Analysis of regularized least squares for functional linear regression model

Hongzhi Tong, Michael Ng

PII: S0885-064X(18)30063-3
DOI: https://doi.org/10.1016/j.jco.2018.08.001
Reference: YJCOM 1374

To appear in: *Journal of Complexity*

Received date : 14 July 2018
Revised date : 28 July 2018
Accepted date : 4 August 2018

Analysis of Regularized Least Squares for Functional Linear Regression Model

Hongzhi Tong[*] Michael Ng[†]

Abstract

In this paper, we study and analyze the regularized least squares for functional linear regression model. The approach is to use the reproducing kernel Hilbert space framework and the integral operators. We show with a more general and realistic assumption on the reproducing kernel and input data statistics that the rate of excess prediction risk by the regularized least squares is minimax optimal.

Key words: Regularized least squares, Functional linear regression, Reproducing kernel Hilbert space, Learning rate

AMS classification: 60K35, 62J05

1 Introduction

There are increasing cases in practice where the data are collected in the form of random functions or curves. This type of data is becoming more prevalent throughout science, engineering and financial market, as automated on-line data collection facilities are becoming more ubiquitous. Many classical statistical tools and models for multivariate analysis, such as principal components analysis, canonical correlation analysis and linear model are then extended to the infinite-dimensional functional domain. In this paper, we consider the functional linear model

$$Y = \alpha_0 + \int_{\mathbb{I}} X(t)\beta_0(t)dt + \epsilon. \tag{1}$$

where Y is a scalar response, $X : \mathbb{I} \to \mathbb{R}$ is a square integrable functional predictor defined over compact domain $\mathbb{I} \subset \mathbb{R}$, α_0 is the intercept, $\beta_0 : \mathbb{I} \to \mathbb{R}$ is the slope function,

[*] School of Statistics, University of International Business and Economics, Beijing 100029, P. R. China.

[†] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. Research supported in part by HKRGC GRF 12302715, 12306616, 12200317, and HKBU RC-ICRS/16-17/03.

and ϵ is the random noise with mean 0 and finite variance σ^2. Functional linear model was introduced by J.O. Ramsay and C.J. Dalzell [11] and first written in its commonly encountered form (1) by T. Hastie and C. Mallows [7]. Some recent research on the statistical analysis of (1) includes [2, 3, 6, 15, 16]. In this paper, we focus on the random design where X is a path of a square integrable stochastic process defined over $\mathbb{I}$ and is independent of ϵ. Without loss of much generality, throughout the paper we assume $\mathbb{E}(X) = 0$ and the intercept $\alpha_0 = 0$, since the intercept can be easily estimated.

Let $\mathcal{L}^2$ be the Hilbert space of square integrable functions on $\mathbb{I}$ (with respect to Lebesgue measure) with standard inner production $< u, v >= \int_{\mathbb{I}} u(s)v(s)ds$ and norm $\|u\| = \left(\int_{\mathbb{I}} u^2(s)ds\right)^{1/2}$. The goal of prediction is to recover the functional η_0:

$$\eta_0(X) = \int_{\mathbb{I}} X(t)\beta_0(t)dt =< \eta_0, X >$$

based on a training sample $\{(X_i, Y_i) : i = 1, \cdots, n\}$ consisting of n independent copies of (X, Y). Define the risk for a prediction η as

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

where (X^*, Y^*) is a copy of (X, Y) independent of the training data, and $\mathbb{E}^*$ represents expectations taken over X^* and Y^* only. Let $\hat{\eta}$ be a prediction rule constructed from the training data. Then, its accuracy can be naturally measured by the excess risk:

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

In this paper, we study the prediction problem in the reproducing kernel Hilbert space (RKHS) framework under which the unknown slope function β_0 is assumed to reside in an RKHS $\mathcal{H}_K$ with a reproducing kernel K [9]. A reproducing kernel $K : \mathbb{I} \times \mathbb{I} \to \mathbb{R}$ is a real, symmetric, continuous, and nonnegative definite function. There is a one-to-one correspondence between a reproducing kernel K and an RKHS $\mathcal{H}_K$ which is a linear functional space endowed with an inner product $< \cdot, \cdot >_K$ such that for any $t \in \mathbb{I}, K(t, \cdot) \in \mathcal{H}_K$, and

$$f(t) =< K(t, \cdot), f >_K, \quad \forall f \in \mathcal{H}_K.$$

We then estimate β_0 via the following regularized least square scheme

$$\beta_{n,\lambda} = \arg \min_{\beta \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \int_{\mathbb{I}} \beta(t)X_i(t)dt]^2 + \lambda \|\beta\|_K^2 \right\}, \tag{2}$$

where $\lambda > 0$ is a regularization parameter. Given the estimate $\beta_{n,\lambda}$, the prediction $\eta_{n,\lambda}$ is obtained by

$$\eta_{n,\lambda}(X) = \int_{\mathbb{I}} X(t)\beta_{n,\lambda}(t)dt.$$

We consider the integral operator associated with kernel K as

$$\mathtt{L}_K(f)(t) = \int_{\mathbb{I}} K(s,t)f(s)ds$$

for $f \in \mathcal{L}^2$. Because of non-negative function of K, the square root operator $\mathtt{L}_K^{1/2}$ of $\mathtt{L}_K$ can be constructed, see Section 2. The covariance function of X is also a symmetric, square integrable, and non-negative definite function defined on $\mathbb{I} \times \mathbb{I}$ as

$$C(s,t) = \mathbb{E}(X(s)X(t)).$$

Hence, the integral operator

$$\mathtt{L}_C(f)(t) = \int_{\mathbb{I}} C(s,t)f(s)ds$$

is well defined and non-negative definite. The main contribution of this paper is to show that if

$$\kappa^2 := ess\sup \|\mathtt{L}_K^{1/2} X\|^2 < \infty, \tag{3}$$

and

$$Tr((\mathtt{L}_K^{1/2}\mathtt{L}_C\mathtt{L}_K^{1/2} + \lambda\mathtt{I})^{-1}\mathtt{L}_K^{1/2}\mathtt{L}_C\mathtt{L}_K^{1/2}) \leq c\lambda^{-\theta}, \quad \forall\lambda > 0. \tag{4}$$

where $0 < \theta \leq 1$ and c is a positive constant, then we have for any $0 < \delta < 1$, we have with confidence at least $1 - \delta$,

$$\mathcal{E}(\eta_{n,\lambda}) - \mathcal{E}(\eta_0) \leq \frac{C(\log\frac{4}{\delta})^4}{\delta^2} n^{\frac{-1}{1+\theta}}, \tag{5}$$

where C is a positive constant. Condition (3) coincides that the range of $\mathtt{L}_K^{1/2}$ on the sample paths of the stochastic process forms a bounded subset of $\mathcal{L}^2$ almost surely. The trace of the operator $(\mathtt{L}_K^{1/2}\mathtt{L}_C\mathtt{L}_K^{1/2} + \lambda\mathtt{I})^{-1}\mathtt{L}_K^{1/2}\mathtt{L}_C\mathtt{L}_K^{1/2}$ is called effective dimension in learning theory (see [17]). Condition (4) reflects the convergence of the eigenvalues of $\mathtt{L}_K$ and $\mathtt{L}_C$, as well as how their eigenfunctions align each other.

The outline of this paper is given as follows. In Section 2, we review basic preliminaries. In Section 3, we show our main minimax optimal results and explain the conditions we assume here. Finally, some concluding remarks are given in Section 4.

2 The Regularization Model

2.1 Preliminaries

For a real, symmetric, square integrable, and nonnegative definite function $R : \mathbb{I}\times\mathbb{I} \to \mathbb{R}$, we define an integral operator $\mathtt{L}_R : \mathcal{L}^2 \to \mathcal{L}^2$ as

$$\mathtt{L}_R(f)(t) = \int_{\mathbb{I}} R(s,t)f(s)ds.$$

It is well-known (see Proposition 11.20 in [9]) that $\mathtt{L}_R$ is a Hilbert-Schmidt operator on $\mathcal{L}^2$, and thus is compact. The spectral theorem implies that there exists a set of orthonormalized eigenfunctions $\{\psi_k^R : k \geq 1\}$ and a sequence of eigenvalues $\theta_1^R \geq \theta_2^R \geq \cdots \geq 0$ such that

$$R(s,t) = \sum_{k\geq 1} \theta_k^R \psi_k^R(s)\psi_k^R(t), \quad \forall s,t \in \mathbb{I},$$

and

$$\mathtt{L}_R(\psi_k^R) = \theta_k^R \psi_k^R, \quad k = 1, 2, \cdots .$$

Then the square root operator of $\mathtt{L}_R$ is defined by

$$\mathtt{L}_R^{1/2}(\psi_k^R) = \mathtt{L}_{R^{1/2}}(\psi_k^R) = \sqrt{\theta_k^R}\psi_k^R,$$

where

$$R^{1/2}(s,t) = \sum_{k\geq 1} \sqrt{\theta_k^R}\psi_k^R(s)\psi_k^R(t), \quad \forall s,t \in \mathbb{I}.$$

Define

$$(R_1R_2)(s,t) = \int_{\mathbb{I}} R_1(s,u)R_2(u,t)du.$$

Then $\mathtt{L}_{R_1}\mathtt{L}_{R_2} = \mathtt{L}_{R_1R_2}$.

2.2 The Formulation

For brevity, we shall write

$$\mathtt{T} = \mathtt{L}_K^{1/2}\mathtt{L}_C\mathtt{L}_K^{1/2}$$

Let τ_k and φ_k be the eigenvalues and eigenfunctions of $\mathtt{T}$ as a compact operator on $\mathcal{L}^2$. Then $\{\varphi_k : k \geq 1\}$ form an orthonormal basis of $\mathcal{L}^2$ and

$$\mathtt{T}(f) = \sum_{k\geq 1} \tau_k < f, \varphi_k > \varphi_k, \quad \forall f \in \mathcal{L}^2.$$

Define the empirical covariance function as

$$C_n(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t)$$

and

$$\mathtt{T}_n = \mathtt{L}_K^{1/2}\mathtt{L}_{C_n}\mathtt{L}_K^{1/2},$$

where $\mathtt{L}_{C_n}$ is an integral operator such that for any $f \in \mathcal{L}^2$,

$$\mathtt{L}_{C_n} f(t) = \int_{\mathbb{I}} C_n(s,t)f(s)ds.$$

Recall that $\mathtt{L}_K^{1/2}(\mathcal{L}^2) = \mathcal{H}_K$. Therefore, there exist $f_0, f_{n,\lambda} \in \mathcal{L}^2$ such that

$$\beta_0 = \mathtt{L}_K^{1/2} f_0 \quad \text{and} \quad \beta_{n,\lambda} = \mathtt{L}_K^{1/2} f_{n,\lambda}.$$

In the following, we assume that $\mathcal{H}_K$ is dense in $\mathcal{L}^2$, which ensures that f_0 and $f_{n,\lambda}$ are uniquely defined. This assumption can be satisfied when K is a universal kernel, such as Gaussian kernel, see [14, 13].

Now (2) can be rewritten as

$$f_{n,\lambda} = \arg\min_{f \in \mathcal{L}^2} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - < X_i, \mathtt{L}_K^{1/2} f >)^2 + \lambda \|f\|^2 \right\}.$$

It is not hard to see that

$$f_{n,\lambda} = (\mathtt{T}_n + \lambda \mathtt{I})^{-1}(\mathtt{T}_n f_0 + g_n),$$

where $\mathtt{I}$ is the identity operator and

$$g_n = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \mathtt{L}_K^{1/2} X_i.$$

It is also clear that

$$\mathcal{E}(\eta_{n,\lambda}) - \mathcal{E}(\eta_0) = \|\mathtt{T}^{1/2}(f_{n,\lambda} - f_0)\|^2.$$

Recall that

$$Y_i = < X_i, \mathtt{L}_K^{1/2} f_0 > + \epsilon_i.$$

Then we obtain

$$\begin{aligned} \|\mathtt{T}^{1/2}(f_{n,\lambda} - f_0)\| \;\; \leq \;\; & \|\mathtt{T}^{1/2}(\mathtt{T}_n + \lambda \mathtt{I})^{-1} g_n\| + \\ & \|\mathtt{T}^{1/2}((\mathtt{T}_n + \lambda \mathtt{I})^{-1} \mathtt{T}_n f_0 - f_0)\|. \end{aligned} \tag{6}$$

In the next section, we estimate the terms of (6) so that $\mathcal{E}(\eta_{n,\lambda}) - \mathcal{E}(\eta_0)$ can be obtained in (5).

3 The Analysis

The main technical tool in our analysis is the integral operators, which is an important approach in learning theory, see [12]. Some techniques used here are also closely connected with the recent papers [5, 8] and references therein. We first define a quantity measuring the learning complexity of kernel regression, the effective dimension of $\mathtt{T}$ [17, 4], to be the trace of the operator $(\mathtt{T} + \lambda \mathtt{I})^{-1}\mathtt{T}$ as follows:

$$\mathtt{D}(\lambda) = Tr((\mathtt{T} + \lambda \mathtt{I})^{-1}\mathtt{T}).$$

We consider that $\|\cdot\|_{op}$ stands for the usual operator norm, that is, $\|\mathtt{U}\|_{op} = \sup_{f:\|f\|=1} \|\mathtt{U}f\|$ for an operator $\mathtt{U} : \mathcal{L}^2 \to \mathcal{L}^2$. Under the assumption in (3), we show the following results.

Theorem 3.1. *Under the assumption in (3), for any $0 < \delta < 1$, with confidence at least $1-\delta$, there holds*

$$\|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}(\mathtt{T}-\mathtt{T}_n)\|_{op} \leq B_{n,\lambda}\log(2/\delta),$$

where

$$B_{n,\lambda} = \frac{2\kappa}{\sqrt{n}}\left\{\frac{\kappa}{\sqrt{n}\lambda} + \sqrt{\mathtt{D}(\lambda)}\right\}.$$

To show the above theorem, we need the following probability inequality stated in [10].

Theorem 3.2. *Let $\mathcal{H}$ be a Hilbert space and ξ be a random variable with values in $\mathcal{H}$. Assume that $\|\xi\|_{\mathcal{H}} \leq M$ almost surely. Let $\{\xi_1, \xi_2, \cdots, \xi_n\}$ be a sample of n independent observations for ξ. Then for any $0 < \delta < 1$,*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}(\xi)]\right\|_{\mathcal{H}} \leq \frac{2M\log(2/\delta)}{n} + \sqrt{\frac{2\mathbb{E}(\|\xi\|_{\mathcal{H}}^2)\log(2/\delta)}{n}}.$$

with confidence at least $1-\delta$.

Proof of Theorem 3.1: Consider the random variable

$$\xi = (\mathtt{T}+\lambda\mathtt{I})^{-1/2} < \mathtt{L}_K^{1/2}X, \cdot > \mathtt{L}_K^{1/2}X.$$

It takes values in $HS(\mathcal{L}^2)$, the Hilbert space of Hilbert-Schmidt operators on $\mathcal{L}^2$, with inner product $< \mathtt{A}, \mathtt{B} >_{HS} = Tr(\mathtt{B}^T\mathtt{A})$. The norm is given by $\|\mathtt{A}\|_{HS}^2 = \sum_i \|\mathtt{A}e_i\|^2$ where $\{e_i\}$ is an orthonormal basis of $\mathcal{L}^2$. The space $HS(\mathcal{L}^2)$ is a subspace of the space of bounded linear operators on $\mathcal{L}^2$, with the norm relations

$$\|\mathtt{A}\|_{op} \leq \|\mathtt{A}\|_{HS}, \quad \|\mathtt{AB}\|_{HS} \leq \|\mathtt{A}\|_{HS}\|\mathtt{B}\|_{op}. \tag{7}$$

Recall the set of eigenfunctions $\{\varphi_k : k \geq 1\}$ of $\mathtt{T}$ form an orthonormal basis of $\mathcal{L}^2$. By the definition of the HS norm, we have

$$\begin{aligned}
\|\xi\|_{HS}^2 &= \sum_k \|(\mathtt{T}+\lambda\mathtt{I})^{-1/2} < \mathtt{L}_K^{1/2}X, \varphi_k > \mathtt{L}_K^{1/2}X\|^2 \\
&\leq \|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}\|_{op}^2\|\mathtt{L}_K^{1/2}X\|^2 \sum_k | < \mathtt{L}_K^{1/2}X, \varphi_k > |^2 \\
&\leq \lambda^{-1}\|\mathtt{L}_K^{1/2}X\|^4 \\
&\leq \frac{\kappa^4}{\lambda}.
\end{aligned}$$

It follows that

$$\|\xi\|_{HS} \leq \frac{\kappa^2}{\sqrt{\lambda}}.$$

Also $\mathtt{L}_K^{1/2}X \in \mathcal{L}^2$ can be expanded by the orthonormal basis $\{\varphi_l\}_l$ as

$$\mathtt{L}_K^{1/2}X = \sum_l < \mathtt{L}_K^{1/2}X, \varphi_l > \varphi_l.$$

Hence

$$\begin{aligned}
\|\xi\|_{HS}^2 &= \sum_k \| < \mathtt{L}_K^{1/2}X, \varphi_k > \sum_l < \mathtt{L}_K^{1/2}X, \varphi_l > (\mathtt{T} + \lambda\mathbf{1})^{-1/2}\varphi_l\|^2 \\
&\leq \sum_k | < \mathtt{L}_K^{1/2}X, \varphi_k > |^2 \| \sum_l < \mathtt{L}_K^{1/2}X, \varphi_l > \frac{1}{\sqrt{\lambda + \tau_l}}\varphi_l\|^2 \\
&\leq \|\mathtt{L}_K^{1/2}X\|^2 \sum_l \frac{| < \mathtt{L}_K^{1/2}X, \varphi_l > |^2}{\lambda + \tau_l} \\
&\leq \kappa^2 \sum_l \frac{| < \mathtt{L}_K^{1/2}X, \varphi_l > |^2}{\lambda + \tau_l}.
\end{aligned}$$

We can see that

$$\begin{aligned}
\mathbb{E}\|\xi\|_{HS}^2 &\leq \kappa^2 \sum_l \frac{\left\langle \mathbb{E} < \mathtt{L}_K^{1/2}X, \varphi_l > \mathtt{L}_K^{1/2}X, \varphi_l \right\rangle}{\lambda + \tau_l} \\
&= \kappa^2 \sum_l \frac{\langle \mathtt{T}\varphi_l, \varphi_l \rangle}{\lambda + \tau_l} \\
&= \kappa^2 \sum_l \frac{\tau_l}{\lambda + \tau_l} \\
&= \kappa^2 \mathtt{D}(\lambda).
\end{aligned}$$

Applying Theorem 3.2 to the random variable ξ with $M = \frac{\kappa^2}{\sqrt{\lambda}}$, we know by (7) that with confidence at least $1 - \delta$,

$$\begin{aligned}
\|(\mathtt{T} + \lambda\mathtt{I})^{-1/2}(\mathtt{T} - \mathtt{T}_n)\|_{op} &= \left\| \frac{1}{n} \sum_{i=1}^n [\mathbb{E}(\xi) - \xi_i] \right\|_{op} \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n [\mathbb{E}(\xi) - \xi_i] \right\|_{HS} \\
&\leq \frac{2\kappa^2 \log(2/\delta)}{n\sqrt{\lambda}} + \sqrt{\frac{2\kappa^2 \mathtt{D}(\lambda) \log(2/\delta)}{n}} \\
&\leq B_{n,\lambda} \log(2/\delta).
\end{aligned}$$

The theorem follows. □

We note that if $\mathtt{A}$ and $\mathtt{B}$ are invertible operators on a Banach space, one can see the following decomposition of the operator product

$$\mathtt{B}\mathtt{A}^{-1} = (\mathtt{B}-\mathtt{A})\mathtt{B}^{-1}(\mathtt{B}-\mathtt{A})\mathtt{A}^{-1} + (\mathtt{B}-\mathtt{A})\mathtt{B}^{-1} + \mathtt{I}. \tag{8}$$

By considering $\mathtt{A} = \mathtt{T}_n + \lambda\mathtt{I}$ and $\mathtt{B} = \mathtt{T} + \lambda\mathtt{I}$ in (8), and applying Theorem 3.1, we have the following results.

Theorem 3.3. *Under the assumption in (3), for any* $0 < \delta < 1$ *with confidence at least* $1-\delta$*, there holds*

$$\|(\mathtt{T}+\lambda\mathtt{I})(\mathtt{T}_n+\lambda\mathtt{I})^{-1}\|_{op} \leq \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^2.$$

Moreover, the confidence set is the same as that in Theorem 3.1.

Proof: We apply (8) to the operator $\mathtt{A} = \mathtt{T}_n + \lambda\mathtt{I}$ and $\mathtt{B} = \mathtt{T} + \lambda\mathtt{I}$. By applying the bounds $\|(\mathtt{T}_n+\lambda\mathtt{I})^{-1}\|_{op} \leq 1/\lambda$ and $\|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}\|_{op} \leq 1/\sqrt{\lambda}$ gives

$$\begin{aligned}\|(\mathtt{T}+\lambda\mathtt{I})(\mathtt{T}_n+\lambda\mathtt{I})^{-1}\|_{op} \quad &\leq \quad \frac{1}{\lambda}\|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}(\mathtt{T}-\mathtt{T}_n)\|_{op}^2 + \\ &\qquad \frac{1}{\sqrt{\lambda}}\|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}(\mathtt{T}-\mathtt{T}_n)\|_{op}+1\end{aligned}$$

Here we have used the fact that

$$\|\mathtt{U}\mathtt{V}\|_{op} = \|(\mathtt{U}\mathtt{V})^T\|_{op} = \|\mathtt{V}^T\mathtt{U}^T\|_{op} = \|\mathtt{V}\mathtt{U}\|_{op} \tag{9}$$

for any self-adjoint operators $\mathtt{U},\mathtt{V}$ on Hilbert spaces. The application of Theorem 3.1 yields with the same confident set in Theorem 3.1,

$$\begin{aligned}\|(\mathtt{T}+\lambda\mathtt{I})(\mathtt{T}_n+\lambda\mathtt{I})^{-1}\|_{op} \quad &\leq \quad \left(\frac{1}{\sqrt{\lambda}}\|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}(\mathtt{T}-\mathtt{T}_n)\|_{op}+1\right)^2 \\ &\leq \quad \left(\frac{B_{n,\lambda}\log(2/\delta)}{\sqrt{\lambda}}+1\right)^2.\end{aligned}$$

The theorem follows. □

Next we can bound the first term on the right-hand side of (6) using Theorem 3.3.

Theorem 3.4. *Under the assumption in (3), for any* $0 < \delta < 1$*, with confidence at least* $1-\delta$*, there holds*

$$\|\mathtt{T}^{1/2}(\mathtt{T}_n+\lambda\mathtt{I})^{-1}g_n)\| \leq \frac{\sigma(\log\frac{4}{\delta})^2}{\kappa\delta}\left(\frac{B_{n,\lambda}}{\sqrt{\lambda}}+1\right)^2 B_{n,\lambda}.$$

Proof: By applying the fact (see e.g. [1, Lemma A.7]) that

$$\|A^\gamma B^\gamma\|_{op} \le \|AB\|_{op}^\gamma, \quad 0 < \gamma < 1,$$

for positive operators A and B on Hilbert spaces and (9), we have

$$\begin{aligned} & \|\mathtt{T}^{1/2}(\mathtt{T}_n + \lambda \mathbf{1})^{-1} g_n)\| \\ \le\ & \|(\mathtt{T} + \lambda\mathtt{I})^{1/2}(\mathtt{T}_n + \lambda\mathtt{I})^{-1/2}\|_{op}\|(\mathtt{T}_n + \lambda\mathtt{I})^{-1/2}(\mathtt{T} + \lambda\mathtt{I})^{1/2}\|_{op}\|(\mathtt{T} + \lambda\mathtt{I})^{-1/2} g_n\| \\ \le\ & \|(\mathtt{T} + \lambda\mathtt{I})(\mathtt{T}_n + \lambda\mathtt{I})^{-1}\|_{op}\|(\mathtt{T} + \lambda\mathtt{I})^{-1/2} g_n\|. \end{aligned}$$

By using Theorem 3.3, we have with confidence at least $1 - \delta/2$

$$\|(\mathtt{T} + \lambda\mathtt{I})(\mathtt{T}_n + \lambda\mathtt{I})^{-1}\|_{op} \le \left(\frac{B_{n,\lambda}\log(4/\delta)}{\sqrt{\lambda}} + 1\right)^2. \tag{10}$$

To estimate $\|(\mathtt{T} + \lambda\mathtt{I})^{1/2} g_n\|$, we recall that

$$g_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i \mathtt{L}_K^{1/2} X_i$$

and consider the random variable ξ defined by

$$\xi = (\mathtt{T} + \lambda\mathtt{I})^{-1/2}(\epsilon \mathtt{L}_K^{1/2} X).$$

It takes values in $\mathcal{L}^2$ and satisfies

$$\mathbb{E}(\xi) \;=\; \mathbb{E}[\epsilon(\mathtt{T} + \lambda\mathtt{I})^{-1/2}\mathtt{L}_K^{1/2} X] = \mathbb{E}[\mathbb{E}[\epsilon(\mathtt{T} + \lambda\mathtt{I})^{-1/2}\mathtt{L}_K^{1/2} X|X]] = 0,$$

and

$$\begin{aligned} \mathbb{E}\|\xi\|^2 &= \mathbb{E}\|\epsilon\sum_l < \mathtt{L}_K^{1/2} X, \varphi_l > (\mathtt{T} + \lambda\mathtt{I})^{-1/2}\varphi_l\|^2 \\ &= \sigma^2 \sum_l \frac{\mathbb{E}\left(< \mathtt{L}_K^{1/2} X, \varphi_l >\right)^2}{\lambda + \tau_l} \\ &= \sigma^2 \sum_l \frac{< \mathtt{T}\varphi_l, \varphi_l >}{\lambda + \tau_l} \\ &= \sigma^2 \mathtt{D}(\lambda). \end{aligned}$$

Hence we obtain

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

By applying Markov inequality, we have with confidence at least $1-\delta/2$,

$$\|(\mathtt{T}+\lambda\mathtt{I})^{-1/2}g_n\| \;\leq \frac{2\sigma}{\delta}\sqrt{\frac{\mathtt{D}(\lambda)}{n}} \leq \;\frac{\sigma}{\kappa\delta}B_{n,\lambda}.$$

This together with (10) proves the theorem. □

Next we estimate the second term of the right-hand side in (6).

Theorem 3.5. *Under the assumption in (3), with the same confidence set of Theorem 3.1, there holds*

$$\|\mathtt{T}^{1/2}((\mathtt{T}_n+\lambda\mathtt{I})^{-1}\mathtt{T}_nf_0-f_0)\| \leq \|f_0\|\sqrt{\lambda}\left(\frac{B_{n,\lambda}\log(4/\delta)}{\sqrt{\lambda}}+1\right).$$

Proof: Since $(\mathtt{T}_n+\lambda\mathtt{I})^{-1}\mathtt{T}_nf_0-f_0=-\lambda(\mathtt{T}_n+\lambda\mathtt{I})^{-1}f_0$, we get from Theorem 3.3 with confidence at least $1-\delta/2$,

$$\begin{aligned}
&\|\mathtt{T}^{1/2}((\mathtt{T}_n+\lambda\mathtt{I})^{-1}\mathtt{T}_nf_0-f_0)\|\\
\leq\;& \lambda\|(\mathtt{T}+\lambda\mathtt{I})^{1/2}(\mathtt{T}_n+\lambda\mathtt{I})^{-1/2}\|_{op}\|(\mathtt{T}_n+\lambda\mathtt{I})^{-1/2}\|_{op}\|f_0\|\\
\leq\;& \|f_0\|\sqrt{\lambda}\|(\mathtt{T}+\lambda\mathtt{I})(\mathtt{T}_n+\lambda\mathtt{I})^{-1}\|_{op}^{1/2}\\
\leq\;& \|f_0\|\sqrt{\lambda}\left(\frac{B_{n,\lambda}\log(4/\delta)}{\sqrt{\lambda}}+1\right). \qquad (11)
\end{aligned}$$

Note that (10) and (11) hold simultaneously with probability at least $1-\delta/2$, we thus derive the error bound. □

Next we derive explicit learning rates of regularized least squares for functional linear regression, we need to quantify the increment of $\mathtt{D}(\lambda)$ with a parameter $0<\theta\leq 1$ and a constant $c>0$ as in (4).

Theorem 3.6. *Under the assumption (3) and (4), for any $0<\delta<1$, by taking $\lambda=(\frac{1}{n})^{\frac{1}{1+\theta}}$, we have with confidence at least $1-\delta$,*

$$\mathcal{E}(\eta_{n,\lambda})-\mathcal{E}(\eta_0) \leq \frac{C(\log\frac{4}{\delta})^4 n^{\frac{-1}{1+\theta}}}{\delta^2},$$

where C is a constant independent of n or δ.

Proof: Let $\lambda=(\frac{1}{n})^{\frac{1}{1+\theta}}$, it is easy to check $B_{n,\lambda}\leq 2\kappa(\kappa+\sqrt{c})\sqrt{\lambda}$. Since Theorems 3.4 and 3.5 hold simultaneously with probability at least $1-\delta$, we get from (6)

$$\begin{aligned}
&\mathcal{E}(\eta_{n,\lambda})-\mathcal{E}(\eta_0)\\
=\;& \|\mathtt{T}^{1/2}(f_{n\lambda}-f_0)\|^2\\
\leq\;& 2\left(\|\mathtt{T}^{1/2}(\mathtt{T}_n+\lambda\mathtt{I})^{-1}g_n\|^2+\|\mathtt{T}^{1/2}((\mathtt{T}_n+\lambda\mathtt{I})^{-1}\mathtt{T}_nf_0-f_0)\|^2\right)\\
\leq\;& 2\frac{(\log\frac{4}{\delta})^4\lambda}{\delta^2}\left(\frac{\sigma^2(2\kappa(\kappa+\sqrt{c})+1)^6}{\kappa^2}+(2\kappa(\kappa+\sqrt{c})+1)^2\|f_0\|^2\right).
\end{aligned}$$

This proves the theorem with

$$C = 2\left(\frac{\sigma^2(2\kappa(\kappa+\sqrt{c})+1)^6}{\kappa^2} + (2\kappa(\kappa+\sqrt{c})+1)^2\|f_0\|^2\right).$$

□

Here we give some remarks for the above theorem. In terms of methodology, the analysis in this paper is most closely related of that of [3], where a minimax rate of convergence of the excess risk is derived under the assumption that the eigenvalues τ_k of $\mathtt{T}$ decay as follows:

$$\tau_k \leq c_1 k^{-2r}, \tag{12}$$

for some constant $c_1 > 0$ and $1/2 < r < \infty$[1]. We note that our assumption in (4) for $0 < \theta < 1$ is more general than that in (12). In fact, if (12) is satisfied, it is easy to check that

$$\begin{aligned}\mathcal{D}(\lambda) &= \sum_{l}^{\infty}\frac{\tau_l}{\lambda+\tau_l} \leq \sum_{l}^{\infty}\frac{c_1 l^{-2r}}{\lambda + c_1 l^{-2r}} = \sum_{l}^{\infty}\frac{c_1}{c_1+\lambda l^{2r}}\\ &\leq \int_0^\infty \frac{c_1}{c_1+\lambda t^{2r}}dt \leq c'\lambda^{-\frac{1}{2r}}.\end{aligned}$$

It implies that condition in (4) holds with $\theta = \frac{1}{2r}$. We thus conclude the results in the following theorem.

Theorem 3.7. *Suppose the eigenvalues $\{\tau_k : k \geq 1\}$ of $\mathtt{T}$ satisfy $\tau_k \leq c_1 k^{-2r}$ for some constant $c_1 > 0$ and $1/2 < r < \infty$. Under the assumption in (3), for any $0 < \delta < 1$, we have with confidence at least $1-\delta$,*

$$\mathcal{E}(\eta_{n,\lambda}) - \mathcal{E}(\eta_0) \leq \frac{C'(\log\frac{4}{\delta})^4}{\delta^2} n^{\frac{-2r}{1+2r}}.$$

The convergence rate presented in Theorem 3.7 is the same as the minimax rate in [3]. It is interesting to note in [3] that the following assumption

$$\mathbb{E}\left(\int_{\mathbb{I}} X(t)f(t)dt\right)^4 \leq c_2\left(\mathbb{E}\left(\int_{\mathbb{I}} X(t)f(t)dt\right)^2\right)^2 \tag{13}$$

for all $f \in \mathcal{L}^2$ with $c_2 > 0$, is required in order to derive the minmax rate for (2). We know by Cauchy inequality that

$$\left(\mathbb{E}\left(\int_{\mathbb{I}} X(t)f(t)dt\right)^2\right)^2 \leq \mathbb{E}\left(\int_{\mathbb{I}} X(t)f(t)dt\right)^4. \tag{14}$$

[1] The original assumption in [3] requires eigenvalues $\{\tau_k : k \geq 1\}$ satisfy $\tau_k \leq c' k^{-2r}$ for some $0 < r < \infty$ and c' is a positive constant. But the proof of [3, Lemma 1] requires $r > 1/2$.

It is clear to see that (13) is an opposite inequality to (14). Therefore, (13) may be very difficult to verify except for Gaussian data X. Compared with (13), our assumption (3) is more realistic. In fact,

$$\begin{aligned}\mathbb{E}\|\mathtt{L}_K^{1/2}X\|^2 &= \mathbb{E}<\mathtt{L}_K X, X>=\mathbb{E}\int_{\mathbb{I}\times\mathbb{I}} K(s,t)X(s)X(t)dsdt\\ &= \int_{\mathbb{I}\times\mathbb{I}} K(s,t)C(s,t)dsdt\\ &\leq \|K\|_{\mathcal{L}^2(\mathbb{I}\times\mathbb{I})}\cdot\|C\|_{\mathcal{L}^2(\mathbb{I}\times\mathbb{I})}.\end{aligned}$$

By Markov inequality, for any $0<\eta<1$, with confidence at least $1-\eta$,

$$\|\mathtt{L}_K^{1/2}X\|^2 \leq \frac{\|K\|_{\mathcal{L}^2(\mathbb{I}\times\mathbb{I})}\cdot\|C\|_{\mathcal{L}^2(\mathbb{I}\times\mathbb{I})}}{\eta}.$$

This implies $\|\mathtt{L}_K^{1/2}X\|^2$ is a bounded set with high probability. From a practical point of view, although condition (3) is not satisfied for any non-degenerate Gaussian process, real-data processes are bounded as usual. Hence, results of the paper can be considered as complimentary results to [3].

4 Concluding Remarks

In this paper, we have derived the minimax rate of of regularized least squares for functional linear regression. Our required assumptions are more general and realistic than those in the literature. On the other hand, we focus on scalar response in the functional linear model. It would be interesting to consider multiple responses arising from multilinear model and study the corresponding regularized least squares setting as a future research work. Such a setting can have useful applications in image and multi-dimensional signal processing.

Acknowledgements

The authors thank the referees for their valuable comments and suggestions.

References

[1] G. Blanchard, N. Krämer, Optimal learning rates for kernel conjugate gradient regression, In: NIPs, 226-234, 2010.

[2] T. Cai and P. Hall, Prediction in function linear regression, *Ann. Statist.*, 34, 2159-2179, 2006.

[3] T. Cai and M. Yuan, Minimax and adaptive prediction for function linear regression, *J. Amer. Statist. Assoc.*, 107, 1201-1216, 2012.

[4] A. Caponnetto and E. DeVito, Optimal rates for the regularized least squares algorithm, *Foundations of Computational Mathematics*, 7, 331-368, 2007.

[5] Z. C. Guo, S. B. Lin and D. X. Zhou, Learning theory of distributed spectral algorithms, *Inverse Problems*, 33(7), 074009, 2017.

[6] P. Hall and G. L. Horowitz, Methodology and convergence rates for function linear regression, *Ann. Statist.*, 35, 70-91, 2007.

[7] T. Hastie and C. Mallow, A discussion of "A statistical view of some chemometrics regression tools" by I. E. Frank and J. H. Friedman, *Technometrics*, 35, 140-143, 1993.

[8] S. Lu, P., Mathe and S. V. Pereverzev, Balancing principle in supervised learning for a general regularization scheme, *Applied and Computational Harmonic Analysis*, https://doi.org/10.1016/j.acha.2018.03.001.

[9] V. Paulsen and M. Raghupathi, *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*, Cambridge University Press, 2016.

[10] I. Pinelis, Optimum bounds for the distributions of martingales in Banach spaces, *The Annals of Probality*, 22, 679-1706, 1994.

[11] J. O. Ramsay and C. J. Dalzell, Some tools for functional data analysis (with Discussion), *Journal of the Royal Statistical Society*, Series B, 53, 539C572, 1991.

[12] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their approximations, *Constr. Approx.*, 26, 153-172, 2007.

[13] I. Steinwart, Support vector machines are universally consistent, *J. Complexity*, 18, 768-791, 2002.

[14] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, *J. Mach. Learning Res.*, 2, 67-93, 2001.

[15] F. Yao, H. G. Müller and J. L. Wang, Function linear regression analysis for longitudinal data, *Ann. Statist.*, 33, 2873-2903, 2005.

[16] M. Yuan and T. Cai, A reproducing kernel Hilbert space approach to function linear regression, *Ann. Statist.*, 38, 3412-3444, 2010.

[17] T. Zhang, Learning bounds for kernel regression using effective data dimensionality, *Neural Computation*, 17, 2077-2098, 2004.