2023, Vol. 1, pp. 33–47
doi: 10.30970/ana.2023.1.33

Convergence analysis of kernel conjugate gradient for functional linear regression

N. Gupta[1], S. Sivananthan[1], B.K. Sriperumbudur[2]

[1]*Indian Institute of Technology Delhi, India,*
[2]*Pennsylvania State University, USA*

Анотація. В роботі аналізується збіжність алгоритму на основі спряженого градієнта для функціональної лінійної моделі в системі відтворюючого ядра гільбертового простору, використовуючи результати ранньої зупинки в регуляризації проти надмірної підгонки. Ми встановлюємо швидкості збіжності залежно від умови регулярності функції нахилу та швидкості спадання власних значень операторної композиції коваріації та оператора ядра. Наші швидкості збіжності відповідають мінімаксній швидкості, доступній у літературі.

Abstract. In this paper, we discuss the convergence analysis of the conjugate gradient-based algorithm for the functional linear model in the reproducing kernel Hilbert space framework, utilizing early stopping results in regularization against over-fitting. We establish the convergence rates depending on the regularity condition of the slope function and the decay rate of the eigenvalues of the operator composition of covariance and kernel operator. Our convergence rates match the minimax rate available from the literature.

1 Introduction

The functional linear regression (FLR) model is one of the fundamental tools for analyzing functional data, introduced by Ramsay and Dalzell [22]. The model gained popularity due to its simplicity in dealing with high-dimensional functional data. For example, it is widely used in medicine, chemometrics, and economics [11,12,21,23]. Mathematically, the FLR model is stated as

$$Y = \int_S X(t)\beta^*(t)\,dt + \epsilon,$$

where Y is a real-valued random variable, $(X(t); t \in S)$ is a continuous time process, β^* is an unknown slope function and ϵ is a zero mean random noise, independent of X, with finite variance σ^2. Throughout the paper, we assume that X and β^* are in $L^2(S)$, and S is a compact subset of $\mathbb{R}^d$. In the context of the slope function, it is evident that

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

The goal is to construct an estimator $\hat{\beta}$ to approximate the slope function β^* using observed empirical data $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$, where X_i's are i.i.d. copies of random function X. The main approach in estimating the slope function β^* is based on the representation of the estimator function and functional data in terms of certain basis functions. In this paper, we utilize the framework of reproducing kernel Hilbert space (RKHS) to construct an estimator function $\hat{\beta}$ using the conjugate gradient method.

In [7], the authors used penalized B-spline basis functions to represent the estimator function and also introduced an alternate smooth version of functional principal component analysis (FPCA) to construct $\hat{\beta}$. A Fourier basis approach was explored in [17] and the FPCA-based approach is

2020 *Mathematics Subject Classification*: 62R10, 62G20, 65F22.
Key words: reproducing kernel Hilbert space, conjugate gradient, covariance operator, functional linear regression.

investigated in [5, 14, 19]. One of the profound choices for the basis functions in the FPCA method is to use the eigenfunctions obtained from the covariance operator of the given data. However, Cai and Yuan [6] demonstrated that this choice may not be suitable for all cases, as shown with the example of Canadian weather data. This observation strongly motivated the researchers to explore alternative choices of basis functions.

It is well-known in learning theory that kernel methods represent predictor functions using data-driven kernel functions, resulting in good generalization error (see [3, 9, 10, 20]). Cai and Yuan [26] proposed utilizing the kernel method approach, wherein the estimator is expressed as a linear combination of kernel functions. The method achieves optimal rates under the assumption that the slope function β^* belongs to the RKHS. Later in [6], they used the regularization technique to achieve optimal rates without the Sacks Ylvisaker condition, which was a necessary assumption in [26]. Further analysis of the FLR model within the framework of RKHS has been studied and discussed in [2, 24, 25, 27]. Since the computational complexity of these techniques is $O(n^3)$, they incur high computational costs when dealing with large datasets.

To address this shortcoming, Blanchard and Krämer [4] employed the conjugate gradient method in the kernel ridge regression method, by utilizing an early stopping rule that also serves as a form of regularization. This reduces the computational complexity to $O\left(n^2 m\right)$ for m number of iterations [18]. Inspired by their work, in this paper, we propose an estimator $\hat{\beta}$ for the FLR model by employing the conjugate gradient (CG) approach with an early stopping rule. We specifically focus on the CG method due to its outstanding computational characteristics, setting it apart from other approaches. Since it aggressively targets the reduction of residual errors, it is commonly observed in practical applications that the CG method achieves convergence in significantly fewer iterations compared to other gradient descent techniques, as discussed in the context of kernel learning by [13] and [8]. We obtain a convergence rate for $\|\hat{\beta}-\beta^*\|_{L^2(S)}$ and show it to align with the minimax rates of the FLR model [6, 27], thereby establishing the minimax optimality of our estimator.

The paper is organized as follows. In Section 2, we present the necessary background for the conjugate gradient method for the FLR model in the RKHS setting and explain some important properties of certain orthogonal polynomials. In Section 3, we discuss our assumptions and provide convergence rates of the CG method. We present the supplementary results, which will be used to prove the main theorem, in the final section, Section 4.

2 Preliminaries and Notations

Let $\mathcal{H}$ be a Hilbert space of real-valued functions on a compact subset S of $\mathbb{R}^d$. We say that $\mathcal{H}$ is RKHS if for every $x \in S$, the pointwise evaluation map $f \mapsto f(x)$ is continuous on $\mathcal{H}$. As a consequence of the Riesz representation theorem, there is a unique kernel function $k : S \times S \to \mathbb{R}$, called the *reproducing kernel* such that $k\left(s, \cdot\right) \in \mathcal{H}$ for any $s \in S$ satisfies the reproducing property:

$$f\left(s\right) = \left\langle k(s, \cdot), f\right\rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

It is easy to see that the associated kernel function k is symmetric and positive definite. Conversely, for a given symmetric and positive definite function k, we can construct a unique RKHS with k as the reproducing kernel. For a detailed study of RKHS, we refer the reader to [1].

We assume that k is continuous; then the associated RKHS $\mathcal{H}$ is separable and the embedding operator (inclusion operator) $J : \mathcal{H} \to L^2\left(S\right)$, which is defined as $(Jf)(x) = \langle k(x, \cdot), f\rangle_{\mathcal{H}}$ is compact. The adjoint operator $J^* : L^2\left(S\right) \to \mathcal{H}$ is given by

$$(J^* g)(x) = \int_S k(x, t) g(t)\, dt.$$

We denote the integral operator, $T := JJ^* : L^2\left(S\right) \to L^2\left(S\right)$ and the covariance operator $C := \mathbb{E}\left[X \otimes X\right] : L^2\left(S\right) \to L^2\left(S\right)$, where $\otimes$ is the $L^2\left(S\right)$ tensor product.

Given $(X_i, Y_i)_{i=1}^n$ i.i.d. copies of (X, Y), our estimator is defined as

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \langle \beta, X_i\rangle_{L^2}\right]^2\\ &= \arg\min_{\beta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \langle J\beta, X_i\rangle_{L^2}\right]^2\\ &= \arg\min_{\beta\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \langle \beta, J^*X_i\rangle_{\mathcal{H}}\right]^2.\end{aligned}$$

A solution for this optimization problem can be obtained by solving

$$J^*\hat{C}_nJ\hat{\beta} = J^*\hat{R}, \tag{2.1}$$

where

$$\hat{C}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \otimes X_i \quad \text{and} \quad \hat{R} := \frac{1}{n}\sum_{i=1}^{n} Y_iX_i.$$

We denote $\Lambda := T^{\frac{1}{2}}CT^{\frac{1}{2}}$ and $\Lambda_n := T^{\frac{1}{2}}\hat{C}_nT^{\frac{1}{2}}$.

The fundamental idea behind the conjugate gradient method is to restrict the optimization problem to a set of subspaces (data dependent), known as Krylov subspaces, defined as

$$\begin{aligned}\mathcal{K}_m\left(J^*\hat{R}, J^*\hat{C}_nJ\right) :&= \operatorname{span}\left\{J^*\hat{R}, J^*\hat{C}_nJJ^*\hat{R}, \left(J^*\hat{C}_nJ\right)^2 J^*\hat{R}, \ldots, \left(J^*\hat{C}_nJ\right)^{m-1} J^*\hat{R}\right\}\\ &= \left\{p\left(J^*\hat{C}_nJ\right)J^*\hat{R} \quad : \quad p \in \mathcal{P}_{m-1}\right\},\end{aligned}$$

where $\mathcal{P}_{m-1}$ is a set of real polynomials of degree at most $m-1$. Then the CG solution after m iterations is

$$\hat{\beta}_m = \arg \min_{\beta\in\mathcal{K}_m\left(J^*\hat{R}, J^*\hat{C}_nJ\right)} \left\|J^*\hat{R} - J^*\hat{C}_nJ\beta\right\|_{\mathcal{H}}. \tag{2.2}$$

The iterated solution, because the problem is restricted to the Krylov subspaces, will take the form $\hat{\beta}_m = q_m\left(J^*\hat{C}_nJ\right)J^*\hat{R}$ with q_m being a polynomial of degree at most $m-1$. Associated with each iterated polynomial q_m, we have a residual polynomial defined as $p_m(x) = 1 - xq_m(x) \in \mathcal{P}_m^0$, where $\mathcal{P}_m^0$ is a set of real polynomials of degree at most m and having constant term equal to 1.

Since the construction of the estimator involves forward multiplication through the residual polynomial p_m, it is essential to understand certain fundamental characteristics of these polynomials.

Suppose $(\xi_{n,i}, e_{n,i})_{i\in I}$ is an eigenvalue-eigenfunction pair for the operator $\hat{\Lambda}_n$ with $\xi_{n,i}$ in $[0, \kappa_\Lambda]$, $i \in I$, where $\{e_{n,i} : i \in I\}$ is an orthonormal system in $L^2(S)$ and κ_Λ is a constant that bounds the kernel function of the integral operator Λ. For $u > 0$, denote $F_u = 1_{[0,u)}\left(\hat{\Lambda}_n\right)$ as the orthogonal projector onto the space spanned by $\{e_{n,i} : i \in I,\ \xi_{n,i} < u\}$. For each integer $l \geq 0$, we will introduce measure $\mu_n^{(l)}$ which is defined as

$$\mu_n^{(l)} := \sum_{i\in I} \xi_{n,i}^l \left\langle T^{\frac{1}{2}}\hat{R}, e_{n,i}\right\rangle_{L^2}^2 \delta_{\xi_{n,i}},$$

where δ_x is Dirac measure centered at x. In particular, for $l = 0$ we use the convention $0^0 = 1$.

Associated to each measure $\mu_n^{(l)}$, $l \geq 0$, we define the scalar product of two polynomials as

$$\begin{aligned}[p,q]_{(l)} &:= \int_0^{\kappa_\Lambda} p\left(t\right) q\left(t\right) d\mu_n^{(l)}\left(t\right) \\ &= \left\langle p\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R}, \left(\hat{\Lambda}_n\right)^l q\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} \right\rangle_{L^2} \\ &= \sum_{i\in I} p\left(\xi_{n,i}\right) q\left(\xi_{n,i}\right) \left(\xi_{n,i}\right)^l \left\langle T^{\frac{1}{2}}\hat{R}, e_{n,i}\right\rangle_{L^2}^2 .\end{aligned}$$

For $l = 0$, we see that

$$[p,q]_{(0)} = \left\langle p\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R}, q\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} \right\rangle_{L^2} .$$

Since $\hat{\Lambda}_n$ is a finite rank operator, it has only a finite number of non-zero eigenvalues. Consequently, we observe that the measure $\mu_n^{(l)}$ has finite support of cardinality, independent of l. Indeed, if $\xi_{n,j} = 0$ for some $j \in I$, then we have

$$\begin{aligned} T^{\frac{1}{2}}\hat{C}_n T^{\frac{1}{2}} e_{n,j} = 0 &\implies \left\langle T^{\frac{1}{2}}\hat{C}_n T^{\frac{1}{2}} e_{n,j}, e_{n,j}\right\rangle_{L^2} = 0 \\ &\implies \frac{1}{n}\sum_{i=1}^n \left\langle T^{\frac{1}{2}} X_i, e_{n,j}\right\rangle_{L^2}^2 = 0 \\ &\implies \left\langle T^{\frac{1}{2}} X_i, e_{n,j}\right\rangle_{L^2} = 0, \quad \forall\, 1 \leq i \leq n. \end{aligned}$$

Now we see that

$$\left\langle T^{\frac{1}{2}}\hat{R}, e_{n,j}\right\rangle_{L^2} = \frac{1}{n}\sum_{i=1}^n Y_i \left\langle T^{\frac{1}{2}} X_i, e_{n,j}\right\rangle_{L^2} = 0.$$

This concludes that $\mu_n^{(l)}$ has finite support of cardinality (independent of l), let's say $n_\gamma \leq n$. It is clear from (2.2) that q_m minimizes

$$\left\| J^*\hat{R} - J^*\hat{C}_n J q\left(J^*\hat{C}_n J\right) J^*\hat{R}\right\|_{\mathcal{H}} = \left\| T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{C}_n T^{\frac{1}{2}} q\left(T^{\frac{1}{2}}\hat{C}_n T^{\frac{1}{2}}\right) T^{\frac{1}{2}}\hat{R}\right\|_{L^2}$$

over $q \in \mathcal{P}_{m-1}$. Equivalently, consider $p\left(x\right) = 1 - xq\left(x\right)$, then p_m minimizes

$$\left\| p\left(T^{\frac{1}{2}}\hat{C}_n T^{\frac{1}{2}}\right) T^{\frac{1}{2}}\hat{R}\right\|_{L^2} = [p,p]_{(0)} \tag{2.3}$$

over $p \in \mathcal{P}_m^0$. In other words, we can say that p_m is the orthogonal projection of origin onto the affine subspace $\mathcal{P}_m^0 \subset \mathcal{P}_m$ for the scalar product $[\cdot,\cdot]_{(0)}$. We take $q_0 = 0, p_0 = 1$ for $m = 0$. Because of the properties of projections, p_m is orthogonal to $\mathcal{P}_m^0$. Since $\mathcal{P}_m^0 = 1 + \pi\mathcal{P}_{m-1}$ is parallel to $\pi\mathcal{P}_{m-1}$, where $\left(\pi(q)\right)(x) = xq(x)$ is a shift operator. So we have $0 = [p_m, \pi q]_{(0)} = [p_m, q]_{(1)}$ for any $q \in \mathcal{P}_{m-1}$ which concludes that $p_0, \ldots, p_{n_{\gamma-1}}$ is an orthogonal sequence of polynomials with respect to $[\cdot,\cdot]_{(1)}$. For $m = n_\gamma$, we can see that $[\cdot,\cdot]_{(l)}$ is semidefinite product on $\mathcal{P}_{n_\gamma}$ and p_{n_γ} is unique element of $\mathcal{P}_{n_\gamma}^0$ satisfying $\left[p_{n_\gamma}, p_{n_\gamma}\right]_{(0)} = 0$. Hence, for $m = n_\gamma$, unicity of the solution holds and $\left[p_{n_\gamma}, p_m\right]_{(1)} = 0$ for all $m \leq n_\gamma$. By applying the representer theorem to (2.1),

$$\hat{\beta} \in \operatorname{span}\left\{ \int_S k(\cdot,t) X_i(t)\, dt : i = 1, \ldots, n \right\},$$

i,e., there exists a $\alpha := (\alpha_1, \dots, \alpha_n)^\top \in \mathbb{R}^n$ such that $\hat{\beta} = \sum_{i=1}^n \alpha_i \int_S k(\cdot, t) X_i(t)\, dt$. Using this in (2.1), we can solve $\mathbf{K}\alpha = \mathbf{y}$ to get a solution of (2.1), where

$$\mathbf{K} \in \mathbb{R}^{n\times n} \text{ with } [\mathbf{K}]_{ij:} = \int_S \int_S k(s,t) X_i(t) X_j(s)\, dt\, ds$$

and $\mathbf{y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$. We refer the reader to [15] for the iterative methodology to create the CG estimator of $\hat{\beta}$ in (2.1).

The following lemma, which lists several properties of orthogonal polynomials, is proven in [4]. It will be used frequently throughout the remainder of the paper.

Lemma 2.1. *(Lemma 5.2 in [4]) Let m be any integer satisfying $1 \le m \le n_\gamma$.*

1. *The polynomial p_m has exactly m distinct roots belonging to $(0, \kappa_\Lambda]$, denoted by $(x_{k,m})_{1\le k\le m}$ in increasing order.*
2. *p_m is positive, decreasing and convex on the interval $[0, x_{1,m})$.*
3. *Define the function φ_m on the interval $[0, x_{1,m})$ as*

$$\varphi_m(x) = p_m(x) \left(\frac{x_{1,m}}{x_{1,m} - x} \right)^{\frac{1}{2}}.$$

Then it holds

$$[p_m, p_m]_{(0)}^{\frac{1}{2}} = \left\| p_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}} \hat{R} \right\|_{L^2} \le \left\| F_{x_{1,m}} \varphi_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}} \hat{R} \right\|_{L^2},$$

and furthermore, for any $\nu \ge 0$,

$$\sup_{x\in[0,x_{1,m}]} x^\nu \varphi_m^2(x) \le \nu^\nu \left| p_m'(0) \right|^{-\nu}. \tag{2.4}$$

4. *Denote $p_0^{(2)}, p_1^{(2)}, \dots, p_{n_\gamma-1}^{(2)}$ the unique sequence of orthogonal polynomial with respect to $[\cdot,\cdot]_2$ and with constant term equal to 1. This sequence enjoys properties (1) and (2) above with $\left(x_{k,m}^{(2)}\right)_{1\le k\le m}$ denoting the distinct roots of $p_m^{(2)}$ in increasing order. Then it holds that $x_{1,m} \le x_{1,m}^{(2)}$. Finally, the following holds:*

$$0 \le p_{m-1}'(0) - p_m'(0) = \frac{[p_{m-1}, p_{m-1}]_{(0)} - [p_m, p_{m-1}]_{(0)}}{\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}} \le \frac{[p_{m-1}, p_{m-1}]_{(0)}}{\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}}.$$

3 The Main Result

In this section, we present the convergence rate of the conjugate gradient method in functional linear regression. Our proofs are inspired by the ideas of [4]. The analysis depends on the eigenvalue behaviour of the operator $\Lambda = T^{\frac{1}{2}} C T^{\frac{1}{2}}$ which indicates the behaviour of eigenvalues of the kernel operator T and the covariance operator C. First, we begin with a list of assumptions that are required for our convergence rate analysis.

Assumption 3.1. (*Source Condition*) *There exists $g \in L^2$ such that $\beta^* = T^{\frac{1}{2}} \left(T^{\frac{1}{2}} C T^{\frac{1}{2}} \right)^\alpha g$, where α is any positive real number.*

Note that the assumption implies $\beta^* \in \mathcal{H}$ with additional smoothness. In [27], the authors use this source condition to derive the minimax and faster convergence rate for the Tikhonov regularization with $0 < \alpha \le \frac{1}{2}$.

Assumption 3.2. (*Decay Condition*) *For some $s \in (0,1)$,*

$$i^{-\frac{1}{s}} \lesssim \xi_i \lesssim i^{-\frac{1}{s}} \quad \forall \quad i \in I,$$

where $(\xi_i, e_i)_{i\in I}$ is the eigenvalue-eigenvector pair of operator Λ and the symbol $\lesssim$ means that there exist constants $b, B > 0$ such that $bi^{-\frac{1}{s}} \leq \xi_i \leq Bi^{-\frac{1}{s}}$ for all $i \in I$.

This decay of eigenvalues is related to the effective dimensionality because under this assumption we have that $\mathcal{N}(\lambda) := trace(\Lambda\,(\Lambda + \lambda I)^{-1}) \leq D^2\,(\kappa_\Lambda \lambda)^{-s}$ for all $\lambda \in (0,1]$ and an appropriate choice of $D > 0$.

Assumption 3.3. (*Fourth Moment Condition*) *For any $f \in L^2\,(S)$,*

$$\mathbb{E}\,\langle X, f\rangle_{L^2}^4 \leq c_0 \left(\mathbb{E}\,\langle X, f\rangle_{L^2}^2\right)^2 \text{ for some constant } c_0 > 0.$$

We define the early stopping rule to stop the CG method at an early stage $m^* \ll n$. This is mainly used for its implicit regularization property. The early stopping, defined as m^*, is the first iteration for which the residual term is less than some predefined threshold. Now we state and prove our main theorem.

Theorem 3.1. *Let $\alpha > 0$, $\tau > 0$ and $\mathbb{E}\,\|X\|^4 < \infty$. Suppose Assumptions 3.1–3.3 hold, and stopping rule holds with threshold*

$$\Omega = (2+\tau)\, n^{-\frac{\alpha+1}{1+s+2\alpha}}.$$

Then for large enough n and $\lambda = c\,(\alpha,\delta)\, n^{-\frac{1}{1+s+2\alpha}}$, it holds with probability at least $1-\delta$ that

$$\left\|\hat{\beta}_{m^*} - \beta^*\right\|_{L^2} \lesssim n^{-\frac{\alpha}{1+s+2\alpha}},$$

where $c\,(\alpha,\delta)$ is a constant that depends only on α and δ.

Proof. Let $\lambda \geq \left(\frac{4c_1^2}{n}\right)^{\frac{1}{s+1}}$ and define $F_u^\perp := (I - F_u)$. We start by considering the error term:

$$\begin{aligned}
\left\|\hat{\beta}_m - \beta^*\right\|_{L^2} &= \left\|J\hat{\beta}_m - \beta^*\right\|_{L^2}\\
&= \left\|Jq_m\left(J^*\hat{C}_nJ\right)J^*\hat{R} - \beta^*\right\|_{L^2}\\
&= \left\|T^{\frac{1}{2}}q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\Lambda^\alpha g\right\|_{L^2}\\
&\leq \left\|T^{\frac{1}{2}}\right\|_{op}\left\|q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - \Lambda^\alpha g\right\|_{L^2}\\
&\lesssim \left\|q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - \Lambda^\alpha g\right\|_{L^2}\\
&\leq \left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - \Lambda^\alpha g\right)\right\|_{L^2} + \left\|F_u^\perp\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - \Lambda^\alpha g\right)\right\|_{L^2}\\
&\leq \underbrace{\left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{C}_n\beta^*\right)\right\|_{L^2}}_{Term-1}\\
&\qquad + \underbrace{\left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{C}_n\beta^* - \Lambda^\alpha g\right)\right\|_{L^2}}_{Term-2}\\
&\qquad\quad + \underbrace{\left\|F_u^\perp\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R} - \Lambda^\alpha g\right)\right\|_{L^2}}_{Term-3}.
\end{aligned}$$

In the third step we have used that $Jq_m\left(J^*\hat{C}_nJ\right)J^* = T^{\frac{1}{2}}q_m\left(T^{\frac{1}{2}}\hat{C}_nJ\right)T^{\frac{1}{2}}$ which can be derived easily using spectral representation. Further, we will estimate each term separately using Lemma 4.1 and Lemma 4.2.

Estimation of Term-1:

$$\begin{aligned}
&\left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\left(\hat{R}-\hat{C}_n\beta^*\right)\right)\right\|_{L^2}\\
\leq&\left\|F_uq_m\left(\hat{\Lambda}_n\right)\left(\hat{\Lambda}_n+\lambda I\right)^{\frac{1}{2}}\right\|_{op}\left\|\left(\hat{\Lambda}_n+\lambda I\right)^{-\frac{1}{2}}(\Lambda+\lambda I)^{\frac{1}{2}}\right\|_{op}\\
&\times\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left(\hat{R}-\hat{C}_n\beta^*\right)\right\|_{L^2}\\
\lesssim&\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left\|F_uq_m\left(\hat{\Lambda}_n\right)\left(\hat{\Lambda}_n+\lambda I\right)^{\frac{1}{2}}\right\|_{op}\qquad\text{(by Lemmas 4.1, 4.2)}\\
\leq&\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\sup_{x\in[0,u]}x^{\frac{1}{2}}q_m(x)+\lambda^{\frac{1}{2}}\sup_{x\in[0,u]}q_m(x)\right)\\
\leq&\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\left(\sup_{x\in[0,u]}q_m(x)\right)^{\frac{1}{2}}\left(\sup_{x\in[0,u]}xq_m(x)\right)^{\frac{1}{2}}+\lambda^{\frac{1}{2}}\left|p_m^{'}(0)\right|\right)\\
\leq&\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\left|p_m^{'}(0)\right|^{\frac{1}{2}}+\lambda^{\frac{1}{2}}\left|p_m^{'}(0)\right|\right).
\end{aligned}$$

Following argument ensures the last inequality: If $m=0$, we use $p_m\equiv 1$ and $q_m\equiv 0$, so the inequality follows for any $u>0$. If $m\geq 1$, we use p_m is decreasing and convex in $[0,u]$, $q_m(x)\leq\left|p_m^{'}(0)\right|$ for all $x\in[0,u)$ and $xq_m(x)\leq 1$ for all $x\in[0,u)$ as $u\leq x_{1,m}$.

Estimation of Term-2: Using Lemma 4.4, and the fact that $|p_m(x)|\leq 1$ for all $x\in[0,u)$, we get

$$\begin{aligned}
&\left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{C}_n\beta^*-\Lambda^\alpha g\right)\right\|_{L^2}\\
=&\left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{C}_nT^{\frac{1}{2}}\Lambda^\alpha g-\Lambda^\alpha g\right)\right\|_{L^2}\\
=&\left\|F_u\left(q_m\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n-I\right)\Lambda^\alpha g\right\|_{L^2}=\left\|F_up_m\left(\hat{\Lambda}_n\right)\Lambda^\alpha g\right\|_{L^2}\\
\leq&2\left(\sup_{t\in[0,u]}t^\alpha p_m(t)+\max\{\alpha,1\}Z_\alpha(\lambda)\sup_{t\in[0,u]}p_m(t)\right)\\
\leq&2\left(u^\alpha+\max\{\alpha,1\}Z_\alpha(\lambda)\right).
\end{aligned}$$

Estimation of Term-3:

$$\begin{aligned}
&\left\|F_u^\perp\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R}-\Lambda^\alpha g\right)\right\|_{L^2}\\
=&\left\|F_u^\perp\left(\hat{\Lambda}_n\right)^{-1}\hat{\Lambda}_n\left(q_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R}-\Lambda^\alpha g\right)\right\|_{L^2}\\
\leq&\left\|F_u^\perp\left(\hat{\Lambda}_n\right)^{-1}\left(\hat{\Lambda}_nq_m\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R}-T^{\frac{1}{2}}\hat{R}+T^{\frac{1}{2}}\hat{R}-\hat{\Lambda}_n\Lambda^\alpha g\right)\right\|_{L^2}
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| F_u^{\perp}\left(\hat{\Lambda}_n\right)^{-1}\left(\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right)\right\|_{L^2} + \left\| F_u^{\perp}\left(\hat{\Lambda}_n\right)^{-1}\left(T^{\frac{1}{2}}\hat{R} - \hat{\Lambda}_n \Lambda^{\alpha} g\right)\right\|_{L^2} \\
&\leq \frac{1}{u}\left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2} + \left\| F_u^{\perp}\left(\hat{\Lambda}_n\right)^{-1}\left(\hat{\Lambda}_n + \lambda I\right)^{\frac{1}{2}}\left(\hat{\Lambda}_n + \lambda I\right)^{-\frac{1}{2}}\right. \\
&\qquad\qquad\qquad\qquad \left.\times \left(\Lambda + \lambda I\right)^{\frac{1}{2}}\left(\Lambda + \lambda I\right)^{-\frac{1}{2}}\left(T^{\frac{1}{2}}\hat{R} - \hat{\Lambda}_n \Lambda^{\alpha} g\right)\right\|_{L^2} \\
&\leq \frac{1}{u}\left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2} + \frac{(u+\lambda)^{\frac{1}{2}}}{u}\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}.
\end{aligned}
$$

The last inequality follows from Lemmas 4.1, 4.2.

So by combining all three terms, we get

$$
\begin{aligned}
\left\|\hat{\beta}_m - \beta^*\right\|_{L^2} \lesssim &\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\frac{(\tilde{u}+\lambda)^{\frac{1}{2}}}{\tilde{u}} + \left(u^{\alpha} + \max\{\alpha, 1\} Z_{\alpha}(\lambda)\right) \\
&+ \frac{1}{u}\left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2},
\end{aligned}
$$

where $\tilde{u} = \min\left\{u, \left|p_m^{'}(0)\right|^{-1}\right\}$. Now we define our stopping rule as

$$
m^* = \inf\left\{m \geq 0 \quad : \left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \leq \Omega\right\},
$$

Then

$$
\left\|\hat{\beta}_{m^*} - \beta^*\right\|_{L^2} \lesssim \sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\frac{(\tilde{u}+\lambda)^{\frac{1}{2}}}{\tilde{u}} + \left(u^{\alpha} + \max\{\alpha, 1\} Z_{\alpha}(\lambda)\right) + \frac{1}{u}\Omega. \tag{3.1}
$$

We still have to bound $\left|p_{m^*}^{'}(0)\right|$ and for that we will use Lemma 4.5. First we will bound $\left|p_{m^*-1}^{'}(0)\right|$. We assume that $0 < u < x_{1,m-1} \leq x_{1,m-1}^{(2)}$ (see Lemma 2.1). Consider

$$
\begin{aligned}
&[p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} \\
&= \left\|p_{m-1}\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \\
&\leq \left\|p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \quad (\text{As } p_m \text{ minimizes (2.3) over } p \in \mathcal{P}_m^0) \\
&\leq \left\|F_u p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R}\right\|_{L^2} + \left\|F_u^{\perp} p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \\
&\leq \left\|F_u T^{\frac{1}{2}}\hat{R}\right\|_{L^2} + u^{-\frac{1}{2}}\left\|p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n^{\frac{1}{2}} T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \\
&\leq \left\|F_u\left(T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{C}_n\beta^*\right)\right\|_{L^2} + \left\|F_u T^{\frac{1}{2}}\hat{C}_n\beta^*\right\|_{L^2} + u^{-\frac{1}{2}}\left\|p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n^{\frac{1}{2}} T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \\
&\leq \left\|F_u\left(\hat{\Lambda}_n + \lambda I\right)^{\frac{1}{2}}\right\|\left\|\left(\hat{\Lambda}_n + \lambda I\right)^{-\frac{1}{2}}(\Lambda + \lambda I)^{\frac{1}{2}}\right\|\left\|(\Lambda + \lambda I)^{-\frac{1}{2}}\left(T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{C}_n\beta^*\right)\right\|_{L^2} \\
&\quad + \left\|F_u T^{\frac{1}{2}}\hat{C}_n\beta^*\right\|_{L^2} + u^{-\frac{1}{2}}\left\|p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n^{\frac{1}{2}} T^{\frac{1}{2}}\hat{R}\right\|_{L^2}.
\end{aligned}
$$

For the third last inequality, we have used the fact that $\left|p_{m-1}^{(2)}(x)\right| \leq 1$ for all $x \in \left[0, x_{1,m-1}^{(2)}\right]$ as $p_{m-1}^{(2)}(0) = 1$ and $p_{m-1}^{(2)}$ is non-increasing on $\left[0, x_{1,m-1}^{(2)}\right]$.

From Lemma 4.1, Lemma 4.2 and Lemma 4.4, we get

$$\begin{aligned}
&[p_{m-1},p_{m-1}]_{(0)}^{\frac{1}{2}}\\
=&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left\|F_u\left(\hat{\Lambda}_n+\lambda I\right)^{\frac{1}{2}}\right\|_{op}+\left\|F_u\hat{\Lambda}_n\Lambda^{\alpha}g\right\|_{L^2}+u^{-\frac{1}{2}}\left\|p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n^{\frac{1}{2}}T^{\frac{1}{2}}\hat{R}\right\|_{L^2}\\
\leq&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(u+\lambda\right)^{\frac{1}{2}}+2c(\alpha)\,u\left(u^{\alpha}+Z_{\alpha}(\lambda)\right)\|g\|_{L^2}+u^{-\frac{1}{2}}\left\|p_{m-1}^{(2)}\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n^{\frac{1}{2}}T^{\frac{1}{2}}\hat{R}\right\|_{L^2}\\
=&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(u+\lambda\right)^{\frac{1}{2}}+2c(\alpha)\,u\left(u^{\alpha}+Z_{\alpha}(\lambda)\right)\|g\|_{L^2}+u^{-\frac{1}{2}}\left[p_{m-1}^{(2)},p_{m-1}^{(2)}\right]_{(1)}^{\frac{1}{2}},
\end{aligned}$$

where $c(\alpha)$ is a constant depending on α. Here we have used that $\left|p_{m-1}^{(2)}(x)\right|\leq 1$ for $x\in\left[0,x_{1,m-1}^{(2)}\right]$. Using Assumption 3 and with the choice of $\lambda=c(\alpha,\delta)\,n^{-\frac{1}{1+s+2\alpha}}$, we get that $\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}\lesssim\lambda^{\alpha+\frac{1}{2}}$ and $Z_{\alpha}(\lambda)\leq\lambda^{\alpha}$.

From Lemma 4.5 and the stopping rule $\left\|\hat{\Lambda}_n q_{m^*-1}\left(\hat{\Lambda}_n\right)T^{\frac{1}{2}}\hat{R}-T^{\frac{1}{2}}\hat{R}\right\|_{L^2}>\Omega$, we get

$$\begin{aligned}
&(2+\tau)\,\lambda^{\frac{1}{2}}\lambda^{\frac{1}{2}+\alpha}\\
<&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\left|p_{m^*-1}'(0)\right|^{-\frac{1}{2}}+\lambda^{\frac{1}{2}}\right)\\
&+2\left(\left|p_{m^*-1}'(0)\right|^{-(\alpha+1)}+c(\alpha)\,Z_{\alpha}(\lambda)\left|p_{m^*-1}'(0)\right|^{-1}\right)\|g\|_{L^2}\\
\leq&\sqrt{2}\lambda^{\frac{1}{2}+\alpha}\left(\left|p_{m^*-1}'(0)\right|^{-\frac{1}{2}}+\lambda^{\frac{1}{2}}\right)\\
&+2\left(\left|p_{m^*-1}'(0)\right|^{-(\alpha+1)}+c(\alpha)\,Z_{\alpha}(\lambda)\left|p_{m^*-1}'(0)\right|^{-1}\right)\|g\|_{L^2}.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\tau\lambda^{\frac{1}{2}}\lambda^{\frac{1}{2}+\alpha}\leq c(\alpha)\max\Bigg\{&\lambda^{\frac{1}{2}+\alpha}\left|p_{m^*-1}'(0)\right|^{-\frac{1}{2}},\left|p_{m^*-1}'(0)\right|^{-(\alpha+1)}\|g\|_{L^2},\\
&Z_{\alpha}(\lambda)\left|p_{m^*-1}'(0)\right|^{-1}\|g\|_{L^2}\Bigg\}.
\end{aligned}$$

Following the steps from [4], we see that if the first term attains the maximum

$$\tau\lambda^{\alpha+1}\leq c(\alpha)\,\lambda^{\alpha+\frac{1}{2}}\left|p_{m^*-1}'(0)\right|^{-\frac{1}{2}}\implies\left|p_{m^*-1}'(0)\right|\leq c(\alpha,\tau)\,\lambda^{-1},$$

if the second term attains the maximum

$$\tau\lambda^{\alpha+1}\leq c(\alpha)\left|p_{m^*-1}'(0)\right|^{-(\alpha+1)}\|g\|_{L^2}\implies\left|p_{m^*-1}'(0)\right|\leq c'(\alpha,\tau)\,\lambda^{-1},$$

if the third term attains the maximum

$$\tau\lambda^{\alpha+1}\leq c(\alpha)\,Z_{\alpha}(\lambda)\left|p_{m^*-1}'(0)\right|^{-1}\|g\|_{L^2}\implies\left|p_{m^*-1}'(0)\right|\leq c''(\alpha,\tau)\,\lambda^{-1},$$

as $Z_{\alpha}(\lambda)\lesssim\lambda^{\alpha}$.

From all three cases, we will get that

$$\left|p_{m^*-1}^{'}(0)\right| \leq c_2(\alpha,\tau)\,\lambda^{-1},$$

for some constant $c_2(\alpha,\tau) > 0$.

In the next step we get an bound on $\left|p_{m^*}^{'}(0)\right|$. From Lemma 2.1, we know that

$$\left|p_{m-1}^{'}(0) - p_m^{'}(0)\right| \leq \frac{[p_{m-1}, p_{m-1}]_{(0)}}{\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}}.$$

Define $u := a(\alpha,\tau)\,\lambda$, where $a(\alpha,\tau)$ is a constant. Using this choice we get

$$\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\,(u+\lambda)^{\frac{1}{2}} + 2c(\alpha)\,u\,(u^\alpha + Z_\alpha(\lambda))\,\|g\|_{L^2} \leq c_3(\alpha,\tau)\,\lambda^{\frac{1}{2}}\lambda^{\alpha+\frac{1}{2}}.$$

From the stopping rule, we know that

$$\left\|\hat{\Lambda}_n q_{m^*-1}\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2} > \Omega = (2+\tau)\,\lambda^{\frac{1}{2}}\lambda^{\alpha+\frac{1}{2}}.$$

So now we can see that using all these inequalities we have that

$$\left|p_{m^*}^{'}(0)\right| \leq c_4(\alpha,\tau)\,\lambda^{-1}. \tag{3.2}$$

Using (3.2), the choice of $a(\alpha,\tau)$ can be made accordingly such that

$$u \leq \left|p_{m^*}^{'}(0)\right|^{-1} \leq x_{1,m}.$$

With this inequality, we can choose $\tilde{u} = u$. Now with this choice of $\tilde{u} = a(\alpha,\tau)\,\lambda$, where $a(\alpha,\tau)$ has been taken to satisfy all the conditions on $\tilde{u}$, we will further bound (3.1). Therefore, we get that

$$\begin{aligned}\left\|\hat{\beta}_{m^*} - \beta^*\right\|_{L^2} \lesssim & \sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\,\tilde{u}^{-1}(\lambda+\tilde{u})^{\frac{1}{2}} + 2\left(u^\alpha + \max\{\alpha,1\}\,Z_\alpha(\lambda)\right) \\ & + \frac{1}{u}\left\|\hat{\Lambda}_n q_{m^*}\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2} \leq \lambda^\alpha.\end{aligned}$$

Now by using $\lambda = c(\alpha,\delta)\,n^{-\frac{\alpha+1}{1+s+2\alpha}}$, the result follows. □

In the RKHS framework, the first minimax convergence rate was established by Cai and Yuan [6, 26] for the Tikhonov regularization method with the source condition $\beta^* \in \mathcal{H}$. Later in [27], the results were extended, and the minimax convergence rates derived with the source condition $\beta^* \in \mathcal{R}\left(T^{\frac{1}{2}}\left(T^{\frac{1}{2}}CT^{\frac{1}{2}}\right)^{\alpha}\right)$ for $0 < \alpha \leq \frac{1}{2}$. Our convergence rates for the conjugate gradient method match the minimax rates of [27] and also match with the convergence of rates of [13].

4 Supplementary Results

Technical details in the paper depend on the estimation of the residual term as the CG method works to reduce the residual term as the number of iterations progresses. To simplify the technical part we use the fact that $\left\|J^*\hat{C}_n J\hat{\beta}_m - J^*\hat{R}\right\|_{\mathcal{H}} = \left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right) T^{\frac{1}{2}}\hat{R} - T^{\frac{1}{2}}\hat{R}\right\|_{L^2}$. We introduce several lemmas that aid us to estimate the residual term and to prove the main theorem.

Lemma 4.1. *Under Assumptions 3.2 and 3.3 we get*

$$\left\|\left(\hat{\Lambda}_n-\Lambda\right)(\Lambda+\lambda I)^{-1}\right\|_{op}\leq c_1\left(n\lambda^{1+s}\right)^{-\frac{1}{2}}$$

for some constant $c_1>0$.

We skip the proof of this lemma as it follows from the similar steps of Lemma 2 in [6].

From Lemma 4.1 and for $\lambda\geq\left(\frac{4c_1^2}{n}\right)^{\frac{1}{s+1}}$, we get

$$\left\|\left(\hat{\Lambda}_n-\Lambda\right)(\Lambda+\lambda I)^{-1}\right\|_{op}\leq\frac{1}{2}.$$

As a consequence, we get

$$\begin{aligned}\left\|(\Lambda+\lambda I)\left(\hat{\Lambda}_n+\lambda I\right)^{-1}\right\|_{op}&=\left\|\left[\left(\hat{\Lambda}_n-\Lambda\right)(\Lambda+\lambda I)^{-1}+I\right]^{-1}\right\|_{op}\\&\leq\frac{1}{1-\left\|\left(\hat{\Lambda}_n-\Lambda\right)(\Lambda+\lambda I)^{-1}\right\|_{op}}\\&\leq 2,\quad\forall\lambda\geq\left(\frac{4c_1^2}{n}\right)^{\frac{1}{s+1}}.\end{aligned}$$

Using Corde's inequality, $\left(\|A^\nu B^\nu\|_{op}\leq\|AB\|_{op}^\nu,\ 0\leq\nu\leq 1\right)$, where A and B are self-adjoint positive operators, we get

$$\left\|(\Lambda+\lambda I)^\nu\left(\hat{\Lambda}_n+\lambda I\right)^{-\nu}\right\|_{op}\leq 2^\nu,\quad\forall\nu\in S,\lambda\geq\left(\frac{4c_1^2}{n}\right)^{\frac{1}{s+1}}.\tag{4.1}$$

We will follow the similar ideas from [2, 25] to prove the following lemma.

Lemma 4.2. *For* $\delta>0$, *with at least probability* $1-\delta$, *we have that*

$$\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left(\hat{R}-\hat{C}_n\beta^*\right)\right\|_{L^2}\leq\sqrt{\frac{\sigma^2\mathcal{N}(\lambda)}{n\delta}}.$$

Proof. Define $Z_i:=(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left[Y_iX_i-(X_i\otimes X_i)\beta^*\right]$. Since the slope function β^* satisfies the operator equation $C\beta^*=\mathbb{E}[YX]$, we get that the mean of random variable Z_i is zero, i.e.,

$$\mathbb{E}[Z_i]=(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left[\mathbb{E}[YX]-C\beta^*\right]=0.$$

By Markov's inequality, for any $t>0$

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

Note that

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

and by taking $t=\sqrt{\frac{\mathbb{E}\|Z_1\|_{L^2}^2}{n\delta}}$, with at least probability $1-\delta$, we have

$$\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left(\hat{R}-\hat{C}_n\beta^*\right)\right\|_{L^2}\leq\sqrt{\frac{\mathbb{E}\left[\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left(YX-(X\otimes X)\beta^*\right)\right\|_{L^2}^2\right]}{n\delta}}.\tag{4.2}$$

Consider

$$\begin{aligned}
&\mathbb{E}\left[\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left(YX-(X\otimes X)\beta^*\right)\right\|_{L^2}^2\right]\\
=&\mathbb{E}\left[\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}\left(Y-\langle X,\beta^*\rangle_{L^2}\right)X\right\|_{L^2}^2\right]\\
=&\mathbb{E}\left[\left(Y-\langle X,\beta^*\rangle_{L^2}\right)^2\left\|(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}X\right\|_{L^2}^2\right]\\
=&\mathbb{E}\left[\epsilon^2\left\langle(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}X,(\Lambda+\lambda I)^{-\frac{1}{2}}T^{\frac{1}{2}}X\right\rangle_{L^2}\right]\\
=&\mathbb{E}\left[\epsilon^2\text{trace}\left((\Lambda+\lambda I)^{-1}T^{\frac{1}{2}}(X\otimes X)T^{\frac{1}{2}}\right)\right]\\
=&\mathbb{E}\left[\epsilon^2\right]\text{trace}\left((\Lambda+\lambda I)^{-1}T^{\frac{1}{2}}CT^{\frac{1}{2}}\right)\\
=&\sigma^2\text{trace}\left((\Lambda+\lambda I)^{-1}\Lambda\right)=\sigma^2\mathcal{N}(\lambda).
\end{aligned}$$

With this bound and (4.2), the result follows. □

Lemma 4.3. *Assuming* $\mathbb{E}\|X\|^4<\infty$, *with at least probability* $1-\delta$, *we get*

$$\left\|\hat{C}_n-C\right\|_{HS}\le\sqrt{\frac{\mathbb{E}\|X\|^4}{n\delta}}:=\Delta.$$

Proof. From Chebyshev's inequality, we have that

$$\mathbb{P}\left(\left\|\hat{C}_n-C\right\|_{HS}>\xi\right)\le\frac{\mathbb{E}\left\|\hat{C}_n-C\right\|_{HS}^2}{\xi^2}.$$

Using Theorem 2.5 [16], we get

$$\mathbb{P}\left(\left\|\hat{C}_n-C\right\|_{HS}>\xi\right)\le\frac{\mathbb{E}\|X\|^4}{n\xi^2}.$$

Taking $\xi=\sqrt{\frac{\mathbb{E}\|X\|^4}{n\delta}}$ will conclude the result. □

The bound in the Hilbert-Schmidt is stronger than the operator norm. So the above lemma provides a stronger estimation compared to the estimation in terms of operator norm. The next lemma explains a technical bound involving operator Λ and $\hat{\Lambda}_n$ which will be used repeatedly in our analysis.

Lemma 4.4. *For any* $\nu>0$, *and measurable* $\varphi:[0,\kappa_\Lambda]\to\mathbb{R}$, *it holds with probability greater than* $1-e^{\xi}$ *that*

$$\left\|\varphi\left(\hat{\Lambda}_n\right)\Lambda^\nu\right\|_{op}\lesssim\sup_{t\in[0,\kappa_\Lambda]}t^\nu\varphi(t)+\max\{\nu,1\}Z_\nu(\lambda)\sup_{t\in[0,\kappa_\Lambda]}\varphi(t),$$

where

$$Z_\nu(\lambda)=\begin{cases}\lambda^\nu, & \textit{if } \nu\le 1\\ \kappa_\Lambda^\nu\Delta, & \textit{if } \nu>1\end{cases}.$$

Proof. Proof of this result follows the similar steps of Lemma 5.3 [4]. For $\nu\le 1$, we have

$$\begin{aligned}
\left\|\varphi\left(\hat{\Lambda}_n\right)\Lambda^\nu\right\|_{op}&\le\left\|\varphi\left(\hat{\Lambda}_n\right)\left(\hat{\Lambda}_n+\lambda I\right)^\nu\right\|_{op}\left\|\left(\hat{\Lambda}_n+\lambda I\right)^{-\nu}(\Lambda+\lambda I)^\nu\right\|_{op}\left\|(\Lambda+\lambda I)^{-\nu}\Lambda^\nu\right\|_{op}\\
&\lesssim\left(\sup_{t\in[0,\kappa_\Lambda]}t^\nu\varphi(t)+\lambda^\nu\sup_{t\in[0,\kappa_\Lambda]}\varphi(t)\right).
\end{aligned}$$

For last inequality, we used (4.1) and the fact that $\left\|(\Lambda+\lambda I)^{-\nu}\Lambda^{\nu}\right\|_{op}\le 1$. For $\nu>1$,

$$\begin{aligned}\left\|\varphi\left(\hat{\Lambda}_n\right)\Lambda^{\nu}\right\|_{op} &\le \left\|\varphi\left(\hat{\Lambda}_n\right)\right\|_{op}\left\|\left(\Lambda^{\nu}-\hat{\Lambda}_n^{\nu}\right)\right\|_{op}+\left\|\varphi\left(\hat{\Lambda}_n\right)\hat{\Lambda}_n^{\nu}\right\|_{op}\\ &\le \left\|\left(\Lambda^{\nu}-\hat{\Lambda}_n^{\nu}\right)\right\|_{op}\sup_{t\in[0,\kappa_\Lambda]}\varphi(t)+\sup_{t\in[0,\kappa_\Lambda]}t^{\nu}\varphi(t)\\ &\lesssim \left\|\Lambda-\hat{\Lambda}_n\right\|_{HS}\sup_{t\in[0,\kappa_\Lambda]}\varphi(t)+\sup_{t\in[0,\kappa_\Lambda]}t^{\nu}\varphi(t).\end{aligned}$$

Here we used that for $\nu>1$, x^{ν} is $\nu\kappa_{\Lambda}^{\nu-1}-$ Lipschitz over $[0,\kappa_\Lambda]$ and by Lemma 4.3 we get our result. □

In the following lemma, we will discuss the bound of the residual term that will be used later to bound $\left|p_m^{'}(0)\right|$.

Lemma 4.5. *Under Assumptions 3.1–3.3, $\mathbb{E}\|X\|^4<\infty$ and $\lambda\ge\left(\frac{4c_1^2}{n}\right)^{\frac{1}{s+1}}$, we have that*

$$\begin{aligned}\left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right)T^{\frac12}\hat{R}-T^{\frac12}\hat{R}\right\|_{L^2}\le&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\left|p_m^{'}(0)\right|^{-\frac12}+\lambda^{\frac12}\right)\\ &+2\left(\left|p_m^{'}(0)\right|^{-(\alpha+1)}+c(\alpha)Z_\alpha(\lambda)\left|p_m^{'}(0)\right|^{-1}\right)\|g\|_{L^2}.\end{aligned}$$

Proof. For bounding the residual term, we use Lemma 2.1 at the initial stage to conclude that

$$\begin{aligned}&\left\|\hat{\Lambda}_n q_m\left(\hat{\Lambda}_n\right)T^{\frac12}\hat{R}-T^{\frac12}\hat{R}\right\|_{L^2}\\ =&\left\|p_m\left(\hat{\Lambda}_n\right)T^{\frac12}\hat{R}\right\|_{L^2}\\ \le&\left\|F_{x_{1,m}}\left(\varphi_m\left(\hat{\Lambda}_n\right)T^{\frac12}\hat{R}\right)\right\|_{L^2}\\ =&\left\|F_{x_{1,m}}\left(\varphi_m\left(\hat{\Lambda}_n\right)\left(T^{\frac12}\hat{R}-T^{\frac12}\hat{C}_n\beta^*+T^{\frac12}\hat{C}_n\beta^*\right)\right)\right\|_{L^2}\\ \le&\underbrace{\left\|F_{x_{1,m}}\left(\varphi_m\left(\hat{\Lambda}_n\right)T^{\frac12}\left(\hat{R}-\hat{C}_n\beta^*\right)\right)\right\|_{L^2}}_{Term-A}+\underbrace{\left\|F_{x_{1,m}}\left(\varphi_m\left(\hat{\Lambda}_n\right)T^{\frac12}\hat{C}_n\beta^*\right)\right\|_{L^2}}_{Term-B}.\end{aligned}$$

We will take help from Lemma 4.1 and Lemma 4.2 for the estimation of both terms.

Estimation of Term-A:

$$\begin{aligned}&\left\|F_{x_{1,m}}\left(\varphi_m\left(\hat{\Lambda}_n\right)T^{\frac12}\left(\hat{R}-\hat{C}_n\beta^*\right)\right)\right\|_{L^2}\\ \le&\left\|F_{x_{1,m}}\varphi_m\left(\hat{\Lambda}_n\right)\left(\hat{\Lambda}_n+\lambda I\right)^{\frac12}\right\|_{op}\\ &\times\left\|\left(\hat{\Lambda}_n+\lambda I\right)^{-\frac12}(\Lambda+\lambda I)^{\frac12}\right\|_{op}\left\|(\Lambda+\lambda I)^{-\frac12}\left(T^{\frac12}\hat{R}-T^{\frac12}\hat{C}_n\beta^*\right)\right\|_{L^2}\\ \le&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left\|F_{x_{1,m}}\varphi_m\left(\hat{\Lambda}_n\right)\left(\hat{\Lambda}_n+\lambda I\right)^{\frac12}\right\|_{op}\\ \le&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\sup_{x\in[0,x_{1,m}]}x^{\frac12}\varphi_m(x)+\lambda^{\frac12}\sup_{x\in[0,x_{1,m}]}\varphi_m(x)\right)\\ \le&\sqrt{\frac{2\sigma^2\mathcal{N}(\lambda)}{n\delta}}\left(\left|p_m^{'}(0)\right|^{-\frac12}+\lambda^{\frac12}\right).\end{aligned}\tag{4.3}$$

For the last inequality, we use (2.4) for $\nu = 0, 1$.

Estimation of Term-B:

$$\begin{aligned}
&\left\| F_{x_{1,m}} \left(\varphi_m \left(\hat{\Lambda}_n \right) T^{\frac{1}{2}} \hat{C}_n \beta^* \right) \right\|_{L^2} \\
=&\left\| F_{x_{1,m}} \varphi_m \left(\hat{\Lambda}_n \right) \hat{\Lambda}_n \Lambda^{\alpha} \right\|_{op} \|g\|_{L^2} \\
\leq&2 \left(\sup_{t \in [0, x_{1,m}]} t^{\alpha+1} \varphi_m (t) + c(\alpha) Z_\alpha (\lambda) \sup_{t \in [0, x_{1,m}]} t \varphi_m (t) \right) \|g\|_{L^2} \\
\leq&2 \left(\left| p_m^{'} (0) \right|^{-(\alpha+1)} + c(\alpha) Z_\alpha (\lambda) \left| p_m^{'} (0) \right|^{-1} \right) \|g\|_{L^2} .
\end{aligned} \tag{4.4}$$

where we used Lemma 4.4 and (2.4) in the last inequality. From (4.3) and (4.4), we obtain the result. □

Acknowledgments

The authors thank the referee for his valuable comments and suggestions. S. Sivananthan acknowledges the Science and Engineering Research Board, Government of India, for the financial support through project no. MTR/2022/000383. BKS is partially supported by the National Science Foundation DMS CAREER Award 1945396.

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337-404 (1950)
2. Balasubramanian, K., Müller, H.-G., Sriperumbudur, B.K.: Unified RKHS methodology and analysis for functional linear and single-index models, arXiv preprint. (2022). arXiv: 2206.03975
3. Bauer, F., Pereverzev, S., Rosasco, L.: On regularization algorithms in learning theory. J. Complexity. **23** (1), 52-72 (2007)
4. Blanchard, G., Krämer, N.: Convergence rates of kernel conjugate gradient for random design regression. Anal. Appl. Singap. **14** (6), 763-794 (2016)
5. Cai, T.T., Hall, P.: Prediction in functional linear regression. Ann. Statist. **34** (5), 2159-2179 (2006)
6. Cai, T.T., Yuan, M.: Minimax and adaptive prediction for functional linear regression. J. Amer. Statist. Assoc. **107** (499), 1201-1216 (2012)
7. Cardot, H., Ferraty, F., Sarda, P.: Spline estimators for the functional linear model. Statist. Sinica. **13** (3), 571-591 (2003)
8. Chen, X., Tang, B., Fan, J., Guo, X.: Online gradient descent algorithms for functional data learning. J. Complexity. **70**, Paper No 101635 (14) (2022)
9. Cucker, F., Smale, S.: On the mathematical foundations of learning. Bull. Amer. Math. Soc. N.S. **39** (1), 1-49 (2002)
10. Cucker, F., Zhou, D.-X.: Learning theory: an approximation theory viewpoint, vol. 24. Cambridge University Press, Cambridge, (2007)
11. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Theory and practice. Springer Series in Statistics. Springer, New York (2006)
12. Forni, M., Reichlin, L.: Let's get real: A factor analytical approach to disaggregated business cycle dynamics. The Review of Economic Studies. **65** (3), 453-473 (1998)
13. Guo, X., Guo, Z.-C., Shi, L.: Capacity dependent analysis for functional online learning algorithms. Appl. Comput. Harmon. Anal. **67**, Paper No 101567 (30) (2023)

14. Hall, P., Horowitz, J.L.: Methodology and convergence rates for functional linear regression. Ann. Statist. **35** (1), 70-91 (2007)
15. Hanke, M.: Conjugate Gradient Type Methods for Ill-Posed Problems, vol. 327 of Pitman Research Notes in Mathematics Series. Longman Scientific & Technical, Harlow, (1995)
16. Horváth, L., Kokoszka, P.: Inference for functional data with applications. Springer Series in Statistics. Springer, New York (2012)
17. Li, Y., Hsing, T.: On rates of convergence in functional linear regression. J. Multivariate Anal. **98** (9), 1782-1804 (2007)
18. Lin, J., Cevher, V.: Kernel conjugate gradient methods with random projections. Appl. Comput. Harmon. Anal. **55**, 223-269 (2021)
19. Müller, H.-G., Stadtmüller, U.: Generalized functional linear models. Ann. Statist. **33** (2), 774-805 (2005)
20. Pereverzyev, S.: An introduction to artificial intelligence based on reproducing kernel Hilbert spaces. Compact Textbooks in Mathematics, Birkhäuser (2022)
21. Preda, C., Saporta, G.: Clusterwise PLS regression on a stochastic process. Comput. Statist. Data Anal. **49** (1), 99-108 (2005)
22. Ramsay, J.O., Dalzell, C.J.: Some tools for functional data analysis. J. Roy. Statist. Soc. Ser. B **53** (3), 539-572 (1991)
23. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, second ed. Springer Series in Statistics. Springer, New York (2005)
24. Tong, H.: Functional linear regression with Huber loss. J. Complexity **74**, Paper No 101696 (14) (2023)
25. Tong, H., Ng, M.: Analysis of regularized least squares for functional linear regression model. J. Complexity **49**, 85-94 (2018)
26. Yuan, M., Cai, T.T.: A reproducing kernel Hilbert space approach to functional linear regression. Ann. Statist. **38** (6), 3412-3444 (2010)
27. Zhang, F., Zhang, W., Li, R., Lian, H.: Faster convergence rate for functional linear regression in reproducing kernel Hilbert spaces. Statistics. **54** (1), 167-181 (2020)

Received 04.10.2023
Revised 22.10.2023