij}^2 \le \frac{256}{(ij)^2\pi^4}\left(\sum_m b_m m\right)^2$$

and

$$\tau_j = \sum_i a_i\eta_{ij}^2 = \frac{256}{j^2\pi^4}\left(\sum_m b_m m\right)^2 \sum_i \frac{a_i}{i^2} \lesssim \frac{1}{j^2}$$

for any choice of eigenvalues b_m with $\sum_m b_m m < \infty$. Therefore, $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})) \subset \tilde{\mathcal{H}} \subset \mathcal{H}$, where $\tilde{\mathcal{H}}$ is an RKHS with one degree of smoothness more than that of $\mathcal{H}$.

Remark 4.4. Cai and Yuan (2012) considered the spline kernel

$$k(x,y) = -\frac{1}{3}\left[B_4(x+y) + B_4(|x-y|)\right], \tag{4.3}$$

where B_4 is the fourth Bernoulli polynomial. In this case, it can be shown that $a_i = 2/(i\pi)^4$ and $\phi_i(x) = \cos(i\pi x)$, and $\mathcal{H}$ is an RKHS (in fact, a periodic Sobolev space) of twice differentiable functions which are square integrable on $[0,1]$. For this choice of a_i, $\tilde{\mathcal{H}}$ is the space of thrice differentiable functions which are square integrable on $[0,1]$ for the covariance functions considered in Remark 4.3 and Example 4.

5 Consequences for prediction error in functional linear models

In this section, we investigate the prediction error for the linear predictor $\langle X, \beta^* \rangle$, which is identical to $\langle X, \tilde{\beta}^* \rangle$ in the setting of a functional linear model where $\tilde{\beta}^* = \beta^*$. The prediction error in functional linear models was studied previously in Cai and Yuan (2012), which showed that a reasonable proxy is $\|C^{1/2}(\hat{\beta} - \beta^*)\|$, which they analyzed without invoking a commutativity requirement between T and C, but under the assumption that $\beta^* \in \mathcal{H}$. In the following, we generalize this result as follows. First, in Theorem 5.1 (proved in Section 7.6), we present a master theorem for the prediction error that does not rely on the assumption $\beta^* \in \mathcal{H}$. We specialize this result to non-commutative and commutative settings in Theorems 5.2 (proved in Section 7.7) and 5.3 (proved in Section 7.8), respectively, wherein the non-commutative setting recovers the result of Cai and Yuan (2012), while the commutative setting addresses the scenario of $\beta^* \in L^2(S) \backslash \mathcal{H}$.

Theorem 5.1 (Master theorem for prediction)**.** *Let $\|T^{-\alpha}\beta^*\| < \infty$ for $\alpha \in (0, 1/2]$, i.e., $\tilde{\beta}^* \in \mathscr{R}(T^\alpha)$. Let Θ, $d(\lambda)$ be as defined in (3.2), and $N(\lambda)$ be as defined in (3.3). Then for δ, n and λ satisfying the conditions in (3.4), with probability at least $1 - 3\delta$, we have*

$$\|C^{1/2}(\hat{\beta} - \beta^*)\|^2 \lesssim_p \frac{\sigma^2 N(\lambda)}{n\delta} + \lambda^2 \|T^{-\alpha}\beta^*\|^2 \frac{\|\Theta\| \text{trace}(\Theta)}{n} \\ + \|C^{1/2}T(CT + \lambda I)^{-1}C\beta^* - C^{1/2}\beta^*\|^2.$$

We now present a specialization of the above result when T and C are not commuting and $\alpha = 1/2$.

Theorem 5.2 (Prediction for noncommutative operators)**.** *Suppose $\beta^* \in \mathscr{R}(T^{1/2})$. Let $(\zeta_i)_{i \in \mathbb{N}}$ denote the eigenvalues of $T^{1/2}CT^{1/2}$ with $i^{-b} \lesssim \zeta_i \lesssim i^{-b}$, for some $b > 1$. Then, for*

$$\lambda = n^{-\frac{b}{1+b}}, \quad \text{we have} \quad \|C^{1/2}(\hat{\beta} - \beta^*)\| \lesssim_p n^{-\frac{b}{1+b}}. \tag{5.1}$$

Compared to Theorem 3.3, Theorem 5.2 requires only the weaker assumption that $\beta^* \in \mathscr{R}(T^{1/2})$ instead of $\beta^* \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})^\nu)$, for some $\nu \in (0, 1]$. This is because the prediction error is a weaker notion than the estimation error. The rate obtained in (5.1) was shown to be minimax optimal in Cai and Yuan (2012).

The following result is another specialization of Theorem 5.1, where β^* is relaxed to lie outside $\mathcal{H}$ but T and C are assumed to commute. Thus compared to Theorem 5.2, Theorem 5.3 considers the alternate setting with a weaker assumption on β^* and a stronger assumption on T and C.

Theorem 5.3 (Prediction for commutative operators)**.** *Let $\|T^{-\alpha}\beta^*\| < \infty$ for $\alpha \in (0, 1/2]$. Suppose the operators T and C commute and have simple eigenvalues (i.e., of multiplicity one) denoted by μ_i and ξ_i for $i \in \mathbb{N}$, such that for some $t > 1$ and $c > 1$, they satisfy the condition in (3.6). Then, by setting λ as in (3.7), we obtain*

$$\|C^{1/2}(\hat{\beta} - \beta^*)\| \lesssim_p n^{-\frac{2\alpha t + c}{1 + c + 2t(1-\alpha)}}. \tag{5.2}$$

The above result extends the results of Yuan and Cai (2010) to the case where β^* does not necessarily lie in the RKHS $\mathcal{H}$. When $\alpha = 1/2$, we recover the corresponding result from Yuan and Cai (2010), which matches with (5.1).

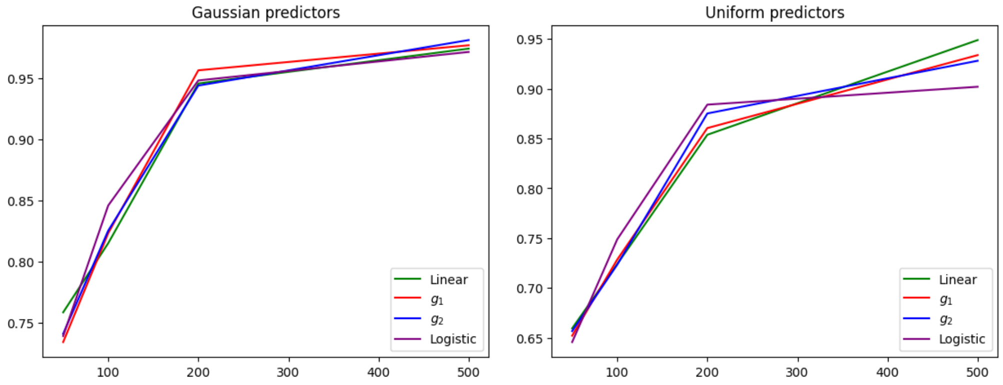

Figure 1: Cosine distance versus sample size: The figure on the left corresponds to the case of Gaussian predictors and figure on the right correspond to the case of uniform predictors.

6 Numerical simulations

In this section, using numerical simulations, we examine the robustness of the proposed method for non-Gaussian predictors while validating the presented theoretical results for Gaussian predictors. To this end, we let $\epsilon \sim N(0,1)$ in (1.1) and follow the setup in Hall and Horowitz (2007) and Cai and Yuan (2012) for $\beta^*(t)$ and $X(t)$, wherein $\beta^*(t) \coloneqq \sum_{j=1}^{50} 4(-1)^{(k+1)}k^{-2}\phi_k(t)$ with $\phi_1(t) \equiv 1$ and $\phi_{k+1}(t) = \sqrt{2}\cos(k\pi\, t)$, $t \in [0,1]$, for $k \geq 1$, and $X(t) = \sum_{k=1}^{50}(-1)^{(k+1)}k^{-2}Z_k\phi_k(t)$, with Z_k being one of the following:

- *Gaussian:* $Z_k \sim N(0,1)$ which leads to Gaussian process predictors, satisfying our assumptions.
- *Non-Gaussian:* $Z_k \sim \text{UNIF}[-3,3]$, which does not satisfy our assumptions.

Following Yang et al. (2017b), we consider four choices for the link function: *(i)* linear, *(ii)* $g_1(u) = g(u) = 3u + 10\sin(u)$, *(iii)* $g_2(u) = g(u) = \sqrt{2}u + 4\exp(-2u^2)$, and *(iv)* logistic, and following Cai and Yuan (2012), the kernel is chosen to be the one defined in (4.3). Using the above, our estimator is constructed as described in the paragraph below (2.4). In the construction of the kernel matrix, we select a grid size of 100 to approximate the integral over $S = [0,1]$. Since our main focus is to compare the estimated direction to the true direction without explicitly providing consideration to ϑ_{g,β^*}, we consider the cosine distance between the estimator and the truth, defined as $1 - \|\hat{\beta}\|^{-1}\|\beta^*\|^{-1}\langle\hat{\beta}(t), \beta^*(t)\rangle$ as a measure of the quality of estimation. Finally, since our main purpose is only to demonstrate the robustness of the proposed approach to any deviations from the assumptions, we set the tuning parameter manually to the best-performing one.

In Figure 1, we show the cosine distance averaged over 1000 simulations. From the results, the following two observations that support our methodology and theory can be made. First, note that despite the true model being not necessarily a linear model, the linear estimator succeeds in estimating the direction. Second, although our methodology and theoretical results are derived under the Gaussian process assumption, the proposed approach works equally well with non-Gaussian predictors. Goldstein and Wei (2019) used a non-Gaussian version of finite-dimensional Stein's identity to explain this observation for the Euclidean setting. In the infinite-dimensional setting, non-Gaussian Stein's identities are not well explored. Deriving such results and providing theoretical support for this empirical observation are left as intriguing future work.

7 Proofs

In this section, we provide the proof of the results of Sections 3-5.

7.1 Proof of Theorem 3.1

The proof proceeds by first decomposing $\|\hat{\beta}-\tilde{\beta}^*\|$ into several terms which are subsequently upper-bounded individually. Recall $\hat{C} := \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i$ and $\hat{R} := \frac{1}{n}\sum_{i=1}^n Y_i X_i$. Define

$$A := \mathfrak{I}^* C \mathfrak{I},\ \ \hat{A} := \mathfrak{I}^* \hat{C} \mathfrak{I},\ \ B = \mathfrak{I}^*\mathfrak{I},\ \text{and}$$

$$\beta_\lambda = [A+\lambda I]^{-1}\,\mathfrak{I}^*\mathbb{E}[YX] = [A+\lambda I]^{-1}\,\mathfrak{I}^* C\tilde{\beta}^*, \tag{7.1}$$

where the last equality in β_λ follows from Stein's identity. By the definition of β_λ in (7.1), we have

$$\begin{aligned}
\hat{\beta}-\beta_\lambda &= (\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-(\hat{A}+\lambda I)\beta_\lambda\right] = (\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\hat{A}\beta_\lambda-\lambda\beta_\lambda\right]\\
&= (\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\hat{A}\beta_\lambda+A\beta_\lambda-\mathfrak{I}^*C\tilde{\beta}^*\right]\\
&= (\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\mathfrak{I}\beta_\lambda+\mathfrak{I}^*C\mathfrak{I}\beta_\lambda-\mathfrak{I}^*C\tilde{\beta}^*\right]\\
&= (\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*C\mathfrak{I}\beta_\lambda-\mathfrak{I}^*C\tilde{\beta}^*-\mathfrak{I}^*\hat{C}\mathfrak{I}\beta_\lambda\right]\\
&= (\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right].
\end{aligned}$$

Based on the above identity, we then have

$$\begin{aligned}
\|\hat{\beta}-\tilde{\beta}^*\| &= \|\mathfrak{I}\hat{\beta}-\tilde{\beta}^*\| = \left\|\mathfrak{I}(\hat{\beta}-\beta_\lambda)+\mathfrak{I}\beta_\lambda-\tilde{\beta}^*\right\|\\
&\le \|\mathfrak{I}(\hat{\beta}-\beta_\lambda)\|+\|\mathfrak{I}\beta_\lambda-\tilde{\beta}^*\| = \|B^{1/2}(\hat{\beta}-\beta_\lambda)\|_{\mathcal{H}}+\|\mathfrak{I}\beta_\lambda-\tilde{\beta}^*\|\\
&= \left\|B^{1/2}(\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right]\right\|_{\mathcal{H}}+\|\mathfrak{I}\beta_\lambda-\tilde{\beta}^*\|.
\end{aligned}$$

Since

$$\begin{aligned}
&B^{1/2}(\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right]\\
&= B^{1/2}(A+\lambda I)^{-1/2}(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1/2}(\hat{A}+\lambda I)^{-1/2}(A+\lambda I)^{1/2}\\
&\qquad\cdot(A+\lambda I)^{-1/2}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right]
\end{aligned}$$

we obtain

$$\begin{aligned}
&\left\|B^{1/2}(\hat{A}+\lambda I)^{-1}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right]\right\|_{\mathcal{H}}\\
&\le \|B^{1/2}(A+\lambda I)^{-1/2}\|\cdot\|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1/2}\|\\
&\qquad\cdot\|(\hat{A}+\lambda I)^{-1/2}(A+\lambda I)^{1/2}\|\\
&\qquad\qquad\cdot\left\|(A+\lambda I)^{-1/2}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*+\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right]\right\|_{\mathcal{H}},
\end{aligned}$$

therefore resulting in

$$\|\hat{\beta}-\tilde{\beta}^*\| \le \underbrace{\|B^{1/2}(A+\lambda I)^{-1/2}\|}_{\texttt{Term 1}}\cdot\underbrace{\|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1/2}\|}_{\texttt{Term 2}}$$

$$\cdot \underbrace{\|(\hat{A}+\lambda I)^{-1/2}(A+\lambda I)^{1/2}\|}_{\texttt{Term 3}} \cdot \Bigg[\underbrace{\Big\|(A+\lambda I)^{-1/2}\Big[\mathfrak{J}^*\hat{R} - \mathfrak{J}^*\hat{C}\tilde{\beta}^*\Big]\Big\|_{\mathcal{H}}}_{\texttt{Term 4}}$$

$$+ \underbrace{\Big\|(A+\lambda I)^{-1/2}\mathfrak{J}^*(C-\hat{C})(\mathfrak{J}\beta_\lambda - \tilde{\beta}^*)\Big\|_{\mathcal{H}}}_{\texttt{Term 5}}\Bigg] + \underbrace{\|\mathfrak{J}\beta_\lambda - \tilde{\beta}^*\|}_{\texttt{Term 6}}. \tag{7.2}$$

Bounding `Term 1`

$$\begin{aligned}
\|B^{1/2}(A+\lambda I)^{-1/2}\|^2 &= \|(A+\lambda I)^{-1/2}B(A+\lambda I)^{-1/2}\| = \|B^{1/2}(A+\lambda I)^{-1}B^{1/2}\| \\
&= \|(A+\lambda I)^{-1/2}B^{1/2}\|^2 \le \|(A+\lambda I)^{-1}B\| = \|(\mathfrak{J}^*C\mathfrak{J}+\lambda I)^{-1}\mathfrak{J}^*\mathfrak{J}\| \\
&= \|\mathfrak{J}^*(CT+\lambda)^{-1}\mathfrak{J}\| = \|\mathfrak{J}^*T^{-1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}\mathfrak{J}\| \\
&\overset{(*)}{=} \|\mathfrak{J}^*T^{1/2}(\Lambda+\lambda I)^{-1}T^{-1/2}\mathfrak{J}\mathfrak{J}^*T^{-1/2}(\Lambda+\lambda I)^{-1}T^{1/2}\mathfrak{J}\|^{1/2} \\
&= \|\mathfrak{J}^*T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-2}T^{1/2}\mathfrak{J}\|^{1/2} \\
&= \|(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}TT^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}\|^{1/2} \\
&= \|(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^2(T^{1/2}CT^{1/2}+\lambda I)^{-1}\|^{1/2} \\
&= \|T(T^{1/2}CT^{1/2}+\lambda I)^{-2}T\|^{1/2},
\end{aligned}$$

where $\Lambda := T^{1/2}CT^{1/2}$ in $(*)$ and the only step with the inequality in the above sequence of calculations, follows from Cordes' inequality (Cordes, 1987). Hence, we have

$$\|B^{1/2}(A+\lambda I)^{-1/2}\| \le \|T(T^{1/2}CT^{1/2}+\lambda I)^{-2}T\|^{1/4}. \tag{7.3}$$

Bounding `Term 2` and `Term 3`:

First note that

$$\begin{aligned}
\|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1/2}\|^2 &= \|(\hat{A}+\lambda I)^{-1/2}(A+\lambda I)(\hat{A}+\lambda I)^{-1/2}\| \\
&= \|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1}(A+\lambda)^{1/2}\| = \|(\hat{A}+\lambda I)^{-1/2}(A+\lambda I)^{1/2}\|^2,
\end{aligned}$$

which implies that `Term 2` = `Term 3`. Next, note that

$$\begin{aligned}
&\|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1/2}\|^2 = \|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1}(A+\lambda)^{1/2}\| \\
&= \left\|\Big[I-(A+\lambda I)^{-1/2}(A-\hat{A})(A+\lambda I)^{-1/2}\Big]^{-1}\right\| \le \frac{1}{1-\Big\|(A+\lambda I)^{-1/2}(A-\hat{A})(A+\lambda I)^{-1/2}\Big\|}.
\end{aligned}$$

Define

$$\begin{aligned}
\Sigma &:= (A+\lambda I)^{-1/2}A(A+\lambda I)^{-1/2} = (A+\lambda I)^{-1/2}\mathfrak{J}^*C\mathfrak{J}(A+\lambda I)^{-1/2} \\
&= \mathbb{E}\Big[(A+\lambda I)^{-1/2}\mathfrak{J}^*(X\otimes X)\mathfrak{J}(A+\lambda I)^{-1/2}\Big],
\end{aligned}$$

and

$$\hat{\Sigma} := (A+\lambda I)^{-1/2}\hat{A}(A+\lambda I)^{-1/2} = (A+\lambda I)^{-1/2}\mathfrak{J}^*\hat{C}\mathfrak{J}(A+\lambda I)^{-1/2}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(A+\lambda I)^{-1/2}\mathfrak{I}^*(X_i \otimes X_i)\mathfrak{I}(A+\lambda I)^{-1/2}.$$

This yields $\|(A+\lambda I)^{-1/2}(A-\hat{A})(A+\lambda I)^{-1/2}\| = \|\hat{\Sigma}-\Sigma\|$. Therefore by Theorem A.3, for any

$$n \geq (r(\Sigma) \vee \tau) \quad \text{and} \quad \tau \geq 1, \tag{7.4}$$

with probability at least $1-e^{-\tau}$, we have

$$\|\hat{\Sigma}-\Sigma\| \;\leq\; K_1\|\Sigma\| \,\frac{\sqrt{r(\Sigma)}+\sqrt{\tau}}{\sqrt{n}} \leq K_1\frac{\sqrt{r(\Sigma)}+\sqrt{\tau}}{\sqrt{n}}, \tag{7.5}$$

where K_1 is a universal constant and we used $\|\Sigma\| = \|(A+\lambda I)^{-1/2}A(A+\lambda I)^{-1/2}\| \leq 1$. The effective rank $r(\Sigma)$ satisfies

$$r(\Sigma) \leq \frac{\text{trace}(\Sigma)}{\|\Sigma\|} = \frac{\text{trace}((A+\lambda I)^{-1/2}A(A+\lambda I)^{-1/2})}{\|(A+\lambda I)^{-1/2}A(A+\lambda I)^{-1/2}\|} = \frac{\text{trace}((A+\lambda I)^{-1}A)}{\sup_i\left[\frac{\lambda_i(A)}{\lambda+\lambda_i(A)}\right]},$$

with $\lambda_i(A)$ denoting the i^{th} eigenvalue of operator A. Observe that

$$\text{trace}((A+\lambda I)^{-1}A) \leq \|(A+\lambda I)^{-1}\|\text{trace}(A) \leq \frac{\text{trace}(A)}{\lambda}.$$

Furthermore,

$$\sup_i\left[\frac{\lambda_i(A)}{\lambda+\lambda_i(A)}\right] \geq \frac{\sup_i \lambda_i(A)}{\lambda+\|A\|} = \frac{\|A\|}{\lambda+\|A\|}.$$

Hence,

$$r(\Sigma) \leq \left(\frac{\lambda+\|A\|}{\|A\|}\right)\frac{\text{trace}(A)}{\lambda} \leq \left(1+\frac{\lambda}{\|A\|}\right)\frac{\text{trace}(A)}{\lambda} = \left(\frac{1}{\lambda}+\frac{1}{\|A\|}\right)\text{trace}(A) \leq \frac{2\text{trace}(A)}{\lambda}, \tag{7.6}$$

where the last inequality holds since $\lambda \leq \|A\|$. Using (7.6) in (7.5), we obtain that with probability at least $1-e^{-\tau}$,

$$\|\hat{\Sigma}-\Sigma\| \leq K_1\left(\sqrt{\frac{2\text{trace}(A)}{\lambda n}}+\sqrt{\frac{\tau}{n}}\right) \leq \frac{1}{2}$$

if

$$\lambda \geq \frac{32K_1^2\text{trace}(A)}{n} \quad \text{and} \quad n \geq 16\tau K_1^2. \tag{7.7}$$

Combining (7.4) and (7.7), we see that we require $\tau \geq 1$, $n \geq \tau$ and $n \geq r(\Sigma)$, which are satisfied as long as $n \geq 2\text{trace}(A)/\lambda$ or equivalently $\lambda \geq 2\text{trace}(A)/n$. Finally, note that we have

$$\text{trace}(A) = \text{trace}(\mathfrak{I}^*C\mathfrak{I}) = \text{trace}(CT) = \text{trace}(T^{1/2}CT^{1/2}),$$

and

$$\|A\| = \|\mathfrak{I}^*C\mathfrak{I}\| = \|C^{1/2}TC^{1/2}\| = \|T^{1/2}CT^{1/2}\|.$$

Hence, as long as

$$\begin{aligned} &(2\vee 32K_1^2)\frac{\operatorname{trace}(T^{1/2}CT^{1/2})}{n} \le \lambda \le \|T^{1/2}CT^{1/2}\|,\\ &\delta \le \frac{1}{e} \quad \text{and} \quad n \ge (1\vee 16K_1^2)\ln\left(\frac{1}{\delta}\right), \end{aligned} \tag{7.8}$$

we have with probability at least $1-\delta$, $\|\hat{\Sigma}-\Sigma\|\le 1/2$. Hence under the conditions in (7.8),

$$\|(A+\lambda I)^{1/2}(\hat{A}+\lambda I)^{-1/2}\| \le \sqrt{\frac{1}{1-\|\hat{\Sigma}-\Sigma\|}} \le \sqrt{2}. \tag{7.9}$$

Bounding Term 4

Define $Z_i := (A+\lambda I)^{-1/2}\left[\mathfrak{I}^* Y_i X_i - \mathfrak{I}^*(X_i\otimes X_i)\tilde{\beta}^*\right]$ so that

$$\mathbb{E}[Z_i] = (A+\lambda I)^{-1/2}\left[\mathfrak{I}^*\mathbb{E}[YX] - \mathfrak{I}^* C\tilde{\beta}^*\right] = 0.$$

By Chebyshev's inequality for Hilbert-valued random variables (see Lemma A.2), with probability at least $1-\delta$, we have

$$\left\|(A+\lambda I)^{-1/2}\left[\mathfrak{I}^*\hat{R} - \mathfrak{I}^*\hat{C}\tilde{\beta}^*\right]\right\|_{\mathcal{H}} \le \sqrt{\frac{\mathbb{E}\left[\left\|(A+\lambda I)^{-1/2}\left[\mathfrak{I}^* YX - \mathfrak{I}^*(X\otimes X)\tilde{\beta}^*\right]\right\|_{\mathcal{H}}^2\right]}{n\delta}}, \tag{7.10}$$

where

$$\begin{aligned} &\mathbb{E}\left[\left\|(A+\lambda I)^{-1/2}\left[\mathfrak{I}^* YX - \mathfrak{I}^*(X\otimes X)\tilde{\beta}^*\right]\right\|_{\mathcal{H}}^2\right] = \mathbb{E}\left[\left\|(A+\lambda I)^{-1/2}\left[\mathfrak{I}^*(Y-\langle X,\tilde{\beta}^*\rangle)X\right]\right\|_{\mathcal{H}}^2\right]\\ &= \mathbb{E}\left[(Y-\langle X,\tilde{\beta}^*\rangle)^2\left\|(A+\lambda I)^{-1/2}\mathfrak{I}^* X\right\|_{\mathcal{H}}^2\right]\\ &= \mathbb{E}\left[(Y-\langle X,\tilde{\beta}^*\rangle)^2\left\langle (A+\lambda I)^{-1/2}\mathfrak{I}^* X, (A+\lambda I)^{-1/2}\mathfrak{I}^* X\right\rangle_{\mathcal{H}}\right]\\ &= \mathbb{E}\left[(Y-\langle X,\tilde{\beta}^*\rangle)^2 \operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]\\ &= \mathbb{E}\left[\left(Y-g(\langle X,\tilde{\beta}^*\rangle)+g(\langle X,\tilde{\beta}^*\rangle)-\langle X,\tilde{\beta}^*\rangle\right)^2 \operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]\\ &\le 2\mathbb{E}\left[\left\{\epsilon^2 + (g(\langle X,\tilde{\beta}^*\rangle)-\langle X,\tilde{\beta}^*\rangle)^2\right\}\operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]\\ &= 2\mathbb{E}[\epsilon^2]\,\mathbb{E}\left[\operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]\\ &\qquad\qquad + 2\mathbb{E}\left[(g(\langle X,\tilde{\beta}^*\rangle)-\langle X,\tilde{\beta}^*\rangle)^2\operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]\\ &= 2\mathbb{E}[\epsilon^2]\operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^* C\mathfrak{I}\right)\\ &\qquad\qquad + 2\mathbb{E}\left[(g(\langle X,\tilde{\beta}^*\rangle)-\langle X,\tilde{\beta}^*\rangle)^2\operatorname{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]\\ &\le 2\sigma^2\operatorname{trace}\left((A+\lambda I)^{-1}A\right) + 2\sqrt{\varkappa}\sqrt{\mathbb{E}\left[\operatorname{trace}^2\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right]}, \end{aligned} \tag{7.11}$$

where we recall that $\varkappa$ is defined in (3.1). Recalling the definition of $N(\lambda)$ from (3.3), we have

$$\operatorname{trace}\left((A+\lambda I)^{-1}A\right) = \operatorname{trace}\left((\mathfrak{I}^* C\mathfrak{I}+\lambda I)^{-1}\mathfrak{I}^* C\mathfrak{I}\right) = \operatorname{trace}\left(\mathfrak{I}^*(C\mathfrak{I}\mathfrak{I}^*+\lambda I)^{-1}C\mathfrak{I}\right)$$

$$
\begin{aligned}
&= \mathrm{trace}\left(T(CT+\lambda I)^{-1}C\right) = \mathrm{trace}\left(CT(CT+\lambda I)^{-1}\right)\\
&= \mathrm{trace}\left(CT^{1/2}(CT^{1/2}+\lambda T^{-1/2})^{-1}\right)\\
&= \mathrm{trace}\left(T^{1/2}CT^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}\right) = N(\lambda).
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
&\mathrm{trace}\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right) = \mathrm{trace}\left(\mathfrak{I}(A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\right)\\
&= \mathrm{trace}\left(\mathfrak{I}(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\right) = \mathrm{trace}\left(\mathfrak{I}\mathfrak{I}^*(C\mathfrak{I}\mathfrak{I}^*+\lambda I)^{-1}(X\otimes X)\right)\\
&= \mathrm{trace}\left(T(CT+\lambda I)^{-1}(X\otimes X)\right) = \mathrm{trace}\left(T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}(X\otimes X)\right)\\
&= \left\langle X, T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}X\right\rangle.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{trace}^2\left((A+\lambda I)^{-1}\mathfrak{I}^*(X\otimes X)\mathfrak{I}\right)\right] = \mathbb{E}\left[\left\langle X, T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}X\right\rangle^2\right]\\
&\overset{(*)}{\le} 3\mathrm{trace}^2\left(T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}C\right) \le 4N^2(\lambda), \qquad (7.12)
\end{aligned}
$$

where $(*)$ follows from Lemma A.4. Combining (7.11) and (7.12) in (7.10), we obtain with probability at least $1-\delta$,

$$
\left\|(A+\lambda I)^{-1/2}\left[\mathfrak{I}^*\hat{R}-\mathfrak{I}^*\hat{C}\tilde{\beta}^*\right]\right\|_{\mathcal{H}} \le \sqrt{\frac{(2\sigma^2+4\sqrt{\varkappa})N(\lambda)}{n\delta}}, \qquad (7.13)
$$

under the assumption that $\mathrm{trace}(C^{1/2})<\infty$, as required by Lemma A.4.

Bounding `Term 5`

Observe that

$$
\begin{aligned}
&\left\|(A+\lambda I)^{-1/2}\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\right\|_{\mathcal{H}}\\
&= \|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{-1/2}\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\|_{\mathcal{H}}\\
&= \|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{1/2}(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{-1}\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\|_{\mathcal{H}}\\
&= \|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{1/2}\mathfrak{I}^*(CT+\lambda)^{-1}(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*))\|_{\mathcal{H}}\\
&= \left\|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{1/2}\mathfrak{I}^*(CT+\lambda I)^{-1}(C-\hat{C})\left(\mathfrak{I}(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{-1}\mathfrak{I}^*C\tilde{\beta}^*-\tilde{\beta}^*\right)\right\|_{\mathcal{H}}\\
&= \left\|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{1/2}\mathfrak{I}^*(CT+\lambda I)^{-1}(C-\hat{C})\left((TC+\lambda I)^{-1}TC\tilde{\beta}^*-\tilde{\beta}^*\right)\right\|_{\mathcal{H}}\\
&= \left\|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{1/2}\mathfrak{I}^*(CT+\lambda)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}\left(TC\tilde{\beta}^*-(TC+\lambda I)\tilde{\beta}^*\right)\right\|_{\mathcal{H}}\\
&= \lambda\left\|(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{1/2}\mathfrak{I}^*(CT+\lambda I)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}\tilde{\beta}^*\right\|_{\mathcal{H}}\\
&\le \lambda\|\mathfrak{I}^*C\mathfrak{I}+\lambda I\|^{1/2}\left\|\mathfrak{I}^*(CT+\lambda I)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}\tilde{\beta}^*\right\|\\
&\le \lambda\left(\left\|T^{1/2}CT^{1/2}\right\|^{1/2}+\sqrt{\lambda}\right)\left\|T^{1/2}(CT+\lambda)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}\tilde{\beta}^*\right\|
\end{aligned}
$$

$$= \lambda \left\| T^{1/2-\alpha} T^{\alpha}(CT+\lambda I)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}T^{\alpha}T^{-\alpha}\tilde{\beta}^*\right\| \left(\left\|T^{1/2}CT^{1/2}\right\|^{1/2}+\sqrt{\lambda}\right)$$
$$\le \lambda\|T\|^{1/2-\alpha}\|T^{\alpha}(CT+\lambda I)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}T^{\alpha}\|\|T^{-\alpha}\tilde{\beta}^*\| \left(\left\|T^{1/2}CT^{1/2}\right\|^{1/2}+\sqrt{\lambda}\right),$$

where $0<\alpha\le\frac{1}{2}$. Now, recalling the definition of Θ from (3.2), and defining

$$\hat{\Theta} := \frac{1}{n}\sum_{i=1}^{n} T^{\alpha}(CT+\lambda I)^{-1}(X_i\otimes X_i)(TC+\lambda I)^{-1}T^{\alpha},$$

we immediately have $\|T^{\alpha}(CT+\lambda I)^{-1}(C-\hat{C})(TC+\lambda I)^{-1}T^{\alpha}\| = \|\hat{\Theta}-\Theta\|$. Hence, by Theorem A.3, we have with probability at least $1-e^{-\tau}$,

$$\|\hat{\Theta}-\Theta\| \le K_2\|\Theta\|\frac{\sqrt{r(\Theta)}+\sqrt{\tau}}{\sqrt{n}},$$

for $\tau\ge 1$ and $n\ge(r(\Theta)\vee\tau)$, where K_2 is a universal constant. In other words, recalling the definition of $d(\lambda)$ from (3.2), for $\delta\le 1/e$, $n\ge d(\lambda)$, and $d(\lambda)\ge\ln(1/\delta)$, with probability at least $1-\delta$, we have

$$\begin{aligned}&\|(A+\lambda I)^{-1/2}\mathfrak{I}^*(C-\hat{C})(\mathfrak{I}\beta_\lambda-\tilde{\beta}^*)\|_{\mathcal{H}}\\ &\le K_3\left(\left\|T^{1/2}CT^{1/2}\right\|^{1/2}+\sqrt{\lambda}\right)\|T\|^{1/2-\alpha}\|T^{-\alpha}\tilde{\beta}^*\|\lambda\|\Theta\|\sqrt{\frac{d(\lambda)}{n}},\end{aligned} \tag{7.14}$$

for some universal constant K_3.

Bounding `Term 6`

Note that

$$\|\mathfrak{I}\beta_\lambda-\tilde{\beta}^*\| = \|(\mathfrak{I}(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{-1}\mathfrak{I}^*C\tilde{\beta}^*-\tilde{\beta}^*\| = \|T(CT+\lambda I)^{-1}C\tilde{\beta}^*-\tilde{\beta}^*\|. \tag{7.15}$$

The claim in Theorem 3.1, immediately follows by combining (7.2), (7.3), (7.9), (7.13), (7.14) and (7.15).

7.2 Proof of Theorem 3.2

The proof of Theorem 3.2 follows by carefully obtaining bounds on the individual terms in the inequality (3.5) of the master theorem (Theorem 3.1). Since T and C commute and have simple eigenvalues, they have the same eigenfunctions. Hence, recalling the definition of Ξ in (3.3), we have

$$\begin{aligned}\|\Xi\| &= \left\|T(T^{1/2}CT^{1/2}+\lambda I)^{-2}T\right\|\\ &= \sup_i\frac{\mu_i^2}{(\mu_i\xi_i+\lambda)^2} \lesssim \sup_i\frac{i^{-2t}}{(i^{-(t+c)}+\lambda)^2} = \left[\sup_i\frac{i^{-t}}{i^{-(t+c)}+\lambda)}\right]^2\\ &\le\left[\lambda^{\frac{t-(t+c)}{t+c}}\right]^2 = \lambda^{-\frac{-2c}{t+c}},\end{aligned} \tag{7.16}$$

where the last inequality follows from Lemma A.6. Next, recalling the definition of $N(\lambda)$ from (3.3) we have

$$\begin{aligned} N(\lambda) &= \operatorname{trace}\left[(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^{1/2}\right] \\ &= \sum_i \frac{\mu_i\xi_i}{\mu_i\xi_i+\lambda} \lesssim \sum_i \frac{i^{-(t+c)}}{i^{-(t+c)}+\lambda} \lesssim \lambda^{-\frac{1}{t+c}}, \end{aligned} \tag{7.17}$$

where the last inequality follows from Lemma A.5. Next, recalling the definition of Θ in (3.2), we have

$$\begin{aligned} \operatorname{trace}(\Theta) &= \operatorname{trace}\left(T^{\alpha}(CT+\lambda I)^{-1}C(TC+\lambda I)^{-1}T^{\alpha}\right) \\ &= \sum_i \frac{\mu_i^{2\alpha}\xi_i}{(\mu_i\xi_i+\lambda)^2} \lesssim \sum_i \frac{i^{-(2\alpha t+c)}}{(i^{-(t+c)}+\lambda)^2} \\ &\lesssim \lambda^{-\frac{1+(t+c)2-(2\alpha t+c)}{t+c}} = \lambda^{-\frac{1+c+2t(1-\alpha)}{t+c}}, \end{aligned}$$

where the last inequality follows from Lemma A.5. Next, we upper bound $\|\Theta\|$ as

$$\begin{aligned} \|\Theta\| &= \sup_i \frac{\mu_i^{2\alpha}\ \xi_i}{(\mu_i\xi_i+\lambda)^2} \lesssim \sup_i \frac{i^{-(2\alpha t+c)}}{(i^{-(t+c)}+\lambda)^2} = \left[\sup_i \frac{i^{-(\alpha t+c/2)}}{i^{-(t+c)}+\lambda}\right]^2 \\ &\leq \left[\lambda^{\frac{\alpha t+c/2-(t+c)}{t+c}}\right]^2 = \lambda^{\frac{2(\alpha-1)t-c}{t+c}}, \end{aligned}$$

where the last inequality follows from Lemma A.6. We also bound $\|\Theta\|$ from below as

$$\|\Theta\| = \sup_i \frac{\mu_i^{2\alpha}\xi_i}{(\mu_i\xi_i+\lambda)^2} \geq \sup_i \frac{\mu_i^{2\alpha}\xi_i}{(\|T^{1/2}CT^{1/2}\|+\lambda)^2} = \frac{\|T^{\alpha}CT^{\alpha}\|}{(\|T^{1/2}CT^{1/2}\|+\lambda)^2}.$$

Hence, we have

$$\|\Theta\| \lesssim \lambda^{-\frac{2(1-\alpha)t+c}{t+c}} \quad \text{and} \quad \operatorname{trace}(\Theta) \lesssim \lambda^{-\frac{1+c+2t(1-\alpha)}{t+c}}, \tag{7.18}$$

and (recalling the definition of $d(\lambda)$ from (3.2))

$$d(\lambda) \leq \frac{\operatorname{trace}(\Theta)}{\|T^{\alpha}CT^{\alpha}\|}(\|T^{1/2}CT^{1/2}\|+\lambda)^2 \lesssim 4\|T^{1/2}CT^{1/2}\|^2\ \lambda^{-\frac{1+c+2t(1-\alpha)}{t+c}}.$$

Hence, the condition $n \gtrsim (d(\lambda) \vee \ln(1/\delta))$ is satisfied if $n \gtrsim \ln(1/\delta)$ and $n \gtrsim \lambda^{-\frac{1+c+2t(1-\alpha)}{t+c}}$ or equivalently $\lambda \gtrsim n^{-\frac{t+c}{1+c+2t(1-\alpha)}}$. Hence, the conditions on n and λ read as

$$n \gtrsim \ln(1/\delta) \quad \text{and} \quad n^{-\frac{t+c}{1+c+2t(1-\alpha)}} \lesssim \lambda \leq \|T^{1/2}CT^{1/2}\|. \tag{7.19}$$

We now handle the bias term $\textsc{bias}(\lambda)$. Denote by $(\psi_i)_{i\in\mathbb{N}}$, the eigenfunctions of the operator T. Recall that the assumption $\tilde{\beta}^* \in \mathscr{R}(T^{\alpha})$, for $\alpha \in (0,1/2]$ implies that $\exists h \in L^2(S)$ such that $T^{\alpha}h = \tilde{\beta}^*$. Using this, we obtain

$$\|T(CT+\lambda I)^{-1}C\tilde{\beta}^* - \tilde{\beta}^*\| = \|T(CT+\lambda I)^{-1}CT^{\alpha}h - T^{\alpha}h\| \tag{7.20}$$

$$
\begin{aligned}
&= \left[\sum_i \left(\frac{\mu_i^{1+\alpha}\xi_i}{\mu_i\xi_i+\lambda} - \mu_i^\alpha\right)^2 \langle \psi_i, h\rangle^2\right]^{1/2} = \left[\sum_i \left(\frac{\lambda\mu_i^{\alpha}}{\mu_i\xi_i+\lambda}\right)^2 \langle \psi_i, h\rangle^2\right]^{1/2} \\
&\leq \lambda \sup_i \left[\frac{\mu_i^\alpha}{\mu_i\xi_i+\lambda}\right] \|h\| \lesssim \lambda \sup_i \left[\frac{i^{-\alpha t}}{i^{-(t+c)}+\lambda}\right] \|T^{-\alpha}\tilde{\beta}^*\| \\
&\lesssim \lambda\lambda^{\frac{\alpha t - t - c}{t+c}} \|T^{-\alpha}\tilde{\beta}^*\| = \lambda^{\frac{\alpha t}{t+c}} \|T^{-\alpha}\tilde{\beta}^*\|,
\end{aligned}
$$

where the last inequality follows from Lemma A.6. Combining (7.16)–(7.20), we obtain

$$
\begin{aligned}
&\|\hat{\beta} - \tilde{\beta}^*\| \\
&\lesssim \lambda^{-\frac{2c}{4(t+c)}} \left[\frac{\lambda^{-\frac{1}{2(t+c)}}}{\sqrt{n}} + \frac{\lambda(1+\sqrt{\lambda})\lambda^{-\frac{2(1-\alpha)t+c}{2(t+c)}}\lambda^{-\frac{1+c+2t(1-\alpha)}{2(t+c)}}}{\sqrt{n}}\right] + \lambda^{\frac{\alpha t}{t+c}} \\
&\lesssim \lambda^{-\frac{2c}{4(t+c)}} \left[\underbrace{\frac{\lambda^{-\frac{1}{2(t+c)}}}{\sqrt{n}}}_{p} + \underbrace{\frac{\lambda^{-\frac{1+2t(1-2\alpha)}{2(t+c)}}}{\sqrt{n}}}_{q}\right] + \lambda^{\frac{\alpha t}{t+c}},
\end{aligned}
$$

as $\sqrt{\lambda} = o(1)$ as $\lambda \to 0$. Also $p = o(q)$ as $\lambda \to 0$. Therefore, we obtain

$$
\|\hat{\beta} - \tilde{\beta}^*\| \lesssim \frac{\lambda^{-\frac{c}{2(t+c)}}\lambda^{-\frac{1+2t(1-2\alpha)}{2(t+c)}}}{\sqrt{n}} + \lambda^{\frac{\alpha t}{t+c}} = \frac{\lambda^{-\frac{1+c+2t(1-2\alpha)}{2(t+c)}}}{\sqrt{n}} + \lambda^{\frac{\alpha t}{t+c}}.
$$

Hence, by picking λ as in (3.7) (which satisfies the condition on λ in (7.19)), we obtain the claim in (3.7).

7.3 Proof of Theorem 3.3

We now prove Theorem 3.3 by carefully upper bounding the terms in Theorem 3.1 under the assumptions of Theorem 3.3. By recalling the definition of Ξ from (3.3), we have

$$
\|\Xi\|^{1/4} \leq \|T\|^{1/2} \left\|(T^{1/2}CT^{1/2}+\lambda I)^{-1}\right\|^{1/2} \lesssim \frac{1}{\sqrt{\lambda}}. \tag{7.21}
$$

Next, recalling the definition of $N(\lambda)$ from (3.3), we have

$$
\begin{aligned}
N(\lambda) &= \operatorname{trace}\left[(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^{1/2}\right] \\
&= \sum_i \frac{\zeta_i}{\zeta_i+\lambda} \lesssim \sum_i \frac{i^{-b}}{i^{-b}+\lambda} \lesssim \lambda^{-\frac{1}{b}}, \qquad (7.22)
\end{aligned}
$$

where the last inequality follows from Lemma A.5. Since

$$
\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}(T^{1/2}CT^{1/2})^\nu) \subset \mathscr{R}\,(T^{1/2}),
$$

we have $\alpha = \frac{1}{2}$. Therefore, it follows from (3.2) that

$$
\operatorname{trace}(\Theta) = \operatorname{trace}\left(T^{1/2}(CT+\lambda I)^{-1}C(TC+\lambda I)^{-1}T^{1/2}\right)
$$

$$\begin{aligned}
&\overset{(*)}{=} \operatorname{trace}\Big(T^{1/2}T^{-1/2}(\Lambda+\lambda I)^{-1}\Lambda(\Lambda+\lambda I)^{-1}T^{-1/2}T^{1/2}\Big)\\
&= \operatorname{trace}\Big(T^{1/2}CT^{1/2}(T^{1/2}CT^{1/2}+\lambda)^{-2}\Big)\\
&= \sum_i \frac{\zeta_i}{(\zeta_i+\lambda)^2} \lesssim \sum_i \frac{i^{-b}}{(i^{-b}+\lambda)^2} \lesssim \lambda^{-\frac{1+b}{b}}, \qquad (7.23)
\end{aligned}$$

where the last inequality follows from Lemma A.5 and $\Lambda := T^{1/2}CT^{1/2}$ in $(*)$. Furthermore, we have the following upper bound on $\|\Theta\|$ as

$$\begin{aligned}
\|\Theta\| &= \Big\|T^{1/2}(CT+\lambda I)^{-1}C(TC+\lambda I)^{-1}T^{1/2}\Big\|\\
&= \Big\|(T^{1/2}CT^{1/2}+\lambda I)^{-1}(T^{1/2}CT^{1/2})(T^{1/2}CT^{1/2}+\lambda I)^{-1}\Big\|\\
&= \sup_i \frac{\zeta_i}{(\zeta_i+\lambda)^2} \lesssim \sup_i \frac{i^{-b}}{(i^{-b}+\lambda)^2} = \left[\sup_i \frac{i^{-b/2}}{i^{-b}+\lambda}\right]^2 \lesssim \frac{1}{\lambda}, \qquad (7.24)
\end{aligned}$$

where the last inequality follows from Lemma A.6. We also have the following lower bound on $\|\Theta\|$ as

$$\|\Theta\| = \sup_i \frac{\zeta_i}{(\zeta_i+\lambda)^2} \geq \frac{\|T^{1/2}CT^{1/2}\|}{(\|T^{1/2}CT^{1/2}\|+\lambda)^2}. \qquad (7.25)$$

Hence, recalling the definition of $d(\lambda)$ from (3.2) and the fact that $\lambda \leq \|T^{1/2}CT^{1/2}\|$, we have

$$d(\lambda) \leq \frac{\lambda^{-\frac{(1+b)}{b}}}{\|T^{1/2}CT^{1/2}\|}\Big(\|T^{1/2}CT^{1/2}\|+\lambda\Big)^2 \leq 4\|T^{1/2}CT^{1/2}\|\lambda^{-\frac{1+b}{b}} \lesssim \lambda^{-\frac{1+b}{b}}.$$

Therefore, the condition $n \gtrsim (d(\lambda) \vee \ln(1/\delta))$ is satisfied if $n \gtrsim \ln(1/\delta)$ and $n \gtrsim \lambda^{-\frac{1+b}{b}}$ or equivalently $\lambda \gtrsim n^{-\frac{b}{1+b}}$. Hence, the conditions on n and λ read as

$$n \gtrsim \ln(1/\delta) \qquad \text{and} \qquad n^{-\frac{b}{1+b}} \lesssim \lambda \leq \|T^{1/2}CT^{1/2}\|.$$

Finally to handle the bias term, the assumption $\tilde{\beta}^* \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})^\nu)$, for $\nu \in (0,1]$ implies that $\exists\, h \in L^2(S)$ such that $T^{1/2}(T^{1/2}CT^{1/2})^\nu h \in \tilde{\beta}^*$. Therefore, we have

$$\begin{aligned}
&\|T(CT+\lambda I)^{-1}C\tilde{\beta}^* - \tilde{\beta}^*\|\\
&\overset{(*)}{=} \|T^{1/2}(\Lambda+\lambda I)^{-1}\Lambda\Lambda^\nu h - T^{1/2}\Lambda^\nu h\|\\
&\leq \|T\|^{1/2}\Big\|(T^{1/2}CT^{1/2}+\lambda I)^{-1}(T^{1/2}CT^{1/2})^{\nu+1}h - (T^{1/2}CT^{1/2})^\nu h\Big\|\\
&\leq \|T\|^{1/2}\sup_i \left|\frac{\zeta_i^{1+\nu}}{\zeta_i+\lambda} - \zeta_i^\nu\right|\|h\| \lesssim \lambda \sup_i \left[\frac{i^{-b\nu}}{i^{-b}+\lambda}\right] \lesssim \lambda\lambda^{\frac{b\nu-b}{b}} = \lambda^\nu, \qquad (7.26)
\end{aligned}$$

where $\Lambda := T^{1/2}CT^{1/2}$ in $(*)$. By combining the bounds in (7.21)–(7.26), we obtain

$$\|\hat{\beta}-\tilde{\beta}^*\| \lesssim \frac{1}{\sqrt{\lambda}}\left[\frac{\lambda^{-\frac{1}{2b}}}{\sqrt{n}} + \frac{\lambda(1+\sqrt{\lambda})\lambda^{-1/2}\lambda^{-\frac{1+b}{2b}}}{\sqrt{n}}\right] + \lambda^\nu \leq \frac{\lambda^{-\left(\frac{1}{2}+\frac{1}{2b}\right)}}{\sqrt{n}} + \lambda^\nu.$$

Thus, by setting λ as in (3.8), we obtain the claim in (3.8).

7.4 Proof of Theorem 3.4

We prove Theorem 3.4 by obtaining a better bound on BIAS(λ) under the assumptions in Theorem 3.4. Define $\Lambda := T^{1/2}CT^{1/2}$. The assumption that $\tilde{\beta}^* \in \mathscr{R}\,(T^{1/2}\Lambda^{\nu})$ implies that $\exists\, h \in L^2(S)$ such that $T^{1/2}\Lambda^{\nu}h = \tilde{\beta}^*$. Hence, we have

$$
\begin{aligned}
\textsc{bias}(\lambda) &= \|T(CT+\lambda I)^{-1}C\tilde{\beta}^* - \tilde{\beta}^*\| = \left\|T^{1/2}(\Lambda+\lambda I)^{-1}\Lambda^{1+\nu}h - T^{1/2}\Lambda^{\nu}h\right\| \\
&= \left\|T^{1/2}\sum_i \frac{\zeta_i^{1+\nu}}{\zeta_i+\lambda}\langle\phi_i,h\rangle\phi_i - T^{1/2}\sum_i \zeta_i^{\nu}\langle\phi_i,h\rangle\phi_i\right\| = \left\|\sum_i\left(\frac{\zeta_i^{1+\nu}}{\zeta_i+\lambda} - \zeta_i^{\nu}\right)\langle\phi_i,h\rangle T^{1/2}\phi_i\right\| \\
&= \lambda\left\|\sum_i \frac{\zeta_i^{\nu}}{\zeta_i+\lambda}\langle\phi_i,h\rangle T^{1/2}\phi_i\right\| = \lambda\left\|\sum_i \frac{\zeta_i^{\nu}}{\zeta_i+\lambda}\langle\phi_i,h\rangle\sum_j\sqrt{\mu_j}\langle\phi_i,\psi_j\rangle\psi_j\right\| \\
&= \lambda\left\|\sum_j\sqrt{\mu_j}\left[\sum_i\frac{\zeta_i^{\nu}}{\zeta_i+\lambda}\langle\phi_i,h\rangle\langle\phi_i,\psi_j\rangle\right]\psi_j\right\| = \lambda\left\{\sum_j\mu_j\left[\sum_i\frac{\zeta_i^{\nu}}{\zeta_i+\lambda}\langle\phi_i,h\rangle\langle\phi_i,\psi_j\rangle\right]^2\right\}^{1/2} \\
&= \lambda\left\{\sum_i\sum_\ell\sum_j\mu_j\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\langle\phi_i,h\rangle\langle\phi_\ell,h\rangle\frac{\zeta_i^{\nu}}{\zeta_i+\lambda}\frac{\zeta_\ell^{\nu}}{\zeta_\ell+\lambda}\right\}^{1/2} \\
&= \lambda\left\{\sum_i\sum_\ell\left[\frac{\zeta_i^{\nu}\zeta_\ell^{\nu}}{(\zeta_i+\lambda)(\zeta_\ell+\lambda)}\sum_j\mu_j\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\right]\langle\phi_i,h\rangle\langle\phi_\ell,h\rangle\right\}^{1/2} \\
&\le \lambda\left\{\left\{\sum_i\sum_\ell\left[\frac{\zeta_i^{\nu}\zeta_\ell^{\nu}}{(\zeta_i+\lambda)(\zeta_\ell+\lambda)}\right]^2\left[\sum_j\mu_j\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\right]^2\right\}^{1/2}\left\{\sum_i\sum_\ell\langle\phi_i,h\rangle^2\langle\phi_\ell,h\rangle^2\right\}^{1/2}\right\}^{1/2} \\
&\le \lambda\|h\|\left\{\sum_{i,\ell}\left[\frac{\zeta_i^{\nu}\zeta_\ell^{\nu}}{(\zeta_i+\lambda)(\zeta_\ell+\lambda)}\right]^2\left[\sum_j\mu_j\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\right]^2\right\}^{1/4} \\
&= \lambda\|h\|\left\{\sum_{i,\ell}\left[\frac{\zeta_i^{\nu}\zeta_\ell^{\nu}\sqrt{\mu_i\mu_\ell}}{(\zeta_i+\lambda)(\zeta_\ell+\lambda)}\right]^2\left[\sum_j\frac{\mu_j}{\sqrt{\mu_i\mu_\ell}}\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\right]^2\right\}^{1/4} \\
&\le \lambda\|h\|\left[\sum_{i,\ell}\left[\frac{\zeta_i^{\nu}\zeta_\ell^{\nu}\sqrt{\mu_i\mu_\ell}}{(\zeta_i+\lambda)(\zeta_\ell+\lambda)}\right]^2\right]^{1/4}\left\{\sup_{i,l}\left[\sum_j\frac{\mu_j}{\sqrt{\mu_i\mu_\ell}}\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\right]^2\right\}^{1/4} \\
&= \lambda\|h\|\left[\sum_i\left(\frac{\zeta_i^{\nu}\sqrt{\mu_i}}{\zeta_i+\lambda}\right)^2\right]^{1/2}\left\{\sup_{i,\ell}\frac{1}{\mu_i\mu_\ell}\left|\sum_j\mu_j\langle\phi_i,\psi_j\rangle\langle\phi_\ell,\psi_j\rangle\right|^2\right\}^{1/4} \\
&\overset{(*)}{\lesssim} \lambda\|h\|\left[\sum_i\left(\frac{\zeta_i^{\nu}\sqrt{\mu_i}}{\zeta_i+\lambda}\right)^2\right]^{1/2} \lesssim \lambda\|h\|\left[\sum_i\frac{i^{-(2b\nu+t)}}{(i^{-b}+\lambda)^2}\right]^{1/2} \\
&\overset{(**)}{\lesssim} \lambda\cdot\lambda^{-(1+2b-2b\nu-t)/2b}\|h\|,
\end{aligned}
$$

where $(*)$ follows from the assumption in (3.9) and $(**)$ from Lemma A.5 when $b \geq 2b\nu + t$ and $2b \geq 2b\nu + t$, i.e., $\nu \leq \frac{1}{2} - \frac{t}{2b}$. Therefore,

$$\textsc{bias}(\lambda) = \|T(CT+\lambda)^{-1}C\tilde{\beta}^* - \tilde{\beta}^*\| \lesssim \lambda^{\frac{2b\nu+t-1}{2b}} \|h\|. \tag{7.27}$$

Combining the bounds in (7.21)–(7.25) from the proof of Theorem 3.3 and (7.27), we obtain

$$\|\hat{\beta} - \tilde{\beta}^*\| \lesssim \frac{\lambda^{-\left(\frac{1}{2}+\frac{1}{2b}\right)}}{\sqrt{n}} + \underbrace{\lambda^{\frac{2b\nu+t-1}{2b}}}_{p}.$$

Note that we have $p = o(\lambda^\nu)$ as $\lambda \to 0$. Hence, by our choice of λ, the result follows.

7.5 Proof of Proposition 4.1

First note that k and c are positive definite kernels which follow from the form of k and c and the assumption that $a_i \geq 0$ for all i and $b_m \geq 0$ for all m. Next, we note that we $k(\cdot, x) = \sum_i a_i\phi_i(x)\phi_i(\cdot) \in \mathcal{H}$, and

$$\langle f, k(\cdot,x)\rangle_{\mathcal{H}} = \sum_i \frac{f_i a_i}{a_i}\phi_i(x) = f(x), \quad \forall x \in [0,1],$$

implying $\mathcal{H}$ is an RKHS induced by the kernel k.

By considering the integral operator $Tf = \int_0^1 k(\cdot,x)f(x)dx$, for $f \in L^2([0,1])$, we have

$$Tf = \int_0^1 \sum_i a_i\phi_i(x)f(x)\phi_i dx = \sum_i a_i \left[\int_0^1 \phi_i(x)f(x)d(x)\right]\phi_i = \sum_i a_i\langle f, \phi_i\rangle\phi_i.$$

This implies $T\phi_j = \sum_i a_i\langle\phi_j,\phi_i\rangle\phi_i = \sum_i a_i\delta_{ij}\phi_i = a_j\phi_j$, i.e., $(a_i,\phi_i)_{i\in\mathbb{N}}$ form the pair of eigenvalues and eigenfunctions of the operator T. Since $\sum_i a_i < \infty$ and $a_i \geq 0$, T is also a positive, trace-class operator.

Since X is a mean-zero Gaussian process with covariance function

$$c(x,y) = \sum_m b_m\psi_m(x)\psi_m(y),$$

it follows from the Karhunen-Loéve theorem that X has a representation of the form

$$X = \sum_m \sqrt{b_m}z_m\psi_m,$$

where $z_m \sim N(0,1)$. Furthermore, we have

$$C = \mathbb{E}[X\otimes X] = \mathbb{E}\left[\sum_{m,s}\sqrt{b_m}\sqrt{b_s}z_m z_s\psi_m\otimes\psi_s\right] = \sum_{m,s}\sqrt{b_m}\sqrt{b_s}\mathbb{E}[z_m z_s]\psi_m\otimes\psi_s$$
$$= \sum_{m,s}\sqrt{b_m}\sqrt{b_s}\delta_{ms}\psi_m\otimes\psi_s = \sum_m b_m\psi_m\otimes\psi_m.$$

This implies that $C\psi_\ell = \sum_m b_m\psi_m\langle\psi_m,\psi_\ell\rangle = b_\ell\psi_\ell$. Since $b_m \geq 0$ and $\sum_m b_m < \infty$, C is also a positive, trace-class operator with $(b_i,\psi_i)_{i\in\mathbb{N}}$ as eigenvalue-eigenfunction pairs.

We next characterize the space $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2}))$. Note that by functional calculus, we have $T^{1/2} = \sum_i \sqrt{a_i}\phi_i \otimes \phi_i$. Hence,

$$CT^{1/2} = \sum_{m,i} b_m \sqrt{a_i} \langle \psi_m, \phi_i \rangle \psi_m \otimes \phi_i = \sum_{m,i} b_m \sqrt{a_i} \theta_{mi} \psi_m \otimes \phi_i$$

and

$$\begin{aligned} T^{1/2}CT^{1/2} &= \left[\sum_j \sqrt{a_j} \phi_j \otimes \phi_j \right] \left[\sum_{m,i} b_m \sqrt{a_i} \theta_{mi} \psi_m \otimes \phi_i \right] \\ &= \sum_{i,j,m} \sqrt{a_i a_j} b_m \theta_{mi} \theta_{mj} \phi_j \otimes \phi_i \\ &= \sum_{i,j} \sqrt{a_i a_j} \left[\sum_m b_m \theta_{mi} \theta_{mj} \right] \phi_j \otimes \phi_i \qquad (7.28) \\ &= \sum_{ij} \sqrt{a_i a_j} \eta_{ij} \phi_j \otimes \phi_i. \end{aligned}$$

By a similar calculation, we obtain

$$T^{1/2}(T^{1/2}CT^{1/2}) = \sum_{ij} a_j \sqrt{a_i} \eta_{ij} \phi_j \otimes \phi_i.$$

Since

$$\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})) = \left\{ \tilde{f} \in L^2([0,1]) | \tilde{f} = T^{1/2}(T^{1/2}CT^{1/2})h : h \in L^2([0,1]) \right\},$$

for any function $\tilde{f} \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2}))$, we have $\tilde{f} = T^{1/2}(T^{1/2}CT^{1/2})h$ for some $h \in L^2([0,1])$. By defining $h_i := \langle h, \phi_i \rangle$ and $\beta_j := \sum_i \sqrt{a_i}\eta_{ij}h_i$, we have

$$\tilde{f} = \sum_{ij} a_j \sqrt{a_i} \eta_{ij} h_i \phi_j = \sum_i a_j \beta_j \phi_j.$$

We now show that $\tilde{f} \in \mathcal{H}$. Consider $\|\tilde{f}\|_{\mathcal{H}}^2 = \sum_j a_j \beta_j^2 \leq \|h\|^2 \sum_j a_j \sum_i a_i \eta_{ij}^2$, where

$$\begin{aligned} \eta_{ij}^2 = \left(\sum_m b_m \theta_{mi} \theta_{mj} \right)^2 &\leq \sum_m b_m^2 \theta_{mi}^2 \sum_m \theta_{mj}^2 = \|\phi_j\|^2 \sum_m b_m^2 \theta_{mi}^2 \\ &= \sum_m b_m^2 \theta_{mi}^2 \leq \sum_m b_m^2 \leq \sum_m b_m. \end{aligned}$$

Hence, $\|\tilde{f}\|_{\mathcal{H}}^2 \leq \|h\|^2 \left(\sum_m b_m\right) \left(\sum_i a_i\right)^2 < \infty$, which implies that we have $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})) \subset \mathcal{H}$.

Next, we show that $\mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2})) \subset \tilde{\mathcal{H}}$. To this end, note that for any function $\tilde{f} \in \mathscr{R}(T^{1/2}(T^{1/2}CT^{1/2}))$, we have that $\tilde{f} = \sum_j a_j \beta_j \phi_j$, and by a similar calculation as above we have that $\|\tilde{f}\|_{\tilde{\mathcal{H}}}^2 < \infty$ where we used $\beta_j^2 \leq \|h\|^2 \tau_j$. Finally, since

$$\sum_i \frac{f_i^2}{a_i} = \sum_i \frac{f_i^2 \tau_i}{a_i \tau_i} \leq \sup_i \tau_i \sum_i \frac{f_i^2}{a_i \tau_i} < \infty.$$

it also follows that $\tilde{\mathcal{H}} \subset \mathcal{H}$.

7.6 Proof of Theorem 5.1

Note that

$$\begin{aligned}\|C^{1/2}(\hat{\beta}-\beta^*)\| &\le \|C^{1/2}(\mathfrak{I}\hat{\beta}-\mathfrak{I}\beta_\lambda)\| + \|C^{1/2}(\mathfrak{I}\beta_\lambda-\beta^*)\| \\ &= \|(\mathfrak{I}^*C\mathfrak{I})^{1/2}(\hat{\beta}-\beta_\lambda)\|_{\mathcal{H}} + \|C^{1/2}(\mathfrak{I}\beta_\lambda-\beta^*)\|.\end{aligned}$$

By defining $A := \mathfrak{I}^*C\mathfrak{I}$ and following the steps similar to the proof of Theorem 3.1 in bounding the first term, we obtain

$$\begin{aligned}\|C^{1/2}(\hat{\beta}-\tilde{\beta}^*)\|^2 \lesssim{}& \underbrace{\|A^{1/2}(A+\lambda I)^{-1/2}\|^2}_{\texttt{Term 7}} \cdot (\texttt{Term 2})^2 \cdot (\texttt{Term 3})^2 \\ &\cdot [\texttt{Term 4} + \texttt{Term 5}]^2 + \underbrace{\|C^{1/2}(\mathfrak{I}\beta_\lambda-\beta^*)\|^2}_{\texttt{Term 8}}.\end{aligned}$$

We will now proceed with bounding `Term 7` and `Term 8`. For `Term 7`, note that

$$\|A^{1/2}(A+\lambda I)^{-1/2}\| \le 1.$$

To bound `Term 8`, note that

$$\begin{aligned}\|C^{1/2}(\mathfrak{I}\beta_\lambda-\beta^*)\| &= \|C^{1/2}\mathfrak{I}(\mathfrak{I}^*C\mathfrak{I}+\lambda I)^{-1}\mathfrak{I}^*C\beta^* - C^{1/2}\beta^*\| \\ &= \|C^{1/2}T(CT+\lambda I)^{-1}C\beta^* - C^{1/2}\beta^*\|.\end{aligned}$$

The result therefore follows by combining the bounds on `Term 7` and `Term 8`, along with the bounds for `Term 2` to `Term 5` from the proof of Theorem 3.1.

7.7 Proof of Theorem 5.2

We first deal with the bias term. Since $\beta^* \in \mathscr{R}(T^{1/2})$, $\exists\, h \in L^2(S)$ such that $\beta^* = T^{1/2}h$. Therefore, we have

$$\begin{aligned}&\|C^{1/2}T(CT+\lambda I)^{-1}C\beta^* - C^{1/2}\beta^*\| \\ &= \|C^{1/2}T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^{1/2}h - C^{1/2}T^{1/2}h\| \\ &= \left\|C^{1/2}T^{1/2}\left[(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^{1/2}h - h\right]\right\| \\ &= \|(T^{1/2}CT^{1/2})^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^{1/2}h - (T^{1/2}CT^{1/2})^{1/2}h\| \\ &= \left[\sum_i \left(\frac{\zeta_i^{3/2}}{\zeta_i+\lambda} - \zeta_i^{1/2}\right)^2 \langle\phi_i, h\rangle^2\right]^{1/2} \le \lambda\|h\| \sup_i \frac{\zeta_i^{1/2}}{\zeta_i+\lambda} \\ &\lesssim \lambda\|h\| \sup_i \frac{i^{-\frac{b}{2}}}{i^{-b}+\lambda} \lesssim \lambda\|h\|\lambda^{-1/2} = \sqrt{\lambda}\|h\|,\end{aligned}$$

where the last inequality follows from Lemma A.6. Since $\alpha = 1/2$, by using (7.22), (7.23) and (7.24) respectively for bounding $N(\lambda)$, trace(Θ) and $\|\Theta\|$, along with the above bound on the bias, we obtain

$$\|C^{1/2}(\hat{\beta}-\beta^*)\|^2 \lesssim \frac{\lambda^{-\frac{1}{b}}}{n} + \lambda.$$

Hence, by setting λ as in (5.1), we obtain the claim in (5.1).

7.8 Proof of Theorem 5.3

We first bound the bias term. Since $\beta^* \in \mathscr{R}(T^\alpha)$, $\exists\, h \in L^2(S)$ such that $\beta^* = T^\alpha h$. Therefore, we have

$$\begin{aligned}
&\|C^{1/2}T(CT+\lambda I)^{-1}C\beta^* - C^{1/2}\beta^*\| \\
&= \|C^{1/2}T(CT+\lambda I)^{-1}CT^\alpha h - C^{1/2}T^\alpha h\| \\
&= \|C^{1/2}T^{1/2}(T^{1/2}CT^{1/2}+\lambda I)^{-1}T^{1/2}CT^\alpha h - C^{1/2}T^\alpha h\| \\
&= \left[\sum_i \left(\frac{\mu_i^{1+\alpha}\xi_i^{3/2}}{\mu_i\xi_i+\lambda} - \xi_i^{1/2}\mu_i^\alpha\right)^2 \langle\phi_i, h\rangle^2\right]^{1/2} \\
&\le \lambda\|h\| \sup_i \frac{\xi_i^{1/2}\mu_i^\alpha}{\mu_i\xi_i+\lambda} \lesssim \lambda\|h\| \sup_i \frac{i^{-(\alpha t+c/2)}}{i^{-(t+c)}+\lambda} \\
&\lesssim \lambda\|h\|\lambda^{\frac{\alpha t+c/2-t-c}{t+c}} = \lambda^{\frac{\alpha t+c/2}{t+c}}\|h\|,
\end{aligned}$$

where the last inequality follows from Lemma A.6. This upper bound on the bias, along with (7.17) and (7.18) from the proof of Theorem 3.2 implies that

$$\|C^{1/2}(\hat{\beta}-\beta^*)\|^2 \lesssim \frac{\lambda^{-\frac{1+2t(1-2\alpha)}{t+c}}}{n} + \lambda^{\frac{2\alpha t+c}{t+c}}.$$

Thus by setting λ as in (3.7), we obtain the claim in (5.2).

Acknowledgements

KB is supported in part by the National Science Foundation (NSF) grant DMS-2053918 and a UC Davis Center for Data Science and Artificial Intelligence Research (CEDAR) seed grant. HGM is supported in part by NSF grant DMS-2014526 and NIH grant UG3-0D023313. BKS is supported in part by the NSF CAREER Award DMS-1945396.

References

D. Babichev and F. Bach. Slice inverse regression with score functions. *Electron. J. Stat.*, 12(1): 1507–1543, 2018.

K. Balasubramanian, H.-G. Müller, and B. K. Sriperumbudur. Supplement to “functional linear and single-index models: A unified approach via gaussian stein identity”. *Bernoulli*, 2024.

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

D. R. Brillinger. A generalized linear model with gaussian regressor variables. In *A Festschrift for Erich L. Lehmann (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.)*, pages 97–114. Wadsworth Publishing Group, 1983.

B. A. Brumback and J. A. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.*, 93(443):961–994, 1998.

T. T. Cai and M. Yuan. Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.*, 107(499):1201–1216, 2012.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statist. Probab. Lett.*, 45:11–22, 1999.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statist. Sinica*, 13(3):571–591, 2003. ISSN 1017-0405.

D. Chen, P. Hall, and H.-G. Müller. Single and multiple index functional regression models with nonparametric link. *Ann. Statist.*, 39:1720–1747, 2011.

Y. Chen and H.-G. Müller. Uniform convergence of local Fréchet regression, with applications to locating extrema and time warping for metric-space valued trajectories. *Ann. Statist.*, 50: 1573–1592, 2023.

H. O. Cordes. *Spectral Theory of Linear Differential Operators and Comparison Algebras*, volume 76. Cambridge University Press, 1987.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39 (1):1–49, 2002.

A. Cuevas, M. Febrero, and R. Fraiman. Linear functional regression: The case of fixed design and functional response. *Canad. J. Statist.*, 30(2):285–300, 2002. ISSN 0319-5724.

E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, F. Odone, and P. Bartlett. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6(5):883–904, 2005.

L. H. Dicker, D. P. Foster, and D. Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electron. J. Stat.*, 11(1):1022–1047, 2017.

R. F. Engle, C. W. Granger, J. Rice, and A. Weiss. Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.*, 81(394):310–320, 1986.

L. Goldstein and X. Wei. Non-Gaussian observations in nonlinear compressed sensing via Stein discrepancies. *Inf. Inference J. IMA*, 8(1):125–159, 2019.

L. Goldstein, S. Minsker, and X. Wei. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Trans. Inform. Theory*, 64(8):5513–5530, 2018.

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

U. Grenander. *Probabilities on Algebraic Structures.* Courier Corporation, 2008.

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

P. Hall and I. Van Keilegom. Two-sample tests in functional data analysis starting from discrete data. *Statist. Sinica*, 17:1511–1531, 2007.

G. He, H.-G. Müller, and J.-L. Wang. Extending correlation and regression from multivariate to functional data. In M. L. Puri, editor, *Asymptotics in Statistics and Probability*, pages 301–315. VSP International Science Publishers, 2000.

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

T. Hsing and H. Ren. An rkhs formulation of the inverse regression dimension-reduction problem. *Ann. Statist.*, 37(2):726–755, 2009.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–601, 2014.

G. M. James. Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 64(3):411–432, 2002. ISSN 1369-7412.

C.-R. Jiang and J.-L. Wang. Functional single index models for longitudinal data. *Ann. Statist.*, 39(1):362–388, 2011.

C.-R. Jiang, W. Yu, and J.-L. Wang. Inverse regression for longitudinal data. *Ann. Statist.*, 42(2): 563–591, 2014.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33:82–95, 1971.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23:110–133, 2017.

H.-H. Kuo and Y.-J. Lee. Integration by parts formula and the stein lemma on abstract wiener space. *Commun. Stoch. Anal.*, 5(2):10, 2011.

B. Li and J. Song. Nonlinear sufficient dimension reduction for functional data. *Ann. Statist.*, 45 (3):1059–1095, 2017.

K.-C. Li. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, 86(414): 316–327, 1991.

K.-C. Li. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. *J. Amer. Statist. Assoc.*, 87(420):1025–1039, 1992.

K.-C. Li and N. Duan. Regression analysis under link violation. *Ann. Statist.*, 17(3):1009–1052, 1989.

Y. Li and T. Hsing. Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *Ann. Statist.*, 38(5):3028–3062, 2010.

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

J. S. Morris. Functional regression. *Annu. Rev. Stat. Appl.*, 2:321–359, 2015.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Ann. Statist.*, 33(2): 774–805, 2005.

H.-G. Müller, U. Stadtmüller, and F. Yao. Functional variance processes. *J. Amer. Statist. Assoc.*, 101:1007–1018, 2006.

Y. Plan and R. Vershynin. The generalized Lasso with non-linear observations. *IEEE Trans. Inform. Theory*, 62(3):1528–1537, 2016.

B. S. Rajput. Gaussian measures on l_p spaces, $1 \leq p \leq \infty$. *J. Multivariate Anal.*, 2(4):382–403, 1972.

B. S. Rajput and S. Cambanis. Gaussian processes and gaussian measures. *Ann. Math. Stat.*, pages 1944–1952, 1972.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 53(3):539–572, 1991. ISSN 0035-9246.

B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In *Proc. of the 14th Annual Conference on Computational Learning Theory*, pages 416–426, London, UK, 2001. Springer-Verlag.

Z. Shang and G. Cheng. Nonparametric inference in generalized functional linear models. *Ann. Statist.*, 43(4):1742–1773, 2015.

H.-H. Shih. On stein's method for infinite-dimensional gaussian approximation in abstract wiener spaces. *J. Funct. Anal.*, 261(5):1236–1283, 2011.

S. Smale and D.-X. Zhou. Shannon sampling ii: Connections to learning theory. *Appl. Comput. Harmon. Anal.*, 19(3):285–302, 2005.

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

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

H. Tong and M. Ng. Analysis of regularized least squares for functional linear regression model. *J. Complexity*, 49:85–94, 2018.

C. Wang and D.-X. Zhou. Optimal learning rates for least squares regularized regression with unbounded sampling. *J. Complexity*, 27(1):55–67, 2011.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annu. Rev. Stat. Appl.*, 3: 257–295, 2016.

Q. Wu, Y. Ying, and D.-X. Zhou. Learning rates of least-square regularized regression. *Found. Comput. Math.*, 6(2):171–192, 2006.

Z. Yang, K. Balasubramanian, and H. Liu. High-dimensional non-Gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning*, pages 3851–3860. PMLR, 2017a.

Z. Yang, K. Balasubramanian, Z. Wang, and H. Liu. Estimating high-dimensional non-Gaussian multiple index models via Stein's lemma. In *Advances in Neural Information Processing Systems*, volume 30, 2017b.

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

H. Zhu, F. Yao, and H. H. Zhang. Structured functional additive regression in reproducing kernel Hilbert spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(3):581–603, 2014.

A Auxiliary results

In this section, we collect some technical results used to prove the main results.

Theorem A.1 (Shih, 2011; Kuo and Lee, 2011)**.** *Let H be a separable Hilbert space, with norm $\|\cdot\|_H$ and inner-product $\langle\cdot,\cdot\rangle_H$. Let $\tilde{H}$ be the completion with respect to $\|\cdot\|_H$. Let p be a Gaussian measure on $\tilde{H}$. Then, for any $h \in H$ and for any once Fréchet-differentiable function $f : \tilde{H} \to \mathbb{R}$, we have $\int_{\tilde{H}} \langle h, x\rangle_H \, dp(x) = \int_{\tilde{H}} \langle \nabla f(x), h\rangle_H \, dp(x)$, where ∇ represents the Fréchet derivative, as long as the expectations are well-defined.*

Lemma A.2 (Chebychev's inequality for Hilbert-valued random variables)**.** *Let $Z_i \in H$, for $i = 1, \dots, n$ be i.i.d. Hilbert-valued random variables such that $\mathbb{E}[Z_i] = 0$. Then*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_H \geq \sqrt{\frac{\mathbb{E}\|Z_1\|_{\mathcal{H}}^2}{n\delta}}\right) \leq \delta.$$

Proof. By Markov's inequality, it is obvious that for any $\epsilon > 0$

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_H \geq \epsilon\right) \leq \frac{\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_H^2}{\epsilon^2}.$$

By noting

$$\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_H^2 = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}\langle Z_i, Z_j\rangle_H = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\,\|Z_i\|_H^2 + \frac{1}{n^2}\sum_{i\neq j}^n \mathbb{E}\langle Z_i, Z_j\rangle_H = \frac{\mathbb{E}\|Z_1\|_H^2}{n}$$

and choosing $\epsilon = \sqrt{\frac{\mathbb{E}\|Z_1\|_H^2}{n\delta}}$ yields the result. □

Theorem A.3 (Koltchinskii and Lounici, 2017)**.** *Let $X_1, \dots, X_n$ be i.i.d. centered Gaussian random variables in a separable Hilbert space H with covariance operator $\Sigma = \mathbb{E}[X \otimes_H X]$. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1} X_i \otimes_H X_i$ be the empirical covariance operator. Define*

$$r(\Sigma) := \frac{(\mathbb{E}\|X\|_H)^2}{\|\Sigma\|_{\text{op}}}.$$

Then for all $\tau \geq 1$ and $n \geq (r(\Sigma) \vee \tau)$, with probability at least $1 - e^{-\tau}$,

$$\|\hat{\Sigma} - \Sigma\|_{\text{op}} \leq K_1\|\Sigma\|_{\text{op}}\frac{\sqrt{r(\Sigma)} + \sqrt{\tau}}{\sqrt{n}},$$

where K_1 is a universal constant independent of Σ, τ and n.

Lemma A.4. *For any bounded, self-adjoint positive operator* Γ *on* $L^2(S)$,

$$\mathbb{E}[\langle X, \Gamma X\rangle^2] \leq 3\mathrm{trace}^2(\Gamma C),$$

assuming $\mathrm{trace}(C^{1/2}) < \infty$ *with* $C = \mathbb{E}[X \otimes X]$.

Proof. First note that by the Karhunen-Loéve expansion, we have $X = \sum_i x_i \varphi_i$, where $(\varphi_i)_{i\in\mathbb{N}}$ and $(\lambda_i)_{i\in\mathbb{N}}$ are the eigenfunctions and eigenvalues of C, and x_i are independent Gaussian random variables with $\mathbb{E}[x_i^2] = \lambda_i$. Hence, we have

$$\begin{aligned}
\mathbb{E}[\langle X, \Gamma X\rangle^2] &= \mathbb{E}\left[\left\langle \sum_i x_i\varphi_i, \sum_j x_j \Gamma\varphi_j \right\rangle^2\right] = \mathbb{E}\left[\sum_{i,j} x_i x_j \langle \varphi_i, \Gamma\varphi_j\rangle\right]^2 \\
&\overset{(*)}{=} \sum_{i,j,k,\ell} \mathbb{E}[x_i x_j x_k x_\ell]\langle \varphi_i, \Gamma\varphi_j\rangle\langle \varphi_k, \Gamma\varphi_\ell\rangle \\
&= \sum_i \mathbb{E}[x_i^4]\langle \varphi_i, \Gamma\varphi_i\rangle^2 + \sum_{i\neq j} \mathbb{E}[x_i^2]\mathbb{E}[x_j^2]\langle \varphi_i, \Gamma\varphi_i\rangle\langle \varphi_j, \Gamma\varphi_j\rangle \\
&= 3\sum_i \lambda_i^2 \langle \varphi_i, \Gamma\varphi_i\rangle^2 + \left(\sum_i \lambda_i\langle \varphi_i, \Gamma\varphi_i\rangle\right)^2 - \sum_i \lambda_i^2\langle \varphi_i, \Gamma\varphi_i\rangle^2 \\
&= 2\sum_i \lambda_i^2 \langle \varphi_i, \Gamma\varphi_i\rangle^2 + \left(\sum_i \lambda_i\langle \varphi_i, \Gamma\varphi_i\rangle\right)^2, \qquad \text{(A.1)}
\end{aligned}$$

where the exchange of expectation and summation in $(*)$ holds by Fubini's theorem if

$$\sum_{i,j,k,\ell} \mathbb{E}|x_i x_j x_k x_\ell||\langle \varphi_i, \Gamma\varphi_j\rangle||\langle \varphi_k, \Gamma\varphi_\ell\rangle| < \infty. \qquad \text{(A.2)}$$

We will later verify (A.2). Note that

$$\begin{aligned}
\mathrm{trace}(\Gamma C) &= \mathrm{trace}\left(\Gamma\left(\sum_i \lambda_i \varphi_i \otimes \varphi_i\right)\right) \\
&= \mathrm{trace}\left(\sum_i \lambda_i\left(\Gamma\varphi_i\right)\otimes\varphi_i\right) = \sum_i \lambda_i\langle\varphi_i, \Gamma\varphi_i\rangle. \qquad \text{(A.3)}
\end{aligned}$$

Furthermore, recall that the Hilbert-Schmidt norm of an operator A is defined as $\|A\|_{HS}^2 := \sum_i \|Ae_i\|^2$, where $(e_i)_{i\in\mathbb{N}}$, is any orthonormal basis for $L^2(S)$. Hence, we have

$$\begin{aligned}
&\left\|\sum_i \lambda_i\langle\varphi_i, \Gamma\varphi_i\rangle\left(\varphi_i\otimes\varphi_i\right)\right\|_{HS}^2 \\
&= \left\langle \sum_i \lambda_i\langle\varphi_i, \Gamma\varphi_i\rangle\left(\varphi_i\otimes\varphi_i\right), \sum_j \lambda_j\langle\varphi_j, \Gamma\varphi_j\rangle\left(\varphi_j\otimes\varphi_j\right)\right\rangle_{HS} \\
&= \sum_{i,j} \lambda_i\lambda_j\langle\varphi_i, \Gamma\varphi_i\rangle\langle\varphi_j, \Gamma\varphi_j\rangle\left\langle\varphi_i\otimes\varphi_i, \varphi_j\otimes\varphi_j\right\rangle_{HS} \\
&= \sum_{i,j} \lambda_i\lambda_j\langle\varphi_i, \Gamma\varphi_i\rangle\langle\varphi_j, \Gamma\varphi_j\rangle\left\langle\varphi_i, \varphi_j\right\rangle^2 = \sum_i \lambda_i^2\langle\varphi_i, \Gamma\varphi_i\rangle^2.
\end{aligned}$$

This means

$$\sum_i \lambda_i^2 \langle \varphi_i, \Gamma\varphi_i\rangle^2 \le \operatorname{trace}^2\left(\sum_i \lambda_i \langle \varphi_i, \Gamma\varphi_i\rangle \varphi_i \otimes \varphi_i\right) = \left(\sum_i \lambda_i \langle \varphi_i, \Gamma\varphi_i\rangle \langle \varphi_i, \varphi_i\rangle\right)^2$$

$$= \left(\sum_i \lambda_i \langle \varphi_i, \Gamma\varphi_i\rangle\right)^2 = \operatorname{trace}^2\left(\Gamma C\right), \tag{A.4}$$

Combining (A.3) and (A.4) in (A.1) yields the result. We will now verify (A.2). Note that

$$\begin{aligned}
&\sum_{i,j,k,\ell} \mathbb{E}|x_i x_j x_k x_\ell| |\langle \varphi_i, \Gamma\varphi_j\rangle| |\langle \varphi_k, \Gamma\varphi_\ell\rangle| \le \|\Gamma\|^2 \sum_{i,j,k,\ell} \mathbb{E}|x_i x_j x_k x_\ell| \\
&\lesssim \|\Gamma\|^2 \left(\sum_i \mathbb{E}[x_i^4] + \sum_{i\neq j} \mathbb{E}[|x_i|^3]\mathbb{E}[|x_j|] + \sum_{i\neq j} \mathbb{E}[x_i^2]\mathbb{E}[x_j^2] \right.\\
&\qquad\qquad \left. + \sum_{i\neq j\neq k\neq \ell} \mathbb{E}[|x_i|]\mathbb{E}[|x_j|]\mathbb{E}[|x_j|]\mathbb{E}[|x_\ell|]\right) \\
&\le \|\Gamma\|^2 \left(3\sum_i \lambda_i^2 + \frac{4}{\pi}\sum_{i\neq j} \lambda_i^{3/2}\sqrt{\lambda_j} + \sum_{i\neq j}\lambda_i\lambda_j + \frac{4}{\pi^2}\left(\sum_i \sqrt{\lambda_i}\right)^4\right) \\
&\lesssim \|\Gamma\|^2 \left(\sum_i \lambda_i^2 + \left(\sum_i \lambda_i^{3/2}\right)\left(\sum_j \sqrt{\lambda_j}\right) + \left(\sum_i \lambda_i\right)^2 + \left(\sum_i \sqrt{\lambda_i}\right)^4\right) \\
&< \infty,
\end{aligned}$$

which completes the proof. □

Lemma A.5. *For $\beta \ge \alpha > 1$, and $\gamma \ge \frac{\alpha}{\beta}$, we have*

$$\sum_{i\in\mathbb{N}} \frac{i^{-\alpha}}{(i^{-\beta}+\lambda)^\gamma} \le \lambda^{-\frac{1+\beta\gamma-\alpha}{\beta}} 2^{1-\frac{\alpha}{\beta}} \int_0^\infty \frac{1}{1+y^\alpha}\, dy.$$

Proof. Note that

$$\begin{aligned}
\sum_{i\in\mathbb{N}} \frac{i^{-\alpha}}{(i^{-\beta}+\lambda)^\gamma} &= \sum_{i\in\mathbb{N}} \frac{i^{\gamma\beta-\alpha}}{(1+\lambda i^\beta)^\gamma} \le \int_0^\infty \frac{x^{\beta\gamma-\alpha}}{(1+\lambda x^\beta)^\gamma}\, dx \\
&= \int_0^\infty \frac{\lambda^{-\frac{(\beta\gamma-\alpha)}{\beta}} y^{\beta\gamma-\alpha}}{(1+y^\beta)^\gamma} \lambda^{-1/\beta}\, dy = \lambda^{-\frac{(1+\beta\gamma-\alpha)}{\beta}} \int_0^\infty \frac{y^{\beta\gamma-\alpha}}{(1+y^\beta)^\gamma}\, dy,
\end{aligned}$$

where

$$\begin{aligned}
\int_0^\infty \frac{y^{\beta\gamma-\alpha}}{(1+y^\beta)^\gamma}\, dy &= \int_0^\infty \frac{y^{\beta\left(\gamma-\frac{\alpha}{\beta}\right)}}{(1+y^\beta)^\gamma}\, dy = \int_0^\infty \left(\frac{y^\beta}{1+y^\beta}\right)^{\gamma-\frac{\alpha}{\beta}} \left(1+y^\beta\right)^{-\frac{\alpha}{\beta}}\, dy \\
&\le \int_0^\infty (1+y^\beta)^{-\frac{\alpha}{\beta}}\, dy \qquad (\because\ \gamma \ge \alpha/\beta).
\end{aligned}$$

Therefore,

$$\int_0^\infty \frac{1}{(1+y^\beta)^{\alpha/\beta}}\,dy = 2^{-\alpha/\beta}\int_0^\infty \left(\frac{1}{2}+\frac{y^\beta}{2}\right)^{-\alpha/\beta} dy \le 2^{-\alpha/\beta}\int_0^\infty \left(\frac{1}{2}+\frac{y^\alpha}{2}\right)^{-1} dy,$$

where the last inequality uses the fact that

$$\left(\frac{1}{2}+\frac{y^\beta}{2}\right)^{\alpha/\beta} \ge \left(\frac{1}{2}+\frac{y^\alpha}{2}\right)$$

by Jensen's inequality, because the function $f(x) = x^{\alpha/\beta}$, $x \in [0,\infty)$ is concave for $\alpha \le \beta$. Hence, we have

$$\int_0^\infty \frac{y^{\beta\gamma-\alpha}}{(1+y^\beta)^\gamma}\,dy \le 2^{1-\frac{\alpha}{\beta}}\int_0^\infty (1+y^\alpha)^{-1}\,dy < \infty$$

as long as $\alpha > 1$. □

Lemma A.6. *For any* $0 \le \alpha \le \beta$,

$$\sup_{i\in\mathbb{N}}\left[\frac{i^{-\alpha}}{i^{-\beta}+\lambda}\right] \le \lambda^{\frac{\alpha-\beta}{\beta}}.$$

Proof. Note that

$$\begin{aligned}
\sup_{i\in\mathbb{N}}\left[\frac{i^{-\alpha}}{i^{-\beta}+\lambda}\right] &= \sup_{i\in\mathbb{N}}\left[\frac{i^{\beta-\alpha}}{1+\lambda i^\beta}\right] \le \sup_{x\in(0,\infty)}\frac{x^{\beta-\alpha}}{1+\lambda x^\beta} = \sup_{t\in(0,\infty)}\frac{(t/\lambda)^{\frac{\beta-\alpha}{\beta}}}{1+t} \\
&= \lambda^{\frac{\alpha-\beta}{\beta}}\sup_{t\in(0,\infty)}\frac{t^{\frac{\beta-\alpha}{\beta}}}{1+t} = \lambda^{\frac{\alpha-\beta}{\beta}}\sup_{t\in(0,\infty)}\left(\frac{t}{1+t}\right)^{\frac{\beta-\alpha}{\beta}}\frac{1}{(1+t)^{1-\frac{(\beta-\alpha)}{\beta}}} \\
&\le \lambda^{\frac{\alpha-\beta}{\beta}}\sup_{t\in(0,\infty)}\frac{1}{(1+t)^{\alpha/\beta}} = \lambda^{\frac{\alpha-\beta}{\beta}},
\end{aligned}$$

which completes the proof. □

Lemma A.7.

$$\int_0^1 \cos(ax)\cos(bx)dx = \frac{b}{b^2-a^2}\cos(a)\sin(b) - \frac{a}{b^2-a^2}\sin(a)\cos(b).$$

Proof. Define $J = \int_0^1 \cos(ax)\cos(bx)dx$. Then, we have

$$\begin{aligned}
J &= \left(\cos(ax)\frac{\sin(bx)}{b}\right)_0^1 + a\int_0^1 \sin(ax)\frac{\sin(bx)}{b}dx \\
&= \frac{1}{b}\cos(a)\sin(b) + \frac{a}{b}\left[\left(\sin(ax)\frac{\cos(bx)}{b}\right)_1^0 + \frac{a}{b}\int_0^1 \cos(ax)\cos(bx)dx\right] \\
&= \frac{1}{b}\cos(a)\sin(b) - \frac{a}{b^2}\sin(a)\cos(b) + \left(\frac{a}{b}\right)^2 J
\end{aligned}$$

from which we get the desired result. □

Lemma A.8. *Let ψ_m be as defined in (4.2). Then, we have*

$$\langle \psi_m(\cdot), \cos(i\pi\cdot)\rangle_{L^2} \le \frac{4}{i\pi} 2^{\lfloor \log_2 m \rfloor/2}.$$

Proof. For a given m, consider $j = \lfloor \log_2 m \rfloor$ and $\ell = m + 1 - 2^j$. Then, we have

$$\begin{aligned}
&\langle \psi_m(\cdot), \cos(i\pi\cdot)\rangle_{L^2} \\
=&\int_0^1 \psi_m(x)\cos(i\pi x)dx \\
=&\int_{\frac{\ell-1}{2^j}}^{\frac{\ell-1/2}{2^j}} 2^{j/2}\cos(i\pi x)dx - \int_{\frac{\ell-1/2}{2^j}}^{\frac{\ell}{2^j}} 2^{j/2}\cos(i\pi x)dx \\
=&\frac{2^{j/2}}{i\pi}\left[\sin(i\pi x)\right]_{\frac{\ell-1}{2^j}}^{\frac{\ell-1/2}{2^j}} - \frac{2^{j/2}}{i\pi}\left[\sin(i\pi x)\right]_{\frac{\ell-1/2}{2^j}}^{\frac{\ell}{2^j}} \\
=&\frac{2^{j/2}}{i\pi}\left[\sin\left(\frac{i\pi(\ell-1/2)}{2^j}\right) - \sin\left(\frac{i\pi(\ell-1)}{2^j}\right)\right] \\
&\qquad - \frac{2^{j/2}}{i\pi}\left[\sin\left(\frac{i\pi\ell}{2^j}\right) - \sin\left(\frac{i\pi(\ell-1/2)}{2^j}\right)\right] \\
=&\frac{2^{j/2}}{i\pi}\left[2\sin\left(\frac{i\pi(\ell-1/2)}{2^j}\right) - \sin\left(\frac{i\pi(\ell-1)}{2^j}\right) - \sin\left(\frac{i\pi\ell}{2^j}\right)\right] \\
\le&\frac{4}{i\pi}2^{j/2},
\end{aligned}$$

thereby completing the proof. □

B Bound on η_{ij} in (4.1)

From (4.1), we have

$$\eta_{ij} = \sum_m b_m \theta_{mi}\theta_{mj} = \frac{1}{\pi^2}\sum_m b_m \frac{\omega_m}{\omega_m^2 - i^2}\frac{\omega_m}{\omega_m^2 - j^2}\sin^2(\pi\omega_m)(-1)^{i+j}$$

which implies

$$|\eta_{ij}| \lesssim \sqrt{\sum_m \left(\frac{\omega_m\sqrt{b_m}}{\omega_m^2 - i^2}\right)^2}\sqrt{\sum_m \left(\frac{\omega_m\sqrt{b_m}}{\omega_m^2 - j^2}\right)^2}.$$

Consider

$$\begin{aligned}
\sum_m \left(\frac{\omega_m\sqrt{b_m}}{\omega_m^2 - j^2}\right)^2 &= \sum_m \frac{\omega_m^2 b_m}{(\omega_m + j)^2(\omega_m - j)^2} \le \frac{1}{j^2}\sum_m \left(\frac{\omega_m}{\omega_m - j}\right)^2 b_m \\
&\lesssim \frac{1}{j^2}\sum_m m^{-(1+\delta)}\left(\frac{am+b}{am+b-j}\right)^2 \\
&= \frac{1}{j^2}\sum_m m^{-(1+\delta)}\left(\frac{m+b/a}{m+(b-j)/a}\right)^2.
\end{aligned}$$

We now consider two cases. First, we consider the case of $b < 0$. Define $c \coloneqq (j-b)/a - \lfloor (j-b)/a \rfloor$ and note that $0 < c < 1$. Then, we have

$$\begin{aligned}
&\sum_m m^{-(1+\delta)} \left(\frac{m + b/a}{m + (b-j)/a} \right)^2 \\
&\lesssim \sum_m m^{-(1+\delta)} \left(\frac{m}{m - (j-b)/a} \right)^2 = \sum_m \frac{m^{1-\delta}}{(m - (j-b)/a)^2} \\
&= \sum_{m \leq \lfloor \frac{(j-b)}{a} \rfloor} \frac{m^{1-\delta}}{(m - (j-b)/a)^2} + \sum_{m = \lfloor \frac{(j-b)}{a} \rfloor + 1}^{\infty} \frac{m^{1-\delta}}{(m - (j-b)/a)^2} \\
&= \sum_{m \leq \lfloor \frac{(j-b)}{a} \rfloor} \frac{m^{1-\delta}}{((j-b)/a - m)^2} + \sum_{m = \lfloor \frac{(j-b)}{a} \rfloor + 1}^{\infty} \frac{m^{1-\delta}}{(m - (j-b)/a)^2} \\
&= \left[\frac{\lfloor \frac{j-b}{a} \rfloor^{1-\delta}}{c^2} + \frac{\lfloor \frac{j-b}{a} - 1 \rfloor^{1-\delta}}{(1+c)^2} + \cdots + \frac{1}{\left(\frac{j-b}{a} - 1 \right)^2} \right] \\
&\qquad + \left[\frac{\left(1 + \lfloor \frac{j-b}{a} \rfloor \right)^{1-\delta}}{(1-c)^2} + \frac{\left(2 + \lfloor \frac{j-b}{a} \rfloor \right)^{1-\delta}}{(2-c)^2} + \cdots \right] \\
&\leq \left\lfloor \frac{j-b}{a} \right\rfloor^{1-\delta} \left(\frac{1}{c^2} + 1 + \frac{1}{2^2} + \frac{1}{3^2} + \cdots \right) + \frac{\left(1 + \lfloor \frac{j-b}{a} \rfloor \right)^{1-\delta}}{(1-c)^2} \\
&\qquad + \sum_{\ell=1}^{\infty} \frac{\left(\ell + 1 + \lfloor \frac{j-b}{a} \rfloor \right)^{1-\delta}}{\ell^2} \\
&\leq \left\lfloor \frac{j-b}{a} \right\rfloor^{1-\delta} \left(\frac{1}{c^2} + \frac{\pi^2}{6} \right) + \begin{cases} \frac{1 + \lfloor \frac{j-b}{a} \rfloor^{1-\delta}}{(1-c)^2} + \sum_{\ell=1}^{\infty} \frac{(\ell+1)^{1-\delta} + \lfloor \frac{j-b}{a} \rfloor^{1-\delta}}{\ell^2}, & \text{for } \delta \leq 1 \\ \frac{1}{(1-c)^2} + \sum_{\ell=1}^{\infty} \frac{1}{\ell^2}, & \text{for } \delta > 1 \end{cases} \\
&\lesssim \left\lfloor \frac{j-b}{a} \right\rfloor^{1-\delta} + C_1,
\end{aligned}$$

where we have used the fact that for $\delta \leq 1$,

$$\sum_{\ell=1}^{\infty} \frac{(\ell+1)^{1-\delta}}{\ell^2} \leq \sum_{\ell=1}^{\infty} \frac{\ell^{1-\delta} + 1}{\ell^2} = \frac{\pi^2}{6} + \sum_{\ell=1}^{\infty} \frac{1}{\ell^{1+\delta}} < \infty,$$

and C_1 is some constant independent of the index j. Hence

$$\begin{aligned}
\sum_m \frac{\omega_m^2 b_m}{(\omega_m + j)^2 (\omega_m - j)^2} &\lesssim \frac{1}{j^2} \left[\left\lfloor \frac{j-b}{a} \right\rfloor^{1-\delta} + C_1 \right] \leq \frac{1}{j^2} \left[\left(\frac{j-b}{a} \right)^{1-\delta} + C_1 \right] \\
&\lesssim j^{-\min(1+\delta, 2)}.
\end{aligned}$$

Next we consider the case of $b > 0$. Note that for $j < b$,

$$\sum_m m^{-(1+\delta)} \left(\frac{m + b/a}{m + (b-j)/a} \right)^2 \leq \sum_m m^{-(1+\delta)} \left(\frac{m + b/a}{m} \right)^2$$

$$
\begin{aligned}
&= \sum_m m^{-(1+\delta)} \left(1 + \frac{b}{am}\right)^2 \\
&\lesssim \sum_m m^{-(1+\delta)} \left[1 + \frac{1}{m^2}\right] := C_2 < \infty.
\end{aligned}
$$

Now, for $j > b$, we have

$$
\begin{aligned}
&\sum_m m^{-(1+\delta)} \left(\frac{m + b/a}{m + (b-j)/a}\right)^2 = \sum_m m^{-(1+\delta)} \frac{(m+b/a)^2}{(m-(j-b)/a)^2} \\
&\lesssim \sum_m m^{-(1+\delta)} \frac{m^2 + (b/a)^2}{(m-(j-b)/a)^2} \\
&= \sum_m \frac{m^{1-\delta}}{(m-(j-b)/a)^2} + \left(\frac{b}{a}\right)^2 \sum_m \frac{m^{-(1+\delta)}}{(m-(j-b)/a)^2} \\
&\leq \left[1 + \left(\frac{b}{a}\right)^2\right] \sum_m \frac{m^{1-\delta}}{(m-(j-b)/a)^2} \lesssim \left\lfloor \frac{j-b}{a} \right\rfloor^{1-\delta} + C_1.
\end{aligned}
$$

Therefore,

$$
\sum_m m^{-(1+\delta)} \left(\frac{m + b/a}{m + (b-j)/a}\right)^2 \lesssim \begin{cases} C_2, \text{ for } j < b, \\ \lfloor \frac{j-b}{a} \rfloor^{1-\delta} + C_1, \text{ for } j > b \end{cases},
$$

and consequently

$$
\sum_m \frac{\omega_m^2 b_m}{(\omega_m + j)^2 (\omega_m - j)^2} \lesssim \begin{cases} j^{-2}, \text{ for } j < b, \\ j^{-\min(1+\delta,2)}, \text{ for } j > b \end{cases} \lesssim j^{-\min(1+\delta,2)}.
$$

Putting everything together yields

$$
\eta_{ij} = (ij)^{-\min\left(1, \frac{1+\delta}{2}\right)}.
$$