Online gradient descent algorithms for functional data learning

Xiaming Chen[1], Bohao Tang[2], Jun Fan[3], and Xin Guo[4]

[1]Department of Computer Science, Shantou University, Shantou, China, chenxm@stu.edu.cn

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, bhtang@jhu.edu

[3]Department of Mathematics, Hong Kong Baptist University, Kowloon, Hong Kong, junfan@hkbu.edu.hk

[4]Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong, x.guo@polyu.edu.hk

December 1, 2021

Abstract

Functional linear model is a fruitfully applied general framework for regression problems, including those with intrinsically infinite-dimensional data. Online gradient descent methods, despite their evidenced power of processing online or large-sized data, are not well studied for learning with functional data. In this paper, we study reproducing kernel-based online learning algorithms for functional data, and derive convergence rates for the expected excess prediction risk under both online and finite-horizon settings of step-sizes respectively. It is well understood that nontrivial uniform convergence rates for the estimation task depend on the regularity of the slope function. Surprisingly, the convergence rates we derive for the prediction task can assume no regularity from slope. Our analysis reveals the intrinsic difference between the estimation task and the prediction task in functional data learning.

Keywords: Learning theory, online learning, gradient descent, reproducing kernel Hilbert space, error analysis

1 Introduction

In this paper, we study the functional linear model

$$Y = \alpha^* + \int_{\mathcal{T}} X(t)\beta^*(t)dt + \varepsilon. \tag{1}$$

Here, the predictor X is a random function on a compact set $\mathcal{T}$ in some Euclidean space. The slope (or coefficient) β^* is an unknown function. We assume that X and β^* are both in the space $(L^2(\mathcal{T}), \langle\cdot,\cdot\rangle, \|\cdot\|)$ of square integrable functions. The number α^* is the constant intercept. The error ε is a zero-mean random variable with variance $\sigma^2 < \infty$, and it is independent of X. We let Y denote the response.

In this paper, for technical simplicity we assume $\alpha^* = 0$, and that the mean function is zero, i.e. $\mathbb{E}[X] = 0$. Consequently, $\mathbb{E}[Y] = 0$.

Let $D = \{(X_i, Y_i)\}_{i=1}^n$ be a sample of independent copies of (X, Y). The prediction problem of functional linear regression is to exploit D and find a linear functional $\hat{\eta}$ on $L^2(\mathcal{T})$ as estimator of the unknown functional η^* on $L^2(\mathcal{T})$,

$$\eta^*(X) = \langle X, \beta^*\rangle = \int_{\mathcal{T}} X(t)\beta^*(t)dt,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product in $L^2(\mathcal{T})$. Denote $\mathcal{E}(\hat{\eta})$ the excess prediction risk of $\hat{\eta}$,

$$\begin{aligned}\mathcal{E}(\hat{\eta}) &= \mathbb{E}_{(X,Y)}\left[(Y - \hat{\eta}(X))^2 - (Y - \eta^*(X))^2\right] \\ &= \mathbb{E}_X\left[(\hat{\eta}(X) - \eta^*(X))^2\right],\end{aligned}$$

where (X, Y) is independent of $\hat{\eta}$ and $\mathbb{E}_{(X,Y)}$ denotes the expectation taken with respect to the distribution of (X, Y). The expectation $\mathbb{E}_X$ is similarly defined.

There is a large literature on function linear models. See [4, 5, 23, 18] and the references therein. Functional principal component analysis (FPCA) is a popular tool for the regression problems [4, 13]. FPCA makes use of the functional principal component representation of X with fast decaying coefficients for estimation, and usually requires strong regularity of the slope function β^*. Another widely applied approach is the reproducing kernel method [23, 5, 15, 19, 14], which represents functions by linear combinations of kernel functions, so that the regression problem is, in computation, reduced to optimization problems over the coefficient vector spaces. The algorithms studied in this paper are designed with reproducing kernels.

In this paper, we study the online stochastic gradient descent scheme which starts from $\beta_1 = 0$ and is then iteratively defined by

$$\beta_{k+1} = \beta_k - \gamma_k \left(\int_{\mathcal{T}} \beta_k(t)X_k(t)dt - Y_k\right)\int_{\mathcal{T}} K(s,\cdot)X_k(s)ds. \tag{2}$$

Here $\gamma_k > 0$ is the step-size. $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is a continuous reproducing kernel (a.k.a. Mercer kernel), which is defined to be continuous, symmetric (i.e., $K(s,t) \equiv K(t,s)$), and positive semi-definite (i.e., the Gramian matrix $(K(t_i, t_j) : 1 \le i, j \le n)$ is positive semi-definite for any $n \ge 1$ and any $t_1, \dots, t_n \in \mathcal{T}$). The function K defines an integral operator $L_K : L^2(\mathcal{T}) \to L^2(\mathcal{T})$,

$$L_K f = \int_{\mathcal{T}} K(s, \cdot) f(s) ds, \tag{3}$$

which is known to be compact and positive semi-definite. See, e.g., [7, Section 4.2], and [20, Section 4.3]. So, (2) can be equivalently written as

$$\beta_{k+1} = \beta_k - \gamma_k \left(\langle \beta_k, X_k \rangle - Y_k \right) L_K X_k. \tag{4}$$

After n iterations, the output estimator $\hat{\eta}_{n+1}$ of the predictor η^* is defined by

$$\hat{\eta}_{n+1}(X) = \langle \beta_{n+1}, X \rangle = \int_{\mathcal{T}} \beta_{n+1}(t) X(t) dt. \tag{5}$$

We study two settings of the step-sizes $\{\gamma_k\}$ and the data set D.

- The *online* setting. In this setting, one takes D as a source of (finite or infinite) sample points and the iterations continue indefinitely, before the possible exhaustion of D. We use a decreasing sequence $\{\gamma_k = \gamma_1 k^{-\mu}\}$ of step-sizes with some $\mu > 0$, and update the estimated predictor after each step of iteration.

- The *finite-horizon* setting. In this setting we assume a finite sample size $n = |D| < \infty$ and use a fixed step-size $\gamma_k \equiv \gamma_0 n^{-\mu}$ that is dependent on n. The iteration is scheduled to terminate after the exhaustion of D.

Let C be the covariance function of X (recall that $\mathbb{E}[X] = 0$),

$$C(s, t) = \mathbb{E}\left[(X(s) - \mathbb{E}[X(s)])(X(t) - \mathbb{E}[X(t)]) \right] = \mathbb{E}[X(s)X(t)].$$

It is easy to verify that C is symmetric and positive semi-definite. In this paper we assume that C is continuous, so C is another Mercer kernel. We define the integral operator L_C by replacing K with C in (3), so L_C is also compact and positive semi-definite.

In the analysis, we make the following assumptions.

(A1) The coefficient β^* satisfies

$$L_C^{1/2} \beta^* = \mathscr{L}^\theta g^*, \quad \text{for some } g^* \in L^2(\mathcal{T}) \text{ and } \theta > 0,$$

where $\mathscr{L} = L_C^{1/2} L_K L_C^{1/2}$. We shall discuss this assumption in Section 3.

(A2) There exists some constant $0 < c < \infty$ such that for any $\beta \in L^2(\mathcal{T})$,

$$\mathbb{E} \left(\int_{\mathcal{T}} \beta(t) X(t) dt \right)^4 \le c \left(\mathbb{E} \left(\int_{\mathcal{T}} \beta(t) X(t) dt \right)^2 \right)^2. \tag{6}$$

One uses vector notation to rewrite (6) as $\mathbb{E}[\langle\beta, X\rangle^4] \le c(\mathbb{E}[\langle\beta, X\rangle^2])^2$. Assumptions (A1) and (A2) are adopted in [9] to establish the convergence of excess risk of a functional data-based classifier under the framework of optimal individualized treatment rules. Assumption (A2) is also used in [23, 5] for the analysis of kernel-based batch learning scheme for functional linear regression. Nonetheless, it remains an interesting open question whether Assumption (A2) is technical or intrinsic, for the convergence of Algorithm (2).

2 Main Results

This paper studies the online scheme (2) for learning the predictor η^* of the functional linear regression model (1). In this section we provide the convergence rates of the expected excess risk. Recall that K and C are both continuous, and $\mathcal{T}$ is compact. So

$$\kappa_1 := \max_{t\in\mathcal{T}} \sqrt{K(t,t)} < \infty, \quad \text{and} \quad \kappa_2 := \max_{t\in\mathcal{T}} \sqrt{C(t,t)} < \infty.$$

In Theorem 1 below, we study the online setting. The data set D is assumed to be a source of finite or infinite sample points, and the estimator $\hat{\eta}_{n+1}$ is obtained with the first n sample points. The iteration (4) can continue until the possible exhaustion of D.

Theorem 1. *Let $\{\hat{\eta}_{n+1} : n \ge 1\}$ be a sequence of estimators defined by (5) and (4) with step-sizes $\gamma_k = \gamma_1 k^{-\mu}$. Assume (A1) with $\theta > 0$, (A2), and let*

$$\mu = \min\left\{\frac{1}{2}, \frac{2\theta}{2\theta+1}\right\} = \begin{cases} \frac{2\theta}{2\theta+1}, & \text{when } 0 < \theta \le 1/2, \\ \frac{1}{2}, & \text{when } \theta > 1/2. \end{cases} \tag{7}$$

If $\gamma_1 \le \mu/[2^{1+\mu}(1+c)(1+\kappa_1^2\kappa_2^2)^2 C_\mu]$ (where C_μ is defined by Lemma 4 below), then for any $n \ge 1$,

$$\mathbb{E}[\mathcal{E}(\hat{\eta}_{n+1})] \le C_1 n^{-\mu} \log(n+1),$$

where C_1 is a constant independent of n, and it will be specified in the proof.

In Theorem 2 below, we study the finite-horizon setting. The data set D is assumed to be finite with size $n = |D| \ge 1$. After n steps of iterations D is exhausted, and the algorithm terminates and outputs the estimator $\hat{\eta}_{n+1}$.

Theorem 2. *Let $\hat{\eta}_{n+1}$ be the estimator defined by (5) and (4) with a finite sample $D = \{(X_i, Y_i)\}_{i=1}^n$ and the constant step-size $\gamma_k \equiv \gamma$. Assume (A1) with $\theta > 0$ and (A2). If $\gamma = \gamma_0 n^{-2\theta/(2\theta+1)}$ with*

$$0 < \gamma_0 \le \frac{1}{2(1+c)(1+\kappa_1^2\kappa_2^2)^2\left(1 + \frac{2\theta+1}{2e\theta}\right)},$$

then we have

$$\mathbb{E}[\mathcal{E}(\hat{\eta}_{n+1})] \le C_2 n^{-2\theta/(2\theta+1)} \log(n+1),$$

where C_2 is a constant independent of n, and it will be specified in the proof.

Remark 1. *The analysis in [23, 5] requires the regularity $\beta^* \in L_K^{1/2}(L^2(\mathcal{T}))$ (that is, β^* resides in the reproducing kernel Hilbert space generated by K). It is well understood in the literature of regression learning that for an algorithm to learn β^* from data (a.k.a. the estimation problem), a non-trivial convergence rate depends on the regularity of β^* [6, 21, 3, 1, 16]. Similar results for classification problems are referred to as no-free-lunch theorems [8, 20]. Our analysis relaxes this assumption. From Theorem 3, we see that by properly selecting the kernel K, Theorems 1 and 2 apply with some $0 < \theta < 1/2$, to any $\beta^* \in L^2(\mathcal{T})$. In fact, if there exist some $\delta > 0$ and $0 < \theta < 1/2$, such that $L_K \succeq \delta L_C^\nu$ with $\nu = \frac{1}{2\theta} - 1$, then one uses Theorem 3 to guarantee (A1) for any $\beta^* \in L^2(\mathcal{T})$. The results in this paper show that it is possible to learn the predictor η^* (in a specific convergence rate) without assuming regularity of the slope function β^*. The regularity requirement on β^* is an intrinsic difference between the prediction problem (for learning η^*) and the estimation problem (for learning β^*) in functional data learning.*

3 Discussions on the Regularity Assumption

In this section we discuss the regularity assumption (A1). Theorem 3 below suggests that (A1) is a mild assumption and it can be satisfied for any $\beta^* \in L^2(\mathcal{T})$ by properly selecting the reproducing kernel K, at least for any $0 < \theta < 1/2$.

For any two self-adjoint operators L_1 and L_2, we write $L_1 \preceq L_2$ (or $L_2 \succeq L_1$) if $L_2 - L_1$ is positive semi-definite. Recall that $\|\cdot\|$ denotes the norm in $L^2(\mathcal{T})$.

Theorem 3. *Assume $L_K \succeq \delta L_C^\nu$ for some $\delta > 0$ and $\nu > 0$. Then, for any $\beta^* \in L^2(\mathcal{T})$, there exists some $g^* \in L^2(\mathcal{T})$ such that $L_C^{1/2}\beta^* = \mathscr{L}^\theta g^*$ and $\|g^*\| \leq \delta^{-\theta}\|\beta^*\|$ with $\theta = 1/(2+2\nu)$. In particular, one has $L_C^{1/2}\beta^* \in \mathscr{L}^{\theta_1}(L^2(\mathcal{T}))$ for any $0 < \theta_1 \leq \theta$.*

Remark 2. *Thanks to the results in [10], the assumption in Theorem 3 can be further relaxed to $L_K^\omega \succeq \delta L_C^\nu$ for $\delta, \omega, \nu > 0$ and $\omega + \nu \geq 1$, with $\theta = 1/(2+2\nu/\omega)$. This relaxation is not trivial for $0 < \omega < 1$. Nonetheless, here we do not expand the details.*

Remark 3. *Since the sum of two Mercer kernels is a Mercer kernel, as long as $K - \delta C$ is a Mercer kernel for some $\delta > 0$, one has already $L_K \succeq \delta L_C$.*

Proof of Theorem 3. Since $L_K \succeq \delta L_C^\nu$, we claim that $\mathscr{L} \succeq \delta L_C^{1+\nu}$. In fact, for any $\beta \in L^2(\mathcal{T})$, $\langle \beta, \mathscr{L}\beta\rangle = \left\langle L_C^{1/2}\beta, L_K(L_C^{1/2}\beta)\right\rangle \geq \delta \left\langle L_C^{1/2}\beta, L_C^\nu(L_C^{1/2}\beta)\right\rangle = \langle \beta, \delta L_C^{1+\nu}\beta\rangle$.

Next, we claim that $\mathscr{L}^{1/(1+\nu)} \succeq \delta^{1/(1+\nu)} L_C$. This follows from the fact that the function $f(x) = x^r$ on $x \in [0, \infty)$ with $0 < r \leq 1$ is operator monotone (i.e., $L_1 \preceq L_2$ implies $f(L_1) \preceq f(L_2)$ for any bounded positive semi-definite operators L_1 and L_2) [17].

For any bounded positive semi-definite operators L_1 and L_2 on $L^2(\mathcal{T})$, $L_1 \preceq L_2$ implies that for any $\beta^* \in L^2(\mathcal{T})$, there exists some $g^* \in L^2(\mathcal{T})$ such that $\|g^*\| \leq \|\beta^*\|$ and $L_1^{1/2}\beta^* = L_2^{1/2}g^*$. This is a standard result with the matrix form available in many linear algebra textbooks, e.g. [2, page 114]. See [11] for a proof for operators on Hilbert spaces. Recall that $\mathscr{L}$ is bounded. The proof is complete. $\square$

4 Proofs of the Main Results

We first symbolically decompose the residual $L_C^{1/2}(\beta_{k+1}-\beta^*)$ after k steps of iterations.

Lemma 1. *Let $\{\beta_k : k \in \mathbb{N}\}$ be defined by (4). One has*

$$L_C^{1/2}(\beta_{k+1}-\beta^*) = -\left[\prod_{i=1}^{k}(I-\gamma_i\mathscr{L})\right] L_C^{1/2}\beta^* + \sum_{i=1}^{k}\gamma_i\left[\prod_{j=i+1}^{k}(I-\gamma_j\mathscr{L})\right] B_i, \tag{8}$$

where I is the identity operator (of which the domain is inferred from the context), the product $\prod_{j=i+1}^{k}$ vanishes to I when $i+1>k$, and B_k is defined by

$$B_k = \mathscr{L}L_C^{1/2}(\beta_k-\beta^*) + (Y_k - \langle X_k, \beta_k\rangle)\, L_C^{1/2}L_K X_k. \tag{9}$$

Proof. By definition (4) of β_k, we have

$$\begin{aligned}
L_C^{1/2}(\beta_{k+1}-\beta^*) &= L_C^{1/2}(\beta_k-\beta^*) + \gamma_k\,(Y_k - \langle X_k,\beta_k\rangle)\, L_C^{1/2}L_K X_k \\
&= (I-\gamma_k\mathscr{L})L_C^{1/2}(\beta_k-\beta^*) + \gamma_k B_k \\
&= (I-\gamma_k\mathscr{L})(I-\gamma_{k-1}\mathscr{L})L_C^{1/2}(\beta_{k-1}-\beta^*) + \gamma_{k-1}(I-\gamma_k\mathscr{L})B_{k-1} + \gamma_k B_k \\
&= -\left[\prod_{i=1}^{k}(I-\gamma_i\mathscr{L})\right] L_C^{1/2}\beta^* + \sum_{i=1}^{k}\gamma_i\left[\prod_{j=i+1}^{k}(I-\gamma_j\mathscr{L})\right] B_i.
\end{aligned}$$

This completes the proof. □

We see that B_k in (9) is just the difference between $(Y_k - \langle X_k,\beta_k\rangle)L_C^{1/2}L_K X_k$ and its mean with respect to the observation (X_k, Y_k),

$$\begin{aligned}
\mathbb{E}_{(X_k,Y_k)}\left[(Y_k-\langle X_k,\beta_k\rangle)L_C^{1/2}L_K X_k\right] =& L_C^{1/2}L_K\mathbb{E}_{X_k}\left[\langle\beta^*-\beta_k, X_k\rangle X_k\right] \\
=& L_C^{1/2}L_K L_C(\beta^*-\beta_k) = \mathscr{L}L_C^{1/2}(\beta^*-\beta_k).
\end{aligned} \tag{10}$$

Therefore, $\mathbb{E}[B_k]=0$.

Lemma 2. *Let $\mathcal{A}$ be a compact positive semi-definite operator on some real separable Hilbert space, such that $\|\mathcal{A}\|_{\mathsf{op}} \le C_*$ for some $C_*>0$. Let $l \le k$ and $\gamma_l, \gamma_{l+1}, \dots \gamma_k \in [0, 1/C_*]$. Then, when $\theta > 0$,*

$$\left\|\mathcal{A}^\theta \prod_{j=l}^{k}(I-\gamma_j\mathcal{A})\right\|_{\mathsf{op}}^2 \le \frac{(\theta/e)^{2\theta}+C_*^{2\theta}}{1+(\sum_{j=l}^{k}\gamma_j)^{2\theta}}. \tag{11}$$

When $\theta = 0$, one has

$$\left\|\prod_{j=l}^{k}(I-\gamma_j\mathcal{A})\right\|_{\mathsf{op}}^2 \le 1. \tag{12}$$

In particular, when $l>k$, the above products vanish to the identity operator, and the sum $\sum_{j=l}^{k}\gamma_j$ vanishes to zero, so the bounds (11) and (12) still hold true.

Proof. The case $l > k$ is trivial and we assume $l \le k$. Bound (12) directly follows the fact $0 \le \gamma_j \|\mathcal{A}\|_{\mathsf{op}} \le \gamma_j C_* \le 1$. Now we assume $\theta > 0$. The case $\sum_{j=l}^k \gamma_j = 0$ is trivial and we assume $\sum_{j=l}^k \gamma_j > 0$.

Define polynomial $\tau(x) = x^\theta \prod_{j=l}^k (1-\gamma_j x)$ on $0 \le x \le C_*$. Then $0 \le \gamma_j x \le \gamma_j C_* \le 1$, so $0 \le 1 - \gamma_j x \le 1$, and one has,

$$0 \le \tau(x) \le C_*^\theta. \tag{13}$$

Recall that for fixed $\theta, A > 0$, the function $x^\theta e^{-Ax}$ defined on $x \in [0, \infty)$ achieves its maximum $\theta^\theta (eA)^{-\theta}$ at $x = \theta/A$. One applies the inequality $1 - x \le e^{-x}$ for $x \ge 0$ to obtain

$$\tau(x) \le x^\theta \prod_{j=l}^k e^{-\gamma_j x} = x^\theta \exp\left\{-x \sum_{j=l}^k \gamma_j\right\} \le \theta^\theta \left(e \sum_{j=l}^k \gamma_j\right)^{-\theta}. \tag{14}$$

Recall that for $a, b, c > 0$, $\min(ab, c) \le \frac{1}{1+b} ab + \frac{b}{1+b} c = b(a+c)/(b+1)$. One lets $a = (\theta/e)^{2\theta}$, $b = (\sum_{j=l}^k \gamma_j)^{-2\theta}$, and $c = C_*^{2\theta}$ to derive from (13) and (14) that,

$$\tau^2(x) \le \frac{(\theta/e)^{2\theta} + C_*^{2\theta}}{1 + (\sum_{j=l}^k \gamma_j)^{2\theta}}.$$

We apply the spectral theorem to derive (11). $\square$

Theorem 4. *Let $\{\beta_k : 1 \le k < N\}$ be defined by (4) for some $N \le \infty$. Assume (A2) and that $\gamma_j \kappa_1^2 \kappa_2^2 \le 1$ for any $j \ge 1$. Then for any $1 \le n < N$,*

$$\mathbb{E}[\mathcal{E}(\hat{\eta}_{n+1})] \le \left\| \left[\prod_{i=1}^n (I - \gamma_i \mathscr{L})\right] L_C^{1/2} \beta^* \right\|^2 + (1+c)(1+\kappa_1^2\kappa_2^2)^2 \sum_{i=1}^n \frac{\gamma_i^2 [\mathbb{E}\mathcal{E}(\hat{\eta}_i) + \sigma^2]}{1 + \sum_{j=i+1}^n \gamma_j}. \tag{15}$$

Furthermore, if we assume(A1), then

$$\mathbb{E}[\mathcal{E}(\hat{\eta}_{n+1})] \le \frac{(\theta/e)^{2\theta} + (\kappa_1\kappa_2)^{4\theta}}{1 + \left(\sum_{j=1}^n \gamma_j\right)^{2\theta}} \|g^*\|^2 + (1+c)(1+\kappa_1^2\kappa_2^2)^2 \sum_{i=1}^n \frac{\gamma_i^2 [\mathbb{E}\mathcal{E}(\hat{\eta}_i) + \sigma^2]}{1 + \sum_{j=i+1}^n \gamma_j}. \tag{16}$$

Proof. Recall that $\eta^*(X) = \langle \beta^*, X \rangle$ and $\hat{\eta}_{n+1}(X) = \langle \beta_{n+1}, X \rangle$. We have

$$\begin{aligned}
\mathcal{E}(\hat{\eta}_{n+1}) =& \mathbb{E}_X \left[(\hat{\eta}_{n+1}(X) - \eta^*(X))^2\right] \\
=& \mathbb{E}_X \left(\int_{\mathcal{T}} (\beta_{n+1}(t) - \beta^*(t)) X(t) dt\right)^2 \\
=& \mathbb{E}_X \left(\int_{\mathcal{T}} \int_{\mathcal{T}} (\beta_{n+1}(t) - \beta^*(t))(\beta_{n+1}(s) - \beta^*(s)) X(t) X(s) dt ds\right) \\
=& \langle \beta_{n+1} - \beta^*, L_C(\beta_{n+1} - \beta^*) \rangle \\
=& \left\| L_C^{1/2} (\beta_{n+1} - \beta^*) \right\|^2.
\end{aligned}$$

Here we have used the definition $C(s,t) = \mathbb{E}[X(t)X(s)]$. By Lemma 1,

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{E}(\hat{\eta}_{n+1})\right] =& \mathbb{E}\left[\left\|L_C^{1/2}(\beta_{n+1}-\beta^*)\right\|^2\right] \\
=& -2\mathbb{E}\left\langle \left[\prod_{i=1}^{n}(I-\gamma_i\mathscr{L})\right] L_C^{1/2}\beta^*, \sum_{i=1}^{n}\gamma_i \left[\prod_{j=i+1}^{n}(I-\gamma_j\mathscr{L})\right] B_i \right\rangle \\
&+ \mathbb{E}\left\|\sum_{i=1}^{n}\gamma_i \left[\prod_{j=i+1}^{n}(I-\gamma_j\mathscr{L})\right] B_i\right\|^2 + \left\|\left[\prod_{i=1}^{n}(I-\gamma_i\mathscr{L})\right] L_C^{1/2}\beta^*\right\|^2. \qquad (17)
\end{aligned}
$$

We use J_1 J_2, and J_3 to denote the three terms in the right-hand side of (17), respectively. Here, J_1 is an expectation where the only randomness comes from B_i's, which have zero mean as we discussed in (10). So,

$$J_1 = 0.$$

Now we study J_2. First, we expand the squared norm and write J_2 as a double sum,

$$J_2 = \sum_{i=1}^{n}\sum_{l=1}^{n}\gamma_i\gamma_l\mathbb{E}\left\langle \left[\prod_{j=i+1}^{n}(I-\gamma_j\mathscr{L})\right] B_i, \left[\prod_{j=l+1}^{n}(I-\gamma_j\mathscr{L})\right] B_l \right\rangle.$$

Among the n^2 summands of J_2, whenever $i > l$, since

$$
\begin{aligned}
&\mathbb{E}_{(X_i,Y_i)}\left\langle \left[\prod_{j=i+1}^{n}(I-\gamma_j\mathscr{L})\right] B_i, \left[\prod_{j=l+1}^{n}(I-\gamma_j\mathscr{L})\right] B_l \right\rangle \\
&= \left\langle \left[\prod_{j=i+1}^{n}(I-\gamma_j\mathscr{L})\right] \mathbb{E}_{(X_i,Y_i)}[B_i], \left[\prod_{j=l+1}^{n}(I-\gamma_j\mathscr{L})\right] B_l \right\rangle = 0,
\end{aligned}
$$

the corresponding summand of J_2 is zero. Similar argument applies to the case $i < l$. Therefore,

$$J_2 = \sum_{i=1}^{n}\gamma_i^2\mathbb{E}\left\|\left[\prod_{j=i+1}^{n}(I-\gamma_j\mathscr{L})\right] B_i\right\|^2. \qquad (18)$$

Write $\tilde{B}_i = L_K^{1/2}L_C(\beta_i-\beta^*) + (Y_i - \langle X_i,\beta_i\rangle)L_K^{1/2}X_i$. Similar to (10), we have $\mathbb{E}[\tilde{B}_i] = 0$ because

$$\mathbb{E}_{(X_i,Y_i)}\left[(Y_i-\langle X_i,\beta_i\rangle)L_K^{1/2}X_i\right] = \mathbb{E}_{X_i}\left[\langle\beta^*-\beta_i, X_i\rangle L_K^{1/2}X_i\right] = L_K^{1/2}L_C(\beta^*-\beta_i).$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[\|\tilde{B}_i\|^2\right] &\le \mathbb{E}\left[(Y_i-\langle X_i,\beta_i\rangle)^2\|L_K^{1/2}X_i\|^2\right] \\
&= \mathbb{E}\left[\|L_K^{1/2}X_i\|^2\mathbb{E}_{Y_i}(Y_i-\langle X_i,\beta_i\rangle)^2\right] \\
&= \mathbb{E}\left[\|L_K^{1/2}X_i\|^2\langle\beta^*-\beta_i, X_i\rangle^2\right] + \sigma^2\mathbb{E}\left[\|L_K^{1/2}X_i\|^2\right].
\end{aligned}
$$

Since L_K is positive and compact, we write $\{\lambda_l : l \in \mathcal{I}\}$ the sequence of all the positive eigenvalues of L_K (arranged non-increasingly and counting multiplicity), where $\mathcal{I}$ is a finite or countable set of indices. Write $\{\phi_l : l \in \mathcal{I}\}$ the corresponding eigenvectors normalized in $L^2(\mathcal{T})$. We have $\|L_K^{1/2} X_i\|^2 = \sum_{l\in\mathcal{I}} \lambda_l \langle X_i, \phi_l\rangle^2$. By our assumption (A2) on the moments of X_i,

$$
\begin{aligned}
\mathbb{E}\left[\|L_K^{1/2}X_i\|^2 \langle \beta^* - \beta_i, X_i\rangle^2\right] &= \sum_{l\in\mathcal{I}} \lambda_l \mathbb{E}\left[\langle \phi_l, X_i\rangle^2 \langle \beta^* - \beta_i, X_i\rangle^2\right] \\
&\leq \sum_{l\in\mathcal{I}} \lambda_l \sqrt{\mathbb{E}\left[\langle \phi_l, X_i\rangle^4\right]} \sqrt{\mathbb{E}\left[\langle \beta^* - \beta_i, X_i\rangle^4\right]} \\
&\leq c \sum_{l\in\mathcal{I}} \lambda_l \mathbb{E}\left[\langle \phi_l, X_i\rangle^2\right] \mathbb{E}\mathbb{E}_{X_i}\left[\langle \beta^* - \beta_i, X_i\rangle^2\right] \\
&= c\mathbb{E}\left[\|L_K^{1/2}X_i\|^2\right] \mathbb{E}\left[\|L_C^{1/2}(\beta^* - \beta_i)\|^2\right].
\end{aligned}
$$

Recall that $\mathbb{E}\left[\|L_K^{1/2}X_i\|^2\right] = \int_\mathcal{T}\int_\mathcal{T} K(s,t)C(s,t)dsdt \leq \kappa_1^2\kappa_2^2$ and $\mathbb{E}\left[\|L_C^{1/2}(\beta^* - \beta_i)\|^2\right] = \mathbb{E}[\mathcal{E}(\hat{\eta}_i)]$. So,

$$
\mathbb{E}\left[\|\tilde{B}_i\|^2\right] \leq \kappa_1^2\kappa_2^2(c\mathbb{E}[\mathcal{E}(\hat{\eta}_i)] + \sigma^2). \tag{19}
$$

To continue the estimation in (18), we define $\mathscr{M} = L_K^{1/2} L_C L_K^{1/2}$. Then $\mathscr{M}$ is also a compact positive semi-definite operator on $L^2(\mathcal{T})$ with $\|\mathscr{M}\|_{\mathsf{op}} \leq \kappa_1^2\kappa_2^2$. Simple calculation shows that $B_i = L_C^{1/2} L_K^{1/2} \tilde{B}_i$ and

$$
\begin{aligned}
\left\|\left[\prod_{j=i+1}^{n} (I - \gamma_j \mathscr{L})\right] B_i\right\|^2 &= \left\|L_C^{1/2} L_K^{1/2} \left[\prod_{j=i+1}^{n} (I - \gamma_j \mathscr{M})\right] \tilde{B}_i\right\| \\
&\leq \left\|\mathscr{M}^{1/2} \prod_{j=i+1}^{n} (I - \gamma_j \mathscr{M})\right\|_{\mathsf{op}}^2 \|\tilde{B}_i\|^2.
\end{aligned}
$$

By (19) and Lemma 2 with $\theta = 1/2$,

$$
\begin{aligned}
J_2 &\leq \sum_{i=1}^{n} \gamma_i^2 \frac{(2e)^{-1} + \kappa_1^2\kappa_2^2}{1 + \sum_{j=i+1}^{n} \gamma_j} \mathbb{E}[\|\tilde{B}_i\|^2] \\
&\leq (1+c)(1+\kappa_1^2\kappa_2^2)^2 \sum_{i=1}^{n} \frac{\gamma_i^2(\mathbb{E}[\mathcal{E}(\hat{\eta}_i)] + \sigma^2)}{1 + \sum_{j=i+1}^{n} \gamma_j},
\end{aligned}
$$

which, together with the estimation $J_1 = 0$ and the expansion (17), proves (15).

With the further assumptions $L_C^{1/2}\beta^* = \mathscr{L}^\theta g^*$ for $\theta > 0$, we estimate J_3 by Lemma 2.

$$
J_3 \leq \left\|\mathscr{L}^\theta \prod_{i=1}^{n} (I - \gamma_i \mathscr{L})\right\|_{\mathsf{op}}^2 \|g^*\|^2 \leq \frac{(\theta/e)^{2\theta} + (\kappa_1\kappa_2)^{4\theta}}{1 + \left(\sum_{j=1}^{n} \gamma_j\right)^{2\theta}} \|g^*\|^2.
$$

The proof is complete. $\square$

Lemma 3. *Let $b \geq 2$, $0 < \mu < 1$, and $a > b^{1-\mu}$. One has*

$$\int_1^b \frac{x^{-2\mu}dx}{a - x^{1-\mu}} \leq \frac{(b/2)^{1-2\mu} - 1}{(1-2\mu)(a - (b/2)^{1-\mu})} + \frac{(b/2)^{-\mu}}{1-\mu} \log \frac{a - (b/2)^{1-\mu}}{a - b^{1-\mu}}, \tag{20}$$

where for the simplicity of notation, at $\mu = 1/2$, the factor $\frac{(b/2)^{1-2\mu}-1}{1-2\mu}$ denotes its limit $\log(b/2)$.

Proof. The estimate is done by separating the integral interval into $[1, b/2]$ and $[b/2, b]$. For the first half,

$$\int_1^{b/2} \frac{x^{-2\mu}dx}{a - x^{1-\mu}} \leq \frac{1}{a - (b/2)^{1-\mu}} \int_1^{b/2} x^{-2\mu}dx = \frac{(b/2)^{1-2\mu} - 1}{(1-2\mu)(a - (b/2)^{1-\mu})}.$$

For the second half, one has

$$\int_{b/2}^b \frac{x^{-2\mu}dx}{a - x^{1-\mu}} \leq \left(\frac{b}{2}\right)^{-\mu} \int_{b/2}^b \frac{x^{-\mu}dx}{a - x^{1-\mu}}$$
$$= \left(\frac{b}{2}\right)^{-\mu} \int_{b/2}^b \frac{\frac{-1}{1-\mu}d(a - x^{1-\mu})}{a - x^{1-\mu}} = \frac{(b/2)^{-\mu}}{1-\mu} \log \frac{a - (b/2)^{1-\mu}}{a - b^{1-\mu}}.$$

□

The following lemma appears a few times in the literature of online learning theory [22, 12, 11]. We include the proof for the sake of completeness.

Lemma 4. *Let $0 < \mu < 1$ and $0 < \gamma_1 \leq 1$. If the step-sizes $\gamma_i = \gamma_1 i^{-\mu}$ for $i \geq 2$, we have for any integer $l \geq 1$,*

$$\sum_{i=1}^{l} \frac{\gamma_i^2}{1 + \sum_{j=i+1}^{l} \gamma_j} \leq C_\mu \gamma_1 \begin{cases} l^{-\mu} \log(l+1), & \text{for } 0 < \mu \leq 1/2, \\ l^{-(1-\mu)}, & \text{for } 1/2 < \mu < 1, \end{cases} \tag{21}$$

where C_μ is a constant only depending on μ and it will be specified in the proof. Consequently, we also have a coarser constant bound

$$\sum_{i=1}^{l} \frac{\gamma_i^2}{1 + \sum_{j=i+1}^{l} \gamma_j} \leq 2^\mu C_\mu \gamma_1 / \mu. \tag{22}$$

Proof. The case $l = 1$ is obvious, where the left-hand side of (21) is γ_1^2, and we only need to set $C_\mu \geq \frac{1}{\log 2}$. Now we assume $l \geq 2$. Note that $i \geq (i+2)/3$ for any $i \geq 1$. We have

$$\begin{aligned}
\sum_{i=1}^{l} \frac{\gamma_i^2}{1 + \sum_{j=i+1}^{l} \gamma_j} =& \gamma_1^2 l^{-2\mu} + \gamma_1 \sum_{i=1}^{l-1} \frac{i^{-2\mu}}{\frac{1}{\gamma_1} + \sum_{j=i+1}^{l} j^{-\mu}} \\
\leq& \gamma_1^2 l^{-2\mu} + \gamma_1 \sum_{i=1}^{l-1} \frac{3^{2\mu}(i+2)^{-2\mu}}{\frac{1}{\gamma_1} + \frac{1}{1-\mu}\left[(l+1)^{1-\mu} - (i+1)^{1-\mu}\right]} \\
\leq& \gamma_1^2 l^{-2\mu} + 9^\mu \gamma_1 \int_1^l \frac{(x+1)^{-2\mu}dx}{\frac{1}{\gamma_1} + \frac{1}{1-\mu}\left[(l+1)^{1-\mu} - (x+1)^{1-\mu}\right]} \\
=& \gamma_1^2 l^{-2\mu} + 9^\mu \gamma_1 (1-\mu) \int_2^{l+1} \frac{x^{-2\mu}dx}{\frac{1-\mu}{\gamma_1} + (l+1)^{1-\mu} - x^{1-\mu}}.
\end{aligned}$$

We apply Lemma 3 to continue the estimation.

$$
\begin{aligned}
\sum_{i=1}^{l} \frac{\gamma_i^2}{1+\sum_{j=i+1}^{l}\gamma_j} \leq & \gamma_1^2 l^{-2\mu} + \frac{9^\mu \gamma_1 (1-\mu)}{\frac{1-\mu}{\gamma_1} + (l+1)^{1-\mu} - \left(\frac{l+1}{2}\right)^{1-\mu}} \times \frac{\left(\frac{l+1}{2}\right)^{1-2\mu} - 1}{1-2\mu} \\
& + 9^\mu \gamma_1 \left(\frac{l+1}{2}\right)^{-\mu} \log \frac{\frac{1-\mu}{\gamma_1} + (l+1)^{1-\mu} - \left(\frac{l+1}{2}\right)^{1-\mu}}{\frac{1-\mu}{\gamma_1}} \\
=: & \gamma_1^2 l^{-2\mu} + J_1' + J_2'.
\end{aligned}
$$

Below we estimate J_1' and J_2'. First,

$$
\begin{aligned}
J_1' \leq & \frac{9^\mu \gamma_1 (1-\mu)}{1-(1/2)^{1-\mu}} (l+1)^{\mu-1} \begin{cases} \frac{(1/2)^{1-2\mu}}{1-2\mu}(l+1)^{1-2\mu}, & \text{when } 0<\mu<1/2, \\ \log\frac{l+1}{2}, & \text{when } \mu = 1/2, \\ \frac{1}{2\mu-1}, & \text{when } 1/2<\mu<1, \end{cases} \\
\leq & C_{\mu,1}\gamma_1 \begin{cases} (l+1)^{-\mu}, & \text{when } 0<\mu<1/2, \\ \frac{\log(l+1)}{\sqrt{l+1}}, & \text{when } \mu = 1/2, \\ (l+1)^{\mu-1}, & \text{when } 1/2<\mu<1, \end{cases}
\end{aligned}
$$

where

$$
C_{\mu,1} = \frac{9^\mu (1-\mu)}{1-(1/2)^{1-\mu}} \begin{cases} \frac{(1/2)^{1-2\mu}}{1-2\mu}, & \text{when } 0<\mu<1/2, \\ 1, & \text{when } \mu = 1/2, \\ \frac{1}{2\mu-1}, & \text{when } 1/2<\mu<1. \end{cases}
$$

For J_2', recall that $\log(l+1) \geq \log 3 \geq 1$ and $\gamma_1 \leq 1$. we have

$$
\begin{aligned}
J_2' \leq & 18^\mu \gamma_1 (l+1)^{-\mu} \log\left[1 + \frac{\gamma_1 (l+1)^{1-\mu}}{1-\mu}(1-(\frac{1}{2})^{1-\mu})\right] \\
\leq & 18^\mu \gamma_1 \left[1-\mu + \log\left(1 + \frac{\gamma_1}{1-\mu}(1-(\frac{1}{2})^{1-\mu})\right)\right](l+1)^{-\mu}\log(l+1) \\
\leq & C_{\mu,2}\gamma_1 (l+1)^{-\mu}\log(l+1),
\end{aligned}
$$

where $C_{\mu,2} = 18^\mu \left(1-\mu+\log\left[1+\frac{1-(1/2)^{1-\mu}}{1-\mu}\right]\right)$. So, when $0<\mu\leq 1/2$, (21) is proved by defining $C_\mu = 1 + C_{\mu,1} + C_{\mu,2}$.

Simple calculation shows that

$$
\max_{1\leq x<\infty} x^{-\mu}\log x = \frac{1}{e\mu}, \quad \text{for any } \mu>0, \tag{23}
$$

where the maximum is achieved at $x = e^{1/\mu}$. Therefore, when $1/2<\mu<1$, $(l+1)^{-\mu}\log(l+1) \leq \frac{1}{e(2\mu-1)}(l+1)^{-1+\mu}$, and (21) is verified by defining $C_\mu = 1 + C_{\mu,1} + \frac{C_{\mu,2}}{e(2\mu-1)}$. Also, from (23) we see that for $0<\mu\leq 1/2$, $l^{-\mu}\log(l+1) \leq 2^\mu (l+1)^{-\mu}\log(l+1) \leq 2^\mu/\mu$, and when $1/2<\mu<1$, $l^{-(1-\mu)} \leq 1 < 2^\mu/\mu$. We have proved (22). □

Without using any specific form of the step-sizes, the lemma below gives a uniform rough estimation on error, which would be used later for deriving finer bounds.

Lemma 5. *Let $\{\beta_k : k \geq 1\}$ be defined by (4), and $N \leq \infty$. Assume (A2). Suppose for all $1 \leq k < N$,*

$$\gamma_k \kappa_1^2 \kappa_2^2 \leq 1, \quad \textit{and}$$
$$\sum_{i=1}^{k} \frac{\gamma_i^2}{1 + \sum_{j=i+1}^{k} \gamma_j} \leq \frac{1}{2(1+c)(1+\kappa_1^2\kappa_2^2)^2}.$$

Then we have

$$\mathbb{E}\left[\mathcal{E}(\hat{\eta}_{k+1})\right] \leq 2\kappa_2^2 \left\|\beta^*\right\|^2 + \sigma^2, \quad \textit{for all } 0 \leq k < N. \tag{24}$$

Proof. We organize the proof with mathematical induction. For $k = 1$, recall $\beta_1 = 0$. So, $\hat{\eta}_1 = 0$. One has

$$\begin{aligned}\mathbb{E}[\mathcal{E}(\hat{\eta}_1)] = \mathbb{E}[\eta^*(X)^2] = \mathbb{E}\left[\left(\int_{\mathcal{T}} \beta^*(s) X(s) ds\right)^2\right] \\ \leq \|\beta^*\|^2 \mathbb{E}\left[\int_{\mathcal{T}} X^2(s) ds\right] \leq \kappa_2^2 \|\beta^*\|^2.\end{aligned}$$

Now assume that (24) holds true for any $k = 1, \ldots, l-1$, with $l < N$. Below we prove that (24) also holds true for $k = l$. In fact, from Theorem 4 and Lemma 2,

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}(\hat{\eta}_{l+1})\right] \leq & \left\|\prod_{i=1}^{l}(I - \gamma_i \mathscr{L})\right\|_{\mathsf{op}}^2 \left\|L_C^{1/2}\right\|_{\mathsf{op}}^2 \left\|\beta^*\right\|^2 \\ & + (1+c)(1+\kappa_1^2\kappa_2^2)^2 \left(\sum_{i=1}^{l} \frac{\gamma_i^2}{1 + \sum_{j=i+1}^{l} \gamma_j}\right) \max_{1 \leq j \leq l} \left(\mathbb{E}\left[\mathcal{E}(\hat{\eta}_j)\right] + \sigma^2\right) \\ \leq & \kappa_2^2 \left\|\beta^*\right\|^2 + \frac{1}{2}\left(2\kappa_2^2 \left\|\beta^*\right\|^2 + 2\sigma^2\right) \\ = & 2\kappa_2^2 \left\|\beta^*\right\|^2 + \sigma^2.\end{aligned}$$

The proof is complete. $\square$

Proof of Theorem 1. We write the two terms at the right-hand side of (16) as J_1^* and J_2^*, respectively. By Assumption (A1), there exists some $g^* \in L^2(\mathcal{T})$, such that $L_C^{1/2}\beta^* = \mathscr{L}^\theta g^*$. We denote $\gamma_i = \gamma_1 i^{-\mu}$ to have

$$\begin{aligned}\sum_{j=1}^{n} \gamma_j \geq \gamma_1 \int_1^{n+1} x^{-\mu} dx = \frac{\gamma_1}{1-\mu}\left[(n+1)^{1-\mu} - 1\right] \\ \geq \frac{\gamma_1}{1-\mu}(1 - 2^{\mu-1})(n+1)^{1-\mu}.\end{aligned}$$

So,

$$J_1^* \leq \frac{C_1^*}{\gamma_1^{2\theta}(n+1)^{2\theta(1-\mu)}}, \quad \text{with } C_1^* = \frac{\left[(\theta/e)^{2\theta} + (\kappa_1\kappa_2)^{4\theta}\right] \left\|g^*\right\|^2}{\left[(1-2^{\mu-1})/(1-\mu)\right]^{2\theta}}.$$

By the setting

$$\gamma_1 \leq \frac{\mu}{2^{1+\mu}(1+c)(1+\kappa_1^2\kappa_2^2)^2 C_\mu},$$

Bound (22) guarantees that

$$\sum_{i=1}^{k} \frac{\gamma_i^2}{1+\sum_{j=i+1}^{k}\gamma_j} \leq \frac{1}{2(1+c)(1+\kappa_1^2\kappa_2^2)^2}$$

for $k \geq 1$. We use Lemma 5 to obtain $\mathbb{E}[\mathcal{E}(\hat{\eta}_k)] \leq 2\kappa_2^2 \|\beta^*\|^2 + \sigma^2$ for any $k \geq 1$. Recall $\mu \leq 1/2$. We apply (21) to obtain

$$\begin{aligned} J_2^* &\leq (1+c)(1+\kappa_1^2\kappa_2^2)^2(2\kappa_2^2\|\beta^*\|^2 + 2\sigma^2)\sum_{i=1}^{n}\frac{\gamma_i^2}{1+\sum_{j=i+1}^{n}\gamma_j} \\ &\leq 2^{-\mu}\mu(\kappa_2^2\|\beta^*\|^2+\sigma^2)n^{-\mu}\log(n+1). \end{aligned}$$

To complete the proof, we set μ as (7) and let

$$C_1 = \frac{C_1^*}{\gamma_1^{2\theta}} + 2^{-\mu}\mu(\kappa_2^2\|\beta^*\|^2+\sigma^2).$$

□

Proof of Theorem 2. We write the two terms at the right-hand side of (16) as J_1^* and J_2^*, respectively. By the setting of step-size, $\gamma_i \equiv \gamma = \gamma_0 n^{-2\theta/(2\theta+1)}$,

$$J_1^* \leq \frac{\left[(\theta/e)^{2\theta} + (\kappa_1\kappa_2)^{4\theta}\right]\|g^*\|^2}{\gamma_0^{2\theta}} n^{-2\theta/(2\theta+1)}.$$

For any $1 \leq k \leq n$,

$$\begin{aligned} \sum_{i=1}^{k}\frac{\gamma_i^2}{1+\sum_{j=i+1}^{k}\gamma_j} &= \sum_{i=1}^{k}\frac{\gamma^2}{1+(k-i)\gamma} = \gamma^2 + \sum_{i=1}^{k-1}\frac{\gamma^2}{1+i\gamma} \\ &\leq \gamma^2 + \gamma\int_0^{k-1}\frac{\gamma dx}{1+x\gamma} = \gamma^2 + \gamma\log(1+(k-1)\gamma). \end{aligned} \tag{25}$$

Recall that $0 < \gamma_0 \leq \left[2(1+c)(1+\kappa_1^2\kappa_2^2)^2(1+(2\theta+1)/(2e\theta))\right]^{-1} < 1$, so $\gamma < 1$. We use (23) to have

$$\gamma\log(1+(k-1)\gamma) \leq \gamma_0 n^{-\frac{2\theta}{2\theta+1}}\log n \leq \gamma_0\frac{2\theta+1}{2e\theta}.$$

So, for any $1 \leq k \leq n$,

$$\sum_{i=1}^{k}\frac{\gamma_i^2}{1+\sum_{j=i+1}^{k}\gamma_j} \leq \gamma_0 + \gamma_0\frac{2\theta+1}{2e\theta} \leq \frac{1}{2(1+c)(1+\kappa_1^2\kappa_2^2)^2}.$$

Also, obviously $\gamma_0\kappa_1^2\kappa_2^2 \le 1$. So by Lemma 5, for any $1 \le k \le n$, $\mathbb{E}[\mathcal{E}(\hat{\eta}_k)] \le 2\kappa_2^2 \|\beta^*\|^2+\sigma^2$. Now we use (25) to have

$$\sum_{i=1}^{n} \frac{\gamma_i^2}{1+\sum_{j=i+1}^{n}\gamma_j} \le \gamma + \gamma\log(n+1) \le \left(\frac{1}{\log 2}+1\right)\gamma\log(n+1).$$

So, we can bound J_2^* as

$$\begin{aligned} J_2^* &\le (1+c)(1+\kappa_1^2\kappa_2^2)^2 \left(\sum_{i=1}^{n} \frac{\gamma_i^2}{1+\sum_{j=i+1}^{n}\gamma_j}\right)\left(2\kappa_2^2\|\beta^*\|^2+\sigma^2\right) \\ &\le \frac{\left(1+\frac{1}{\log 2}\right)\left(2\kappa_2^2\|\beta^*\|^2+\sigma^2\right)}{2\left(1+\frac{2\theta+1}{2e\theta}\right)} n^{-\frac{2\theta}{2\theta+1}}\log(n+1). \end{aligned}$$

We specify C_2 below to complete the proof.

$$C_2 = \frac{[(\theta/e)^{2\theta}+(\kappa_1\kappa_2)^{4\theta}]\|g^*\|^2}{\gamma_0^{2\theta}\log 2} + \frac{\left(1+\frac{1}{\log 2}\right)\left(2\kappa_2^2\|\beta^*\|^2+\sigma^2\right)}{2\left(1+\frac{2\theta+1}{2e\theta}\right)}.$$

□

Acknowledgments

The work by Xiaming Chen is supported partially by Shantou University Scientific Research Start-up Fund Project (No. NTF18022). The work by Jun Fan is partially supported by the Research Grants Council of Hong Kong [Project No. HKBU 12303220] and National Natural Science Foundation of China [Project No. 11801478]. The work by Xin Guo is partially supported by the Research Grants Council of Hong Kong [Project No. PolyU 15304917] and The Hong Kong Polytechnic University [start-up ZE8Q]. The corresponding author is Jun Fan.

References

[1] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

[2] Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[3] Gilles Blanchard and Nicole Mücke. Kernel regression, minimax rates and effective dimensionality: beyond the regular case. *Anal. Appl. (Singap.)*, 18(4):683–696, 2020.

[4] Tony Cai and Peter Hall. Prediction in functional linear regression. *Ann. Statist.*, 34(5):2159–2179, 2006.

[5] Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.*, 107(499):1201–1216, 2012.

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

[7] Felipe Cucker and Ding-Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.

[8] Luc Devroye. Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(2):154–157, March 1982.

[9] Jun Fan, Fusheng Lv, and Lei Shi. An RKHS approach to estimate individualized treatment rules based on functional predictors. *Mathematical Foundations of Computing*, 2(2):169–181, 2019.

[10] Takayuki Furuta. $A \geq B \geq 0$ assures $(B^r A^p B^r)^{1/q} \geq B^{(p+2r)/q}$ for $r \geq 0$, $p \geq 0$, $q \geq 1$ with $(1+2r)q \geq p+2r$. *Proc. Amer. Math. Soc.*, 101(1):85–88, 1987.

[11] Xin Guo, Junhong Lin, and Ding-Xuan Zhou. Convergence of the randomized Kaczmarz algorithm in Hilbert spaces. Under review.

[12] Zheng-Chu Guo and Lei Shi. Fast and strong convergence of online learning algorithms. *Adv. Comput. Math.*, 45(5-6):2745–2770, 2019.

[13] Peter Hall and Joel Horowitz. Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35(1):70–91, 2007.

[14] Bingzheng Li and Zhengzhan Dai. Error analysis on regularized regression based on the maximum correntropy criterion. *Mathematical Foundations of Computing*, 3(1):25–40, 2020.

[15] Ying Lin, Rongrong Lin, and Qi Ye. Sparse regularized learning in the reproducing kernel banach spaces with the ℓ^1 norm. *Mathematical Foundations of Computing*, 3(3):205–218, 2020.

[16] Fusheng Lv and Jun Fan. Optimal learning with Gaussians and correntropy loss. *Anal. Appl. (Singap.)*, 19(1):107–124, 2021.

[17] Gert Pedersen. Some operator monotone functions. *Proc. Amer. Math. Soc.*, 36:309–310, 1972.

[18] Jim Ramsay and Bernard Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

[19] Lei Shi. Distributed learning with indefinite kernels. *Anal. Appl. (Singap.)*, 17(6):947–975, 2019.

[20] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[21] Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[22] Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Found. Comput. Math.*, 8(5):561–596, 2008.

[23] Ming Yuan and Tony Cai. A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.*, 38(6):3412–3444, 2010.