Capacity dependent analysis for functional online learning algorithms

Xin Guo[1], Zheng-Chu Guo[2], and Lei Shi[3]

[1]School of Mathematics and Physics, The University of Queensland, Brisbane, QLD 4072, Australia. xin.guo@uq.edu.au

[2]School of Mathematical Sciences, Zhejiang University, Hangzhou 310058, P. R. China. guozhengchu@zju.edu.cn

[3]School of Mathematical Sciences and Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, Shanghai 200433, and Shanghai Artificial Intelligence Laboratory, 701 Yunjin Road, Shanghai 200232, P. R. China. leishi@fudan.edu.cn

Abstract

This article provides convergence analysis of online stochastic gradient descent algorithms for functional linear models. Adopting the characterizations of the slope function regularity, the kernel space capacity, and the capacity of the sampling process covariance operator, significant improvement on the convergence rates is achieved. Both prediction problems and estimation problems are studied, where we show that capacity assumption can alleviate the saturation of the convergence rate as the regularity of the target function increases. We show that with properly selected kernel, capacity assumptions can fully compensate for the regularity assumptions for prediction problems (but not for estimation problems). This demonstrates the significant difference between the prediction problems and the estimation problems in functional data analysis.

Key words and phrases: Functional data analysis, Stochastic gradient decent, Reproducing kernel Hilbert space, Capacity dependent analysis

1 Introduction

In this paper, we consider a functional linear model

$$Y = \int_{\mathcal{T}} \beta^*(u)X(u)du + \varepsilon. \tag{1.1}$$

Here, $\mathcal{T}$ is a compact subset in a Euclidean space $\mathbb{R}^d$, X is a random function, β^* is an unknown slope function, ε is a centered random noise independent of X, with finite variance $\sigma^2 = \mathrm{Var}(\varepsilon) < \infty$, and $Y \in \mathbb{R}$ is the response. We write $(L^2(\mathcal{T}), \langle\cdot,\cdot\rangle_2, \|\cdot\|_2)$ the space of square integrable functions on $\mathcal{T}$, and assume $X, \beta^* \in L^2(\mathcal{T})$. Then, Model (1.1) can be equivalently written as $Y = \langle\beta^*, X\rangle_2 + \varepsilon$. Without loss of generality, we assume that $\mathcal{T} = [0,1]^d$ throughout the paper.

We study two kinds of learning problems for Model (1.1). The *estimation problem* asks one to recover the unknown slope function β^*, and the *prediction problem* asks one to recover the linear functional on $L^2(\mathcal{T})$, denoted by φ^*, which is given by

$$\varphi^* : f \mapsto \langle\beta^*, f\rangle_2 = \int_{\mathcal{T}} \beta^*(u)f(u)du. \tag{1.2}$$

Mathematically, φ^* is defined with β^*, which in turn is fully determined by φ^* through the Riesz representation theorem. Nonetheless, it is well understood that the two learning problems are different. In particular, the integral in (1.2) brings a smoothing effect, leading to a weaker regularity requirement for the prediction problems [4, 7].

Write $D = \{(x_t, y_t)\}_{t=1}^T$ a sample of independent copies of (X, Y) in Model (1.1). We study both the case of a finite sample $T < \infty$, and the case $T = \infty$ where D models an ongoing indefinite sampling process.

Both prediction and estimation problems can be solved by constructing an estimator $\hat{\beta}$ of the slope function β^*. In the literature, many works have been done on functional principal component analysis (FPCA) [22, 4, 16]. FPCA defines $\hat{\beta}$ with a linear combination of the estimated eigenfunctions of C, which is the covariance function of the random function X. Another approach of constructing $\hat{\beta}$ is the kernel method, which adopts a reproducing kernel K and represents $\hat{\beta}$ by the linear combination of kernel functions [27, 5].

We adopt the kernel method and define $\hat{\beta}$ through stochastic gradient descent approach in this paper. A reproducing kernel K on $\mathcal{T}$ is defined as a function $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ that is symmetric (i.e. $K(u,v) = K(v,u)$ for any $u, v \in \mathcal{T}$) and positive semi-definite (which

requires that the Gram matrix $(K(u_i, u_j))_{i,j=1}^n$ is positive semi-definite for any $n \geq 1$ and any $u_1, \ldots, u_n \in \mathcal{T}$). We further assume that K is continuous, exclude the trivial case $K \equiv 0$, and let $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K, \|\cdot\|_K)$ denote the reproducing kernel Hilbert space (RKHS) associated with K [8, 23]. The stochastic gradient descent algorithm defines a sequence $\{\hat{\beta}_t\}$ of estimators, from $\hat{\beta}_1 = 0$ and then iteratively by

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \eta_t \left(\int_{\mathcal{T}} \hat{\beta}_t(u) x_t(u) du - y_t \right) \int_{\mathcal{T}} K(v, \cdot) x_t(v) dv, \quad \text{for } t \geq 1. \tag{1.3}$$

Here $\eta_t > 0$ is the step-size. Based on the nature of the sample D, we study two settings of the step-sizes $\{\eta_t\}$.

- The *online* setting. We write $|D| = \infty$ and use D to model the outcome of an ongoing and indefinite sampling process. The estimator $\hat{\beta}$ is being updated following the sampling process. For example, we update the estimator to $\hat{\beta} = \hat{\beta}_{t+1}$ after t steps of iterations and before the observation (x_{t+1}, y_{t+1}) is available. For the online setting, the step-sizes $\{\eta_t\}$ are designed to decrease, rendering Algorithm (1.3) more and more conservative against the possible random noise brought by new observations.

- The *finite-horizon* setting. We assume a finite sample D with size $|D| = T < \infty$. A constant step-size $\eta_t \equiv \eta = \eta(T)$ is adopted throughout the iterations (1.3) with $t = 1, \ldots, T$. The sample D is then exhausted and we use $\hat{\beta} = \hat{\beta}_{T+1}$ as the derived estimator. The step-size $\eta(T)$ can be optimized (at least asymptotically) over T, but it could be not trivial later to warm-start the iteration efficiently when new sample points are available.

To measure the estimation performance of $\hat{\beta}$, we use the expected squared $\mathcal{H}_K$ norm $\mathbb{E}[\|\hat{\beta} - \beta^*\|_K^2]$. Write $\hat{\varphi} : L^2(\mathcal{T}) \to L^2(\mathcal{T})$ the estimator of the functional φ^*,

$$\hat{\varphi} : f \mapsto \langle \hat{\beta}, f \rangle_2 = \int_{\mathcal{T}} \hat{\beta}(u) f(u) du.$$

The prediction performance of $\hat{\varphi}$ is measured by the expected excess generalization error $\mathbb{E}[\mathcal{E}(\hat{\varphi})]$. Here for any linear functional φ on $L^2(\mathcal{T})$,

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

where the expectation is taken with respect to the distribution of (X, Y) in Model (1.1).

As a technical instrument, the integral operator $L_K : L^2(\mathcal{T}) \to L^2(\mathcal{T})$ is defined with the reproducing kernel K, by

$$L_K(f) = \int_{\mathcal{T}} K(\cdot, u) f(u) du. \tag{1.4}$$

It is well understood in the literature that L_K is positive semi-definite (thus self-adjoint), and of trace class (so, compact). See, e.g., [23, Theorem 4.27]. The power L_K^r with $r \in (0, \infty)$ is well defined by the spectral theorem. In terms of L_K, the iteration (1.3) is equivalently written as

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \eta_t \left(\langle \hat{\beta}_t, x_t \rangle_2 - y_t \right) L_K x_t, \quad \text{for } t \geq 1. \tag{1.5}$$

For the sake of simplicity we assume $\mathbb{E}[X] = 0$ and $\|X\|_2 = 1$ a.s. Consequently, $\mathbb{E}[Y] = 0$. The covariance function C has the form

$$C(u, v) = \mathbb{E}[X(u)X(v)], \quad \text{for } u, v \in \mathcal{T}.$$

Obviously C is also a reproducing kernel. We further assume that C is continuous, exclude the trivial case $C \equiv 0$, and define the operator L_C on $L^2(\mathcal{T})$ in the same way as (1.4) by substituting K with C. So, L_C is self-adjoint, positive semi-definite, of trace class, and thus compact. The power L_C^r with $r > 0$ is well defined. For any $f, g, h \in L^2(\mathcal{T})$, we define $f \otimes g$ as a rank-one operator on $L^2(\mathcal{T})$ defined by $(f \otimes g)h = \langle g, h \rangle_2 f$. For any linear functional $\varphi(\cdot) = \langle \beta, \cdot \rangle_2$ on $L^2(\mathcal{T})$, the excess generalization error can be written in terms of the norm of $L^2(\mathcal{T})$,

$$\begin{aligned} \mathcal{E}(\varphi) &= \mathbb{E}\left[(Y - \langle \beta, X \rangle_2)^2 - (Y - \langle \beta^*, X \rangle_2)^2 \right] \\ &= \mathbb{E}\left[\langle \beta - \beta^*, X \rangle_2^2 \right] = \mathbb{E}[\langle \beta - \beta^*, X \otimes X(\beta - \beta^*) \rangle_2] \\ &= \| L_C^{1/2}(\beta - \beta^*) \|_2^2. \end{aligned} \tag{1.6}$$

Since $\mathcal{T}$ is compact, we write

$$\kappa = \max_{u \in \mathcal{T}} \sqrt{K(u, u)} \in (0, \infty).$$

Recall that by the positive semi-definiteness, $|K(u, v)| \leq \sqrt{K(u, u)K(v, v)} \leq \kappa^2$ for any $u, v \in \mathcal{T}$. The spectral norm of L_K is bounded by $\|L_K\|_{\mathsf{op}(L^2)} \leq \kappa^2$. For the sake of simplicity we assume $\|L_C\|_{\mathsf{op}(L^2)} \leq 1$.

Modern scalable computing and stochastic optimization techniques make stochastic gradient descent a popular approach across various applications. Theoretical analysis of its convergence is recently extensively studied. The present work aims to establish a novel capacity-dependent convergence analysis for stochastic gradient descent (1.3) which is applied to solve the linear functional model (1.1) in an RKHS. We study prediction problem through the convergence of excess generalization error (1.6) and estimation problem through the strong convergence in an RKHS. Our analysis developed in this paper leads to fast rates for both types of convergence. State-of-the-art convergence rates in RKHS metric are obtained. Our error estimates fully exploit the spectral structure of the operators and the capacity condition encoding the smoothness of kernels and covariance function. Our work provides insights for the applications of kernel methods to functional data analysis, and better understanding of the difference between the estimation problems and the prediction problems in functional linear models.

The remaining sections of the paper are organized as follows. In Section 2, we provide convergence analysis of Algorithm (1.3) with explicit rates, for both excess generalization error and estimation error. Section 3 provides comprehensive discussion on our main assumptions, and literature discussions. Detailed error analysis is put in Sections 4, 5, and 6. In Section 7 we give a simple numerical experiment with an example used in [5].

2 Main Results

In this section we list some main assumptions and present the convergence rates of the stochastic gradient descent algorithm (1.3), in the finite-horizon and online settings, respectively. We provide discussions of the assumptions in Section 3.

Denote $\mathscr{L}_K = L_C^{1/2} L_K L_C^{1/2}$ and $\mathscr{L}_C = L_K^{1/2} L_C L_K^{1/2}$. It is easy to verify that both of the operators $\mathscr{L}_K$ and $\mathscr{L}_C$ are self-adjoint, positive semi-definite, of trace class, and compact.

Assumption 1 (Regularity Condition of the slope β^*)**.** *There exists some g^* in $L^2(\mathcal{T})$ and $r \in (0, \infty)$ such that*

$$L_C^{1/2} \beta^* = \mathscr{L}_K^r g^*.$$

One can see Section 2 and 3 of [20] for more illustrations, and Theorem 3 in [7] for a more insightful description of Assumption 1. In particular, we have Remark 2 in Section 3 that provides some insights.

For any positive semi-definite compact operator L, let $\mathrm{Tr}(L)$ denote the trace of L, i.e., the sum of all the positive eigenvalues (counting multiplicity) of L. In particular, $\mathrm{Tr}(L) < \infty$ if and only if L is of trace class.

Assumption 2 (Capacity Condition)**.**

$$\mathrm{Tr}(\mathscr{L}_K^s) < \infty, \quad \textit{for some } 0 < s \leq 1.$$

Note that since $\mathscr{L}_K$ is a trace-class operator, Assumption 2 with $s = 1$ holds true automatically.

Assumption 3 (Moment Condition)**.** *For Model (1.1), there exist a constant $c_{\mathsf{M}} > 0$ such that for any f in $L^2(\mathcal{T})$,*

$$\mathbb{E}[\langle X, f\rangle_2^4] \leq c_{\mathsf{M}} \left(\mathbb{E}[\langle X, f\rangle_2^2]\right)^2. \tag{2.1}$$

2.1 Analysis of the Prediction Error

In this subsection, we study the estimator $\hat{\varphi} = \langle \hat{\beta}, \cdot \rangle_2$ for the prediction problem and bound the expected excess generalization error.

Theorem 1. *In the online setting, define $\{\hat{\varphi}_t = \langle \hat{\beta}_t, \cdot \rangle_2\}$ through (1.3). Under Assumptions 1 (with $r > 0$), 2 (with $0 < s \leq 1$), and 3, set $\eta_t = \eta_0 t^{-\theta}$ with*

$$\theta = \frac{\min\{2r, 2-s\}}{1 + \min\{2r, 2-s\}} = \begin{cases} \frac{2r}{2r+1}, & \textit{when } 2r \leq 2-s, \\ \frac{2-s}{3-s}, & \textit{when } 2r \geq 2-s. \end{cases} \tag{2.2}$$

If $0 < \eta_0 \leq \min\{1, \kappa^{-2}, C_1^{\mathsf{S}}\}$ (where C_1^{S} is a constant, and it will be specified by (5.20) in the proof), then

$$\mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})] \leq C_1 \begin{cases} (t+1)^{-\theta}, & 0 < s < 1, \\ (t+1)^{-\theta} \log(t+1), & s = 1, \end{cases} \quad \textit{for any } t \geq 1, \tag{2.3}$$

where C_1 is a constant independent of t, and it will be specified by (5.22) in the proof.

For the piecewise definition (2.2), we let the domains overlap on purpose to highlight the continuity of θ on the whole domain $r > 0$ and $0 < s \leq 1$. The index θ as a function of r and s is also visualized in Figure 1. Without Assumption 2 (i.e., case $s = 1$ in (2.3)), the convergence rate $O((t+1)^{-2r/(2r+1)} \log(t+1))$, saturated as $O((t+1)^{-1/2} \log(t+1))$

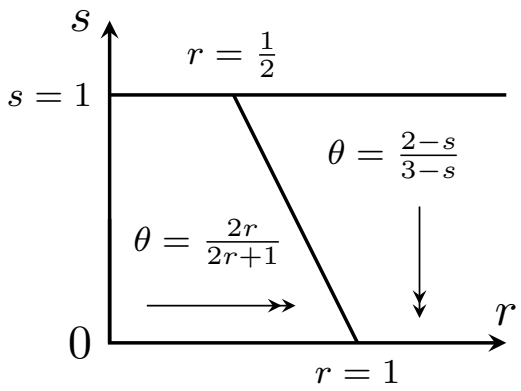

Figure 1: The index θ in Theorem 1 as a function of (r, s). Here the double-headed arrows show the gradient directions.

for $r \geq 1/2$, is also obtained in [7]. Here, r indicates the regularity of the target function β^* as described in Assumption 1. The saturation means beyond $r \in (0, 1/2]$, further improvement of such regularity (i.e., increasing of r) does not help to improve the rate $\mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})]$ converges to zero. In this paper, Theorem 1 suggests that Assumption 2 on capacity with $s < 1$, not only removes the logarithmic factor in the convergence rates, but also uplifts the saturating boundary from $1/2$ to $1/2 + (1 - s)/2$.

In the following, Theorem 2 shows that in the finite-horizon setting, Algorithm (1.3) does not suffer from the above discussed saturation, and the expected prediction error converges to zero in a rate arbitrarily close to $O(T^{-1})$, for sufficiently large r.

Theorem 2. *In the finite-horizon setting with $1 \leq T = |D| < \infty$, define $\hat{\varphi}_{T+1} = \langle \hat{\beta}_{T+1}, \cdot \rangle_2$ through (1.3). Under Assumptions 1 (with $r > 0$), 2 (with $0 < s \leq 1$), and 3, set the constant step-size $\eta_t = \eta_0 T^{-2r/(2r+1)}$, with $0 < \eta_0 \leq \min\{1, \kappa^{-2}, C_2^{\mathsf{S}}\}$ (where C_2^{S} is a constant independent of T, and it will be specified by (5.25) in the proof). Then,*

$$\mathbb{E}[\mathcal{E}(\hat{\varphi}_{T+1})] \leq C_2 \begin{cases} T^{-2r/(2r+1)}, & \text{when } 0 < s < 1, \\ T^{-2r/(2r+1)} \log(T+1), & \text{when } s = 1. \end{cases} \tag{2.4}$$

where the constant C_2 is independent of T, and it will be specified by (5.27) in the proof.

The capacity independent convergence rate $O(T^{-2r/(2r+1)} \log(T+1))$ for $s = 1$ in (2.4) is first derived in [7]. In the finite-horizon setting, the capacity assumption $0 < s < 1$ helps remove the logarithmic factor.

2.2 Analysis of the Estimation Error

In this subsection, we study the estimator $\hat{\beta}$ for the estimation problem. The analysis employs the following Assumption 4 to replace Assumption 1.

Assumption 4 (Regularity Condition of the slope β^*)**.** *There exists some $g^\dagger$ in $L^2(\mathcal{T})$ and $r > 0$, such that*

$$\beta^* = L_K^{1/2} \mathscr{L}_C^r g^\dagger.$$

This assumption implies that the slope β^* lies in the range of $L_K^{1/2}$, i.e., $\beta^* \in \mathcal{H}_K$.

Theorem 3. *In the online setting, define $\{\hat{\beta}_t\}_{t\geq 1}$ through (1.3). Under Assumptions 2 (with $0 < s < 1$), 3, and 4 (with $r > 0$), set step-sizes $\eta_t = \eta_0 t^{-\theta}$ with $0 < \eta_0 \leq \min\{1, \kappa^{-2}, C_3^{\mathsf{S}}\}$ (where C_3^{S} is a constant independent of t, and it will be specified by (6.6) in the proof), and*

$$\theta = \begin{cases} \frac{2r+s}{2r+s+1}, & \text{when } 2r \leq 1-s, \\ 1/2, & \text{when } 2r \geq 1-s. \end{cases} \tag{2.5}$$

Then,

$$\mathbb{E}[\|\hat{\beta}_{t+1} - \beta^*\|_K^2] \leq C_3 \begin{cases} (t+1)^{-2r/(1+s+2r)}, & 2r < 1-s, \\ (t+1)^{-(1-s)/2} \log(t+1), & 2r \geq 1-s, \end{cases} \quad \text{for any } t \geq 1, \tag{2.6}$$

where C_3 is a constant independent of t, and it will be specified by (6.10) in the proof.

In the definition (2.5), θ is a continuous function of $r > 0$ and $0 < s < 1$. So we purposely use two overlapping domains. The power index of the rates in (2.6) will be elucidated in Figure 2.

Remark 1. *Theorem 3 does not work in the capacity independent setting $s = 1$, where the convergence analysis remains an open problem.*

Next we establish unsaturated convergence rates of estimation error for the finite-horizon setting.

Theorem 4. *In the finite-horizon setting with $1 \leq T = |D| < \infty$, define $\{\hat{\beta}_t : 1 \leq t \leq T+1\}$ through (1.3). Under Assumptions 2 (with $0 < s \leq 1$), 3, and 4 (with $r > 0$), set*

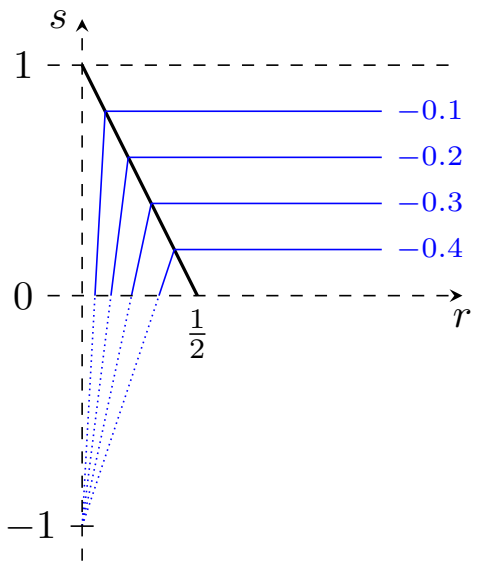

Figure 2: The power index $\tilde{\omega}(r,s)$ of the convergence estimate (2.6) in Theorem 3. The thick black line marks the boundary $2r = 1-s$ of the two regimes. $\tilde{\omega} = -2r/(1+s+2r)$ when $2r \le 1-s$, and $\tilde{\omega} = -(1-s)/2$ when $2r \ge 1-s$. Contours of $\tilde{\omega}$ are plotted in solid blue lines, and further extended by dotted lines.

the constant step-size $\eta_t = \eta_0 T^{-(s+2r)/(1+s+2r)}$, *with* $0 < \eta_0 \le \min\{1, \kappa^{-2}, C_4^{\mathsf{S}}\}$ *(where* C_4^{S} *is a constant independent of* T*, and it will be specified by (6.11) in the proof). Then,*

$$\mathbb{E}[\|\hat{\beta}_{T+1} - \beta^*\|_K^2] \le C_4 T^{-2r/(1+s+2r)}, \tag{2.7}$$

where the constant C_4 *is independent of* T*, and it will be specified by (6.16) in the proof.*

The rates (2.7) in Theorem 4 does not saturate for $r > 0$, and are arbitrarily close to $O(T^{-1})$ for sufficiently large r. For a fixed $r > 0$, with a smaller s one has a faster rate in (2.7). Here, a smaller s indicates a stronger capacity assumption (Assumption 2). As we shall see in Theorem 5 below, a smaller s corresponds to faster eigenvalue decay for $\mathscr{L}_K$ (equivalently, faster eigenvalue decay for $\mathscr{L}_C$), and a smaller hypothesis space $L_K^{1/2}\mathscr{L}_C^r(L^2(\mathcal{T}))$ in Assumption 4.

3 Comparisons and Discussions

There has been rapidly growing literature focusing on stochastic gradient descent and its variants in an RKHS or general Hilbert spaces [26, 9, 21, 18, 14, 2, 11, 12]. We refer the

readers to these papers and references therein. Our paper contributes to the theoretical analysis of functional linear regression in an RKHS that stems from the works of Yuan and Cai [27, 5] which establish capacity dependent analysis for batch learning. As far as we know, the convergence of stochastic gradient descent has not been investigated in the context of functional linear regression in an RKHS till the very recent paper [7] in which the authors conduct capacity independent analysis of the prediction error.

Under the batch learning setting, Yuan and Cai [27] derive the minimax optimal convergence rate $T^{-2s_*/(2s_*+1)}$ of the excess generalization error $\mathcal{E}(\hat{\varphi}_{T+1})$ for prediction, with the regularity assumption $\beta^* \in \mathcal{H}_K$ and capacity assumption on the rates of eigenvalue decay, $\lambda_i(L_K) \sim i^{-2s_1}$ (here $a_i \sim b_i$ means a_i/b_i being uniformly bounded away from zero and infinity as $i \to \infty$) and $\lambda_i(L_C) \sim i^{-2s_2}$, where $s_1, s_2 > 1/2$ and $s_* = s_1 + s_2$. Later, Cai and Yuan [5] derive the same rate with a different capacity assumption $\lambda_i(\mathscr{L}_C) \sim i^{-2s_*}$.

Compared with these works, the strength of our analysis includes that first, our Assumption 2 on capacity, $\mathrm{Tr}(\mathscr{L}_K^s) < \infty$, is way more general. We shall see in Theorem 5 that this is roughly equivalent to the assumption $\lambda_i(\mathscr{L}_K) = O(i^{-1/s})$. We shall see in Remark 3 that although the eigenvalues $\{\lambda_i(\mathscr{L}_K)\}_{i=1}^\infty$ are arranged non-increasingly, in general there is no exact index s_* such that $\lambda_i(\mathscr{L}_K) \sim i^{-2s_*}$ (same for other compact operators including L_K, L_C, and $\mathscr{L}_C$). Second, our analysis supports finer characterizations $L_C^{1/2}\beta^* = \mathscr{L}_K^r(g^*)$ (Assumption 1) and $\beta^* = L_K^{1/2}\mathscr{L}_C^r(g^\dagger)$ (Assumption 4) of slope function regularity. This leads to a better convergence rate $O(T^{-2r/(2r+1)})$ in Theorem 2, than $O(T^{-2s_*/(2s_*+1)})$ when $r > s_*$. Third, we proved the non-trivial convergence rates for the estimation error $\|\hat{\beta} - \beta^*\|_K^2$ in $\mathcal{H}_K$ metric, $O(T^{-2r/(2r+1)})$ (saturated at $r = (1-s)/2$) in the online setting in Theorem 3, and $O(T^{-2r/(1+s+2r)})$ in the finite-horizon setting in Theorem 4. Note that the analysis in [27] only provides a constant rate $O(1)$ for $\|\hat{\beta} - \beta^*\|_K^2$.

It is an interesting problem to replace Assumptions 1 and 4 by general forms of source conditions like $\phi(L_K, \mathscr{L}_C)g^*$ or $\phi(L_C, \mathscr{L}_K)g^*$ as studied in [1].

Next we provide some comments on the main assumptions in Section 2. For any bounded self-adjoint operators A and B on $L^2(\mathcal{T})$, we write $A \succeq B$ (or $B \preceq A$) if $A - B$ is positive semi-definite.

Remark 2. *It is well understood [7, Remark 2] that when $0 < r < 1/2$, if $L_K^\tau \succeq \delta L_C^\nu$ for some $\tau, \delta, \nu > 0$ with $\tau + \nu \geq 1$ and $r = \tau/(2\tau + 2\nu)$, then Assumption 1 is guaranteed by any $\beta^* \in L^2(\mathcal{T})$. That is, with a carefully selected reproducing kernel K, for the prediction error to converge, the capacity assumption (Assumption 2) can fully compensate for the regularity assumption (Assumption 1). Note that the above condition $L_K^\tau \succeq \delta L_C^\nu$ puts*

some requirement on the selection of the reproducing kernel K, but it does not require the one-to-one matching between the eigenfunctions of L_K and L_C, respectively.

Similarly, if $L_C \succeq \delta L_K^{\nu}$ for some $\delta, \nu > 0$, then Assumption 4 with $0 < r \leq 1/2$ is guaranteed when $\beta^* \in L_K^{r(1+\nu)+\frac{1}{2}}(L^2(\mathcal{T}))$. However, Assumption 4 implies $\beta^* \in L_K^{1/2}(L^2(\mathcal{T}))$. Therefore, the regularity assumption for the estimation error to converge, can not be fully compensated for by the capacity assumption. This demonstrates a significant difference between the prediction problems, and the estimation problems in functional data analysis.

In the literature of kernel-based learning algorithms [6, 1, 3, 19, 15, 13, 24, 17], the capacity of the hypothesis space $\mathcal{H}_K$ is usually measured by covering numbers, or the effective dimension [28] $\mathcal{N}_{L_K}(\lambda) = \mathrm{Tr}((L_K + \lambda I)^{-1} L_K)$, where I denotes the identity operator. A typical capacity assumption takes the form $\mathcal{N}_{L_K}(\lambda) = O(\lambda^{-s})$ (as $\lambda \downarrow 0$) for some $0 < s < 1$, and is well understood. The following theorem shows that roughly speaking, Assumption 2 with $0 < s < 1$ is comparable to the assumption $\mathcal{N}_{\mathscr{L}_K}(\lambda) = O(\lambda^{-s})$ as $\lambda \downarrow 0$. The conclusion is well understood [14, 11], but the proof through (3.1), is to our best knowledge not available elsewhere.

Theorem 5. *Let L be a positive semi-definite operator of trace class with infinite positive eigenvalues $\{\lambda_i = \lambda_i(L)\}_{i=1}^{\infty}$ arranged in non-increasing order. Let $0 < s < 1$. We have*

$$\mathrm{Tr}(L^s) = \frac{\sin(\pi s)}{\pi} \int_0^{\infty} \lambda^{s-1} \mathcal{N}_L(\lambda) d\lambda. \tag{3.1}$$

Consequently,

(a). If $\mathrm{Tr}(L^s) < \infty$, then $\mathcal{N}_L(\lambda) = O(\lambda^{-s})$ as $\lambda \downarrow 0$;

(b). If $\mathcal{N}_L(\lambda) = O(\lambda^{-s})$ as $\lambda \downarrow 0$, then $\mathrm{Tr}(L^{s+\epsilon}) < \infty$ for any $\epsilon > 0$;

(c). Moreover, for any fixed $0 < s < 1$, $\mathcal{N}_L(\lambda) = O(\lambda^{-s})$ as $\lambda \downarrow 0$ if and only if $\lambda_i = O(i^{-1/s})$ as $i \to \infty$.

The relations listed in Theorem 5 are summarized in Figure 3.

The case L has only finite positive eigenvalues is trivial, where $\mathcal{N}_L(\lambda) = O(1)$ as $\lambda \downarrow 0$ and $\mathrm{Tr}(L^s) < 0$ for any $s > 0$.

Note that on the one hand, when L^s does not belong to the trace class (equivalently, $\mathrm{Tr}(L^s) = \infty$), Equation (3.1) implies that the integral on its right-hand side diverges to infinity. On the other hand, when this integral diverges to infinity, $\mathrm{Tr}(L^s) = \infty$.

$$\begin{array}{c} \mathrm{Tr}(L^s) < \infty \\ \Downarrow \\ \mathcal{N}_L(\lambda) = O(\lambda^{-s}) \iff \lambda_i(L) = O(i^{-1/s}) \\ \Downarrow \\ \mathrm{Tr}(L^{s+\epsilon}) < \infty \end{array}$$

Figure 3: Summary of the relations listed in Theorem 5.

The bound $\mathcal{N}_L(\lambda) = O(\lambda^{-s})$ as $\lambda \downarrow 0$ does not guarantee $\mathrm{Tr}(L^s) < \infty$. For example, $\lambda_i = i^{-1/s}$ implies $\mathrm{Tr}(L^s) = \infty$, yet we still have

$$\mathcal{N}_L(\lambda) = \sum_{i=1}^{\infty} \frac{1}{1+\lambda i^{1/s}} \leq \int_0^{\infty} \frac{du}{1+\lambda u^{1/s}} = \lambda^{-s} \int_0^{\infty} \frac{du}{1+u^{1/s}} = O(\lambda^{-s}), \quad \text{as } \lambda \downarrow 0.$$

Remark 3. *Note that in general, for a non-increasing sequence $\{a_k\}_{k=1}^{\infty} \subset (0,\infty)$, Theorem 5 does not suggest the existence of some $\gamma > 0$ such that $a_k \sim k^{-\gamma}$. It is easy to construct a non-increasing sequence that stays between $k^{-\gamma_1}$ and $k^{-\gamma_2}$ for any $\gamma_1 > \gamma_2 > 0$. To this end, we define $\{b_k\}$ as $b_1 = 2$ and $b_{k+1} = b_k^{\gamma_1/\gamma_2}$ for $k \geq 1$. We define a function f on $[2,\infty)$, piece-wisely by $f(x) = b_k^{-\gamma_1}$ for $b_k \leq x < b_{k+1}$. Writing $a_k = f(k+1)$ to give*

$$\limsup_{k\to\infty} \frac{a_k}{k^{-\gamma_2}} = \liminf_{k\to\infty} \frac{a_k}{k^{-\gamma_1}} = 1.$$

Proof of Theorem 5. Write $B(u,v)$ the Euler beta function for $u, v > 0$. Recall that for any $a > 0$,

$$\frac{\pi}{\sin(\pi s)} = B(s, 1-s) = \int_0^{\infty} \frac{\xi^{s-1}}{1+\xi} d\xi \overset{\xi=\lambda/a}{=\!=\!=} a^{-s} \int_0^{\infty} \frac{a\lambda^{s-1}}{a+\lambda} d\lambda.$$

So,

$$a^s = \frac{\sin(\pi s)}{\pi} \int_0^{\infty} \lambda^{s-1} \frac{a}{a+\lambda} d\lambda. \tag{3.2}$$

In (3.2) substitute a with all the positive eigenvalues of L respectively, and take the sum to obtain (3.1). Since L is of trace class, $\mathcal{N}_L(\lambda)$ is well defined for each $\lambda > 0$. Obviously λ^{s-1} and $\mathcal{N}_L(\lambda)$ are non-increasing. So when $\mathrm{Tr}(L^s) < \infty$,

$$\lambda^s \mathcal{N}_L(\lambda) = \lambda^{s-1} \mathcal{N}_L(\lambda) \int_0^{\lambda} d\xi < \int_0^{\infty} \xi^{s-1} \mathcal{N}_L(\xi) d\xi = \frac{\pi \mathrm{Tr}(L^s)}{\sin(\pi s)} < \infty,$$

which verifies (a). Now assume $\mathcal{N}_L(\lambda) = O(\lambda^{-s})$ as $\lambda \downarrow 0$. Then there are two constants $0 < \delta, C_1 < \infty$ such that $0 \le \mathcal{N}_L(\lambda) \le C_1\lambda^{-s}$ for any $0 < \lambda \le \delta$. So,

$$\int_0^\delta \lambda^{s+\epsilon-1}\mathcal{N}_L(\lambda)d\lambda \le C_1 \int_0^\delta \lambda^{\epsilon-1}d\lambda < \infty.$$

Since L is in the trace class, when $s + \epsilon \ge 1$, (b) is trivial. Now we assume $s + \epsilon < 1$. Note that $\mathcal{N}_L(\lambda) \le \mathrm{Tr}(L)/\lambda$. So

$$\int_\delta^\infty \lambda^{s+\epsilon-1}\mathcal{N}_L(\lambda)d\lambda \le \mathrm{Tr}(L) \int_\delta^\infty \lambda^{s+\epsilon-2}d\lambda < \infty.$$

The claim (b) is verified by combining the above two bounds together.

Now we verify item (c). When $\lambda_i = O(i^{-1/s})$, there is some constant $C_2 > 0$ such that $\lambda_i \le C_2 i^{-1/s}$ for all $i \ge 1$. Since $u/(u + \lambda)$ is an increasing function of u,

$$\mathcal{N}_L(\lambda) \le \sum_{i=1}^\infty \frac{C_2 i^{-1/s}}{C_2 i^{-1/s} + \lambda} = \sum_{i=1}^\infty \frac{1}{1 + \lambda i^{1/s}/C_2}$$
$$\le \int_0^\infty \frac{du}{1 + \lambda u^{1/s}/C_2} = \left(\frac{\lambda}{C_2}\right)^{-s} \int_0^\infty \frac{du}{1 + u^{1/s}} = O(\lambda^{-s}).$$

This verifies the "if" part. For the "only-if" part, when $\mathcal{N}_L(\lambda) = O(\lambda^{-s})$, it is easy to see that there is some $C_3 > 0$ such that $\mathcal{N}_L(\lambda) \le C_3\lambda^{-s}$ for every $0 < \lambda \le s\lambda_1/(1 - s)$. Since for a fixed λ, $\{\lambda_i/(\lambda_i + \lambda)\}_{i=1}^\infty$ is a non-increasing sequence, $i\lambda_i/(\lambda_i + \lambda) \le C_3\lambda^{-s}$ for each $\lambda \in (0, s\lambda_1/(1 - s)]$ and $i \ge 1$. Therefore,

$$i\lambda_i \le \inf_{\lambda \in (0, s\lambda_1/(1-s)]} C_3\lambda^{-s}(\lambda_i + \lambda) = C_3\lambda_i^{1-s}s^{-s}(1 - s)^{s-1}, \tag{3.3}$$

where the infimum is achieved at $\lambda = s\lambda_i/(1 - s) \in (0, s\lambda_1/(1 - s)]$. From (3.3) one obtains $\lambda_i = O(i^{-1/s})$ as $i \to \infty$, and completes the proof. □

Assumption 3 is quite often adopted in the literature of functional linear regression. For example, if X is a Gaussian process, then (2.1) is satisfied with $c_{\mathsf{M}} = 3$. See [27, 5, 10].

4 Error Decomposition

Our analysis starts with error decomposition. By (1.5) (the equivalent expression of algorithm (1.3)), for any $t \ge 1$,

$$\begin{aligned}\hat{\beta}_{t+1} - \beta^* &= \hat{\beta}_t - \beta^* - \eta_t(\langle\hat{\beta}_t, x_t\rangle_2 - y_t)L_K x_t \\ &= (I - \eta_t L_K L_C)(\hat{\beta}_t - \beta^*) + \mathcal{B}_t,\end{aligned} \tag{4.1}$$

where $\mathcal{B}_t = \eta_t(y_t - \langle\hat{\beta}_t, x_t\rangle_2)L_K x_t + \eta_t L_K L_C(\hat{\beta}_t - \beta^*)$, of which the second term is the conditional mean of the first term,

$$\begin{aligned}\mathbb{E}_{z_t}\left[(y_t - \langle\hat{\beta}_t, x_t\rangle_2)L_K x_t\right] &= \mathbb{E}_{x_t}\left[\langle\beta^* - \hat{\beta}_t, x_t\rangle_2 L_K x_t\right] \\ &= L_K L_C(\beta^* - \hat{\beta}_t). \end{aligned} \tag{4.2}$$

Where $z_t = (x_t, y_t)$, and the expectations $\mathbb{E}_{z_t}$ and $\mathbb{E}_{x_t}$ are taken with respect to the (conditional) distributions of $z_t = (x_t, y_t)$ and x_t, respectively. Equation (4.2) shows that $\mathcal{B}_t$ is mean-zero, $\mathbb{E}_{z_t}[\mathcal{B}_t] = 0$. Then applying induction to (4.1) implies that for any $t \geq 1$,

$$\hat{\beta}_{t+1} - \beta^* = -\left[\prod_{k=1}^{t}(I - \eta_k L_K L_C)\right]\beta^* + \sum_{k=1}^{t}\left[\prod_{j=k+1}^{t}(I - \eta_j L_K L_C)\right]\mathcal{B}_k, \tag{4.3}$$

where and in the following, the product of an empty set of operators is defined as the identity operator, $\prod_{j=t+1}^{t}(I - \eta_j L_K L_C) = I$. Recall that $\mathscr{L}_K = L_C^{1/2} L_K L_C^{1/2}$.

Proposition 6. *Define $\{\hat{\varphi}_t = \langle\hat{\beta}_t, \cdot\rangle_2 : t \geq 1\}$ through (1.3). Then for any $t \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\mathcal{E}(\hat{\varphi}_{t+1})\right] \leq &\left\|\left[\prod_{k=1}^{t}(I - \eta_k\mathscr{L}_K)\right]L_C^{1/2}\beta^*\right\|_2^2 \\ &+ \sum_{k=1}^{t}\eta_k^2\left(\sigma^2 + \mathbb{E}\sqrt{\mathbb{E}_{x_k}\langle\beta^* - \hat{\beta}_k, x_k\rangle_2^4}\right)\left[\mathbb{E}\left\|\left[\prod_{j=k+1}^{t}(I - \eta_j\mathscr{L}_K)\right]L_C^{1/2}L_K x_k\right\|_2^4\right]^{1/2}, \end{aligned} \tag{4.4}$$

where the sum of an empty set is defined as zero.

The techniques of proving Proposition 6 are standard. For example, see [14], [18], and [26] for kernel-based online algorithms. The proof of Proposition 6 follows [7, Theorem 4], and is provided for the sake of completeness.

Proof of Proposition 6. The case $t = 0$ is trivial and we assume $t \geq 1$. For any k,

$$L_C^{1/2}(I - \eta_k L_K L_C) = L_C^{1/2} - \eta_k\mathscr{L}_K L_C^{1/2} = (I - \eta_k\mathscr{L}_K)\,L_C^{1/2}.$$

From (4.3),

$$L_C^{1/2}(\hat{\beta}_{t+1} - \beta^*) = -\left[\prod_{k=1}^{t}(I - \eta_k\mathscr{L}_K)\right]L_C^{1/2}\beta^* + \sum_{k=1}^{t}\left[\prod_{j=k+1}^{t}(I - \eta_j\mathscr{L}_K)\right]L_C^{1/2}\mathcal{B}_k.$$

It follows from (1.6) that

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}(\hat{\varphi}_{t+1})\right] = \mathbb{E}\left[\left\|L_C^{1/2}(\hat{\beta}_{t+1}-\beta^*)\right\|_2^2\right] \\
=&\mathbb{E}\left[\left\|-\left[\prod_{k=1}^{t}(I-\eta_k\mathscr{L}_K)\right]L_C^{1/2}\beta^* + \sum_{k=1}^{t}\left[\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right]L_C^{1/2}\mathcal{B}_k\right\|_2^2\right] \\
=&\left\|\left[\prod_{k=1}^{t}(I-\eta_k\mathscr{L}_K)\right]L_C^{1/2}\beta^*\right\|_2^2 + \mathbb{E}\left[\left\|\sum_{k=1}^{t}\left[\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right]L_C^{1/2}\mathcal{B}_k\right\|_2^2\right] \\
=&: \Upsilon_1^{\mathsf{E}} + \Upsilon_2^{\mathsf{E}} \qquad (4.5)
\end{aligned}$$

where in the expansion of the squared norm, the cross terms vanish because $\mathbb{E}[\mathcal{B}_k] = \mathbb{E}_{z_k}[\mathcal{B}_k] = 0$. The notations Υ_1^{E} and Υ_2^{E} are used only within this proof.

Let $W_1, \ldots, W_t$ be deterministic bounded linear operators. When $k > s$, $\mathcal{B}_s$ is independent of z_k, so $\mathbb{E}[\langle W_k\mathcal{B}_k, W_s\mathcal{B}_s\rangle_2] = \mathbb{E}\mathbb{E}_{z_k}\langle W_k\mathcal{B}_k, W_s\mathcal{B}_s\rangle_2 = \mathbb{E}[\langle W_k\mathbb{E}_{z_k}[\mathcal{B}_k], W_s\mathcal{B}_s\rangle_2] = 0$. We use this trick to expand the squared norm and cancel the cross terms,

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=1}^{t}W_k\mathcal{B}_k\right\|_2^2\right] &= \sum_{k=1}^{t}\mathbb{E}[\|W_k\mathcal{B}_k\|_2^2] + 2\sum_{s=1}^{t-1}\sum_{k=s+1}^{t}\mathbb{E}\left[\langle W_k\mathcal{B}_k, W_s\mathcal{B}_s\rangle_2\right] \\
&= \sum_{k=1}^{t}\mathbb{E}[\|W_k\mathcal{B}_k\|_2^2].
\end{aligned}$$

So we expand the squared norm in Υ_2^{E},

$$\Upsilon_2^{\mathsf{E}} = \sum_{k=1}^{t}\mathbb{E}\left[\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right]L_C^{1/2}\mathcal{B}_k\right\|_2^2\right]. \qquad (4.6)$$

Recall that for any random vector $\mathcal{B}$ with $\mathbb{E}[\|\mathcal{B}\|_2^2] < \infty$, $\mathbb{E}[\|\mathcal{B}-\mathbb{E}[\mathcal{B}]\|_2^2] = \mathbb{E}[\|\mathcal{B}\|_2^2] - \|\mathbb{E}[\mathcal{B}]\|_2^2 \le \mathbb{E}[\|\mathcal{B}\|_2^2]$. Note that $\mathcal{B}_k = \eta_k(y_k - \langle\hat{\beta}_k, x_k\rangle_2)L_Kx_k - \mathbb{E}_{z_k}[\eta_k(y_k - \langle\hat{\beta}_k, x_k\rangle_2)L_Kx_k]$ as we explained in (4.2). So, $\mathbb{E}[\|W_k\mathcal{B}_k\|_2^2] \le \eta_k^2\mathbb{E}[(y_k - \langle\hat{\beta}_k, x_k\rangle_2)^2\|W_kL_Kx_k\|_2^2]$ for any deterministic bounded linear operator W_k. Therefore,

$$\Upsilon_2^{\mathsf{E}} \le \sum_{k=1}^{t}\eta_k^2\mathbb{E}\left[\mathbb{E}_{\varepsilon_k}[(y_k - \langle\hat{\beta}_k, x_k\rangle_2)^2]\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right]L_C^{1/2}L_Kx_k\right\|_2^2\right], \qquad (4.7)$$

Furthermore, we recall that $\mathbb{E}_{\varepsilon_k}[(y_k - \langle \hat{\beta}_k, x_k \rangle_2)^2] = \sigma^2 + \langle \beta^* - \hat{\beta}_k, x_k \rangle_2^2$ and obtain

$$\begin{aligned}\Upsilon_2^{\mathsf{E}} \leq & \sigma^2 \sum_{k=1}^{t} \eta_k^2 \mathbb{E}\left[\left\| \left[\prod_{j=k+1}^{t} (I - \eta_j \mathscr{L}_K) \right] L_C^{1/2} L_K x_k \right\|_2^2 \right] \\ & + \sum_{k=1}^{t} \eta_k^2 \mathbb{E}\left[\langle \beta^* - \hat{\beta}_k, x_k \rangle_2^2 \left\| \left[\prod_{j=k+1}^{t} (I - \eta_j \mathscr{L}_K) \right] L_C^{1/2} L_K x_k \right\|_2^2 \right]. \end{aligned} \tag{4.8}$$

The proof is complete by applying Cauchy-Schwarz inequality to the right-hand side of (4.8). □

Now we consider the error decomposition for estimation error.

Proposition 7. *Let $\{\hat{\beta}_t\}$ be defined by (1.3). Assume $\beta^* \in \mathcal{H}_K$. We have the following error decomposition for any $t \geq 0$.*

$$\begin{aligned}\mathbb{E}\left[\|\hat{\beta}_{t+1} - \beta^*\|_K^2 \right] \leq & \left\| \left[\prod_{k=1}^{t} (I - \eta_k L_K L_C) \right] \beta^* \right\|_K^2 \\ & + \sum_{k=1}^{t} \eta_k^2 \left(\sigma^2 + \mathbb{E} \sqrt{\mathbb{E}_{x_k} \langle \beta^* - \hat{\beta}_k, x_k \rangle_2^4} \right) \left(\mathbb{E} \left\| \left[\prod_{j=k+1}^{t} (I - \eta_j L_K L_C) \right] L_K x_k \right\|_K^4 \right)^{1/2} \end{aligned} \tag{4.9}$$

The proof of Proposition 7 parallels that of Proposition 6. We see similar analysis in the literature for studying kernel-based online algorithms [14].

Proof of Proposition 7. By (4.3) and the fact $\mathbb{E}[\mathcal{B}_k] = 0$ we have

$$\begin{aligned}\mathbb{E}\left[\|\hat{\beta}_{t+1} - \beta^*\|_K^2 \right] &= \mathbb{E}\left[\left\| - \left[\prod_{k=1}^{t} (I - \eta_t L_K L_C) \right] \beta^* + \sum_{k=1}^{t} \left[\prod_{j=k+1}^{t} (I - \eta_j L_K L_C) \right] \mathcal{B}_k \right\|_K^2 \right] \\ &= \left\| \left[\prod_{k=1}^{t} (I - \eta_k L_K L_C) \right] \beta^* \right\|_K^2 + \mathbb{E}\left[\left\| \sum_{k=1}^{t} \left[\prod_{j=k+1}^{t} (I - \eta_j L_K L_C) \right] \mathcal{B}_k \right\|_K^2 \right], \end{aligned} \tag{4.10}$$

where the second term on the right-hand side is further estimated with the trick we used

in (4.7).

$$\mathbb{E}\left[\left\|\sum_{k=1}^{t}\left[\prod_{j=k+1}^{t}(I-\eta_j L_K L_C)\right]\mathcal{B}_k\right\|_K^2\right]=\sum_{k=1}^{t}\mathbb{E}\left[\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j L_K L_C)\right]\mathcal{B}_k\right\|_K^2\right]$$

$$\leq\sum_{k=1}^{t}\eta_k^2\mathbb{E}\left[(y_k-\langle\hat{\beta}_k,x_k\rangle_2)^2\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j L_K L_C)\right]L_K x_k\right\|_K^2\right]$$

$$\leq\sum_{k=1}^{t}\eta_k^2\left(\sigma^2+\mathbb{E}\sqrt{\mathbb{E}_{x_k}\langle\beta^*-\hat{\beta}_k,x_k\rangle_2^4}\right)\left(\mathbb{E}\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j L_K L_C)\right]L_K x_k\right\|_K^4\right)^{1/2}.$$

The proof is complete. □

5 Bounding the Excess Generalization Error

In this section, we study the excess generalization error $\mathcal{E}[\hat{\varphi}]$ and prove Theorems 1 and 2. This is achieved by continuing the estimation (4.4) in Proposition 6 to bound $\mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})]$ in terms of the step-sizes first, which is given in Theorem 10. The factors $\mathbb{E}[\mathcal{E}(\hat{\varphi}_k)]$ in the bound are difficult to avoid directly, and are further bounded by constant through Proposition 11. So, the coarse bound in Proposition 11, although appearing to be a corollary for the derived learning rates in Theorems 1 or 2, is to the contrary essential for deriving such rates. Then the step-sizes are specified in the online setting and finite-horizon setting to derive the learning rates for Theorems 1 and 2 respectively. This technique is widely used for the analysis of online algorithms, for example in [14, 18, 26] for analyzing kernel-based online learning schemes, in [7] for capacity independent analysis for online functional data learning algorithms, and in [11] for analyzing relaxed randomized Kaczmarz algorithms.

5.1 Analysis with General Step-sizes

In this subsection, we study the excess generalization error with minimal assumptions on the step-sizes. The following Lemma 8 is a typical application of the spectral theorem on the polynomial $u^\alpha\prod_{j=a}^{b}(1-\eta_j u)$ for $u\geq 0$. For a detailed proof, see e.g. [7, Lemma 2]. See also [11, 14, 18, 26, 25]. Note that when $b<a$, the sum $\sum_{j=a}^{b}\eta_j$ is defined to be zero.

Lemma 8. *Let A be a compact positive semi-definite operator on a Hilbert space. Let $\{\eta_i\} \subset (0, 1/\|A\|_{\mathsf{op}}]$. Then for any $a \le b$ and $\alpha > 0$, we have*

$$\left\| A^\alpha \prod_{j=a}^{b} (I - \eta_j A) \right\|_{\mathsf{op}}^2 \le \frac{(\alpha/e)^{2\alpha} + \|A\|_{\mathsf{op}}^{2\alpha}}{1 + \left(\sum_{j=a}^{b} \eta_j \right)^{2\alpha}}. \tag{5.1}$$

When $\alpha = 0$, we have

$$\left\| \prod_{j=a}^{b} (I - \eta_j A) \right\|_{\mathsf{op}}^2 \le 1. \tag{5.2}$$

In particular, when $a > b$, recall that the product $\prod_{j=a}^{b}(I - \eta_j A)$ is the identity operator. So the above estimates (5.1) and (5.2) still hold true. □

The following lemma provides an equivalent condition (5.3) to Assumption 3. It is interesting because apparently, Condition (5.3) is much stronger than Assumption 3.

Lemma 9. *Let X be the random function in Model (1.1). Let W be a compact operator (not necessarily self-adjoint or positive) on $L^2(\mathcal{T})$. Under Assumption 3,*

$$\mathbb{E}\left[\|WX\|_2^4 \right] \le c_{\mathsf{M}} \left(\mathbb{E}\left[\|WX\|_2^2 \right] \right)^2. \tag{5.3}$$

Proof. Write W' the adjoint operator of W. Then $W'W$ is a compact positive operator. So we write $\mu_1 \ge \mu_2 \ge \cdots > 0$ as all the positive eigenvalues of $W'W$, counting multiplicity. We use an orthonormal sequence $\{\psi_i\}$ in $L^2(\mathcal{T})$ as the corresponding eigenvectors. Assumption 3 implies that

$$\begin{aligned}
\mathbb{E}\left[\|WX\|_2^4 \right] &= \mathbb{E}\left[\langle X, W'WX \rangle_2^2 \right] \\
&= \mathbb{E}\left[\left(\sum_i \mu_i \langle \psi_i, X \rangle_2^2 \right)^2 \right] = \sum_{i,j} \mu_i \mu_j \mathbb{E}\left[\langle \psi_i, X \rangle_2^2 \langle \psi_j, X \rangle_2^2 \right] \\
&\le \sum_{i,j} \mu_i \mu_j \sqrt{\mathbb{E}[\langle \psi_i, X \rangle_2^4]} \sqrt{\mathbb{E}[\langle \psi_j, X \rangle_2^4]} \le c_{\mathsf{M}} \left(\mathbb{E} \sum_i \mu_i \langle \psi_i, X \rangle_2^2 \right)^2 \\
&= c_{\mathsf{M}} \left(\mathbb{E}\left[\|WX\|_2^2 \right] \right)^2.
\end{aligned}$$

The proof is then completed. □

Theorem 10. *Let $\{\hat{\beta}_t\}$ be defined by (1.3) with step-sizes $\{\eta_t\} \subset (0, \kappa^{-2}]$. Under Assumption 1 (with $r > 0$), Assumption 2 (with $0 < s \leq 1$), and Assumption 3, for any $t \geq 0$,*

$$\mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})] \leq \frac{\|g^*\|_2^2 \left((r/e)^{2r} + \kappa^{4r}\right)}{1 + \left(\sum_{k=1}^t \eta_k\right)^{2r}} + \sum_{k=1}^t \eta_k^2 \left(\sigma^2 + \sqrt{c_{\mathsf{M}}}\mathbb{E}[\mathcal{E}(\hat{\varphi}_k)]\right) \frac{\sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_K^s)\left[(\frac{2-s}{2e})^{2-s} + \kappa^{4-2s}\right]}{1 + \left(\sum_{j=k+1}^t \eta_j\right)^{2-s}}. \tag{5.4}$$

Proof. When $t = 0$, Bound (5.4) is reduced to

$$\mathbb{E}[\mathcal{E}(0)] \leq \|g^*\|_2^2((r/e)^{2r} + \kappa^{4r}). \tag{5.5}$$

We use Assumption 1 to have $\mathcal{E}(0) = \|L_C^{1/2}\beta^*\|_2^2 = \|\mathscr{L}_K^r g^*\|_2^2 \leq \kappa^{4r}\|g^*\|_2^2$, which verifies (5.5).

Now we assume $t \geq 1$. We let J_1 and J_2 denote the two terms in the right-hand side of (4.4) in Proposition 6, respectively. That is, $\mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})] \leq J_1 + J_2$, with

$$J_1 = \left\| \left[\prod_{k=1}^t (I - \eta_k \mathscr{L}_K)\right] L_C^{1/2}\beta^* \right\|_2^2, \text{ and}$$

$$J_2 = \sum_{k=1}^t \eta_k^2 \left(\sigma^2 + \mathbb{E}\sqrt{\mathbb{E}_{x_k}\langle \beta^* - \hat{\beta}_k, x_k\rangle_2^4}\right) \left[\mathbb{E}\left\| \left[\prod_{j=k+1}^t (I - \eta_j \mathscr{L}_K)\right] L_C^{1/2} L_K x_k \right\|_2^4\right]^{1/2}.$$

Assumption 1 gives $L_C^{1/2}\beta^* = \mathscr{L}_K^r g^*$ for some $r > 0$. Recall the assumption $\{\eta_j\} \subset (0, \kappa^{-2}]$. We apply Lemma 8 to bound J_1,

$$J_1 = \left\| \left[\prod_{k=1}^t (I - \eta_k \mathscr{L}_K)\right] \mathscr{L}_K^r g^* \right\|_2^2 \leq \|g^*\|_2^2 \frac{(r/e)^{2r} + \kappa^{4r}}{1 + \left(\sum_{k=1}^t \eta_k\right)^{2r}}.$$

To bound J_2, we apply Assumption 3 (the moment condition),

$$\mathbb{E}\sqrt{\mathbb{E}_{x_k}\langle \beta^* - \hat{\beta}_k, x_k\rangle_2^4} \leq \sqrt{c_{\mathsf{M}}}\mathbb{E}\left[\langle \beta^* - \hat{\beta}_k, x_k\rangle_2^2\right] = \sqrt{c_{\mathsf{M}}}\mathbb{E}[\mathcal{E}(\hat{\varphi}_k)]. \tag{5.6}$$

Recall that for any bounded linear operator A on $L^2(\mathcal{T})$, $\mathbb{E}[\|Ax_t\|_2^2] = \mathbb{E}\mathrm{Tr}(Ax_t \otimes x_t A') = \mathrm{Tr}(AL_C A')$. We apply Lemma 9, Assumption 2 (with $0 < s \leq 1$), and Lemma 8 to obtain

that

$$\begin{aligned}
&\left[\mathbb{E}\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right]L_C^{1/2}L_Kx_k\right\|_2^4\right]^{1/2}\\
\leq&\sqrt{c_{\mathsf{M}}}\mathbb{E}\left\|\left[\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right]L_C^{1/2}L_Kx_k\right\|_2^2=\sqrt{c_{\mathsf{M}}}\mathrm{Tr}\left(\mathscr{L}_K^2\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)^2\right)\\
\leq&\sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_K^s)\left\|\mathscr{L}_K^{1-\frac{s}{2}}\prod_{j=k+1}^{t}(I-\eta_j\mathscr{L}_K)\right\|_{\mathsf{op}(L^2)}^2\leq\sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_K^s)\frac{\left(\frac{2-s}{2e}\right)^{2-s}+\kappa^{4-2s}}{1+\left(\sum_{j=k+1}^{t}\eta_j\right)^{2-s}}. \quad (5.7)
\end{aligned}$$

The proof is complete. □

Proposition 11. *Let $t\geq 1$. Let $\{\hat{\beta}_k\}$ be defined by (1.3) with step-sizes $\{\eta_k\}\subset(0,\kappa^{-2}]$. Suppose that Assumption 2 with $0<s\leq 1$ (in particular, Assumption 2 is not needed when $s=1$) and Assumption 3 hold. In particular, when $t\geq 2$ we assume for any $k\leq t-1$ that*

$$c_{\mathsf{M}}\mathrm{Tr}(\mathscr{L}_K^s)\left[\left(\frac{2-s}{2e}\right)^{2-s}+\kappa^{4-2s}\right]\sum_{l=1}^{k}\frac{\eta_l^2}{1+\left(\sum_{j=l+1}^{k}\eta_j\right)^{2-s}}\leq\frac{1}{2}. \quad (5.8)$$

Then we have a coarse estimation of the expected excess generalization error for $k=1,\cdots,t$,

$$\mathbb{E}\left[\mathcal{E}(\hat{\varphi}_k)\right]\leq 2\left\|\beta^*\right\|_2^2+\frac{\sigma^2}{\sqrt{c_{\mathsf{M}}}}. \quad (5.9)$$

We see that (5.9) only provides a coarse bound $\mathbb{E}\left[\mathcal{E}(\hat{\varphi}_k)\right]=O(1)$. However, the designed purpose of Proposition 11 is to estimate $\mathbb{E}\left[\mathcal{E}(\hat{\varphi}_k)\right]$ in the right-hand side of (5.4) in our convergence analysis, and a bound finer than $O(1)$ would not serve the purpose better, because a constant variance σ^2 is added to $\sqrt{c_{\mathsf{M}}}\mathbb{E}[\mathcal{E}(\hat{\varphi}_k)]$ in (5.4).

Proof of Proposition 11. We organize the proof by induction. Recall that $\hat{\varphi}_1=0$ and $\|L_C\|_{\mathsf{op}(L^2)}\leq 1$. When $t=1$, (5.8) is verified by

$$\mathcal{E}(0)=\|L_C^{1/2}\beta^*\|_2^2\leq\|\beta^*\|_2^2.$$

Let $T\geq 2$, and we assume Proposition 11 holds for $t=1,\ldots,T-1$. To finish the induction, we need only to prove Proposition 11 for $t=T$. That is, we assume (5.8) and

(5.9) for $t = 1, \ldots, T-1$ and need only to prove (5.9) for $t = T$. To this end, we use Proposition 6 to have

$$\mathbb{E}[\mathcal{E}(\hat{\varphi}_T)] \leq \Upsilon_1^T + \sum_{k=1}^{T-1} \eta_k^2 \left(\sigma^2 + \mathbb{E}\sqrt{\mathbb{E}_{x_k} \langle \beta^* - \hat{\beta}_k, x_k \rangle_2^4} \right) \Upsilon_{2,k}^T, \tag{5.10}$$

where

$$\Upsilon_1^T = \left\| \left[\prod_{k=1}^{T-1} (I - \eta_k \mathscr{L}_K) \right] L_C^{1/2} \beta^* \right\|_2^2, \quad \text{and}$$

$$\Upsilon_{2,k}^T = \left[\mathbb{E} \left\| \left[\prod_{j=k+1}^{T-1} (I - \eta_j \mathscr{L}_K) \right] L_C^{1/2} L_K x_k \right\|_2^4 \right]^{1/2}, \quad k = 1, \ldots, T-1.$$

To bound Υ_1^T, we note that $\|I - \eta_k \mathscr{L}_K\|_{\mathsf{op}(L^2)} \leq 1$ because $\eta_k \in (0, \kappa^{-2}]$ and $\|\mathscr{L}_K\|_{\mathsf{op}(L^2)} \leq \kappa^2$. So,

$$\Upsilon_1^T \leq \|L_C^{1/2} \beta^*\|_2^2 \leq \|\beta^*\|_2^2.$$

Next, we follow (5.6), use the induction assumption and Assumption 3 (the moment condition) to obtain

$$\mathbb{E}\sqrt{\mathbb{E}_{x_k} \langle \beta^* - \hat{\beta}_k, x_k \rangle_2^4} \leq \sqrt{c_{\mathsf{M}}} \mathbb{E}[\mathcal{E}(\hat{\varphi}_k)] \leq \sigma^2 + 2\sqrt{c_{\mathsf{M}}} \|\beta^*\|_2^2, \quad k = 1, \ldots, T-1.$$

Then, we follow (5.7) and use Assumption 2 with $0 < s \leq 1$ to obtain

$$\begin{aligned} \Upsilon_{2,k}^T &\leq \sqrt{c_{\mathsf{M}}} \mathbb{E} \left\| \left[\prod_{j=k+1}^{T-1} (I - \eta_j \mathscr{L}_K) \right] L_C^{1/2} L_K x_k \right\|_2^2 \\ &\leq \sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_K^s) \frac{\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s}}{1 + (\sum_{j=k+1}^{T-1} \eta_j)^{2-s}}, \quad k = 1, \ldots, T-1. \end{aligned}$$

We continue (5.10) and use Condition (5.8) for $k = 1, \ldots, T-1$ to have

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\hat{\varphi}_T)] &\leq \|\beta^*\|_2^2 + \sum_{k=1}^{T-1} \eta_k^2 (2\sigma^2 + 2\sqrt{c_{\mathsf{M}}} \|\beta^*\|_2^2) \sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_K^s) \frac{\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s}}{1 + (\sum_{j=k+1}^{T-1} \eta_j)^{2-s}} \\ &\leq \|\beta^*\|_2^2 + \frac{\sigma^2}{\sqrt{c_{\mathsf{M}}}} + \|\beta^*\|_2^2. \end{aligned}$$

This completes the proof. □

5.2 Analysis in Online and Finite-horizon Settings of Step-sizes

In this subsection we study the excess generalization error in the online and finite-horizon settings of step-sizes, respectively. The following Lemma 12 is commonly used in the literature [7, 26, 11] with smaller ranges of parameters θ and ν. In this paper, we need coverage of the whole domain $\nu > 0$ and $0 < \theta < 1$, and the proof is not elsewhere available to our best knowledge.

Lemma 12. *Let $t \geq 1$, $\eta_k = \eta_0 k^{-\theta}$ with $\eta_0 > 0$ and $0 < \theta < 1$. For any $\nu > 0$,*

$$\left(\sum_{k=1}^{t} \eta_k\right)^{-\nu} \leq \left(\frac{\eta_0(1-2^{\theta-1})}{1-\theta}\right)^{-\nu} (t+1)^{-\nu(1-\theta)}, \tag{5.11}$$

and

$$\sum_{k=1}^{t} \frac{\eta_k^2}{1+\left(\sum_{j=k+1}^{t} \eta_j\right)^{\nu}} \leq C^{\mathsf{OL}} \begin{cases} (t+1)^{\omega}\log(t+1), & (\nu,\theta) \in \Omega, \\ (t+1)^{\omega}, & (\nu,\theta) \notin \Omega, \end{cases}$$

where $\Omega = \{(\nu,\theta) : 0 < \nu \leq 1$ and $\theta = 1/2\} \cup \{(\nu,\theta) : \nu = 1$ and $0 < \theta \leq 1/2\}$, C^{OL} is a constant independent of t, and

$$\omega = \omega(\nu,\theta) = \begin{cases} 1-2\theta-\nu+\nu\theta, & 0 < \nu \leq 1 \text{ and } 0 < \theta \leq 1/2, \\ -\theta, & \nu \geq 1 \text{ and } 0 < \theta \leq \nu/(\nu+1), \\ -\nu(1-\theta), & 1/2 \leq \theta < 1 \text{ and } \theta \geq \nu/(\nu+1). \end{cases} \tag{5.12}$$

In particular, when $\nu \geq 1$, $\omega = -\min\{\theta, \nu(1-\theta)\}$. The constant C^{OL} will be specified by (5.15) in the proof.

Lemma 12 is based on Lemma 14 in Appendix. Same as Lemma 14, we purposely allow the domains to overlap in (5.12), to simplify the usage later. We will elucidate the parameter ω and the set Ω by Figure 5 in Appendix.

Proof of Lemma 12. For any $k \geq 0$ and $t \geq k+1$,

$$\sum_{j=k+1}^{t} \eta_j = \eta_0 \sum_{j=k+1}^{t} j^{-\theta} \geq \frac{\eta_0}{1-\theta}\left[(t+1)^{1-\theta} - (k+1)^{1-\theta}\right]. \tag{5.13}$$

We set $k = 0$ to have

$$\sum_{j=1}^{t} \eta_j \geq \frac{\eta_0(1-2^{\theta-1})}{1-\theta}(t+1)^{1-\theta}. \tag{5.14}$$

One raises (5.14) to power $-\nu$ to obtain (5.11).

Recall that for $k \geq 1$, one has $k \geq (k+2)/3$, and

$$\sum_{k=1}^{t} \frac{\eta_k^2}{1+\left(\sum_{j=k+1}^{t}\eta_j\right)^{\nu}} \leq \eta_t^2 + \sum_{k=1}^{t-1} \frac{\eta_0^2 (k+2)^{-2\theta} 3^{2\theta}}{1+\left(\frac{\eta_0}{1-\theta}\right)^{\nu}\left[(t+1)^{1-\theta}-(k+1)^{1-\theta}\right]^{\nu}}.$$

Note that for any $k = 1, \ldots, t-1$,

$$\frac{(k+2)^{-2\theta}}{1+\left[(t+1)^{1-\theta}-(k+1)^{1-\theta}\right]^{\nu}} \leq \int_{k+1}^{k+2} \frac{u^{-2\theta}du}{1+\left[(t+1)^{1-\theta}-u^{1-\theta}\right]^{\nu}}.$$

We use Lemma 14 to have

$$\begin{aligned}
\sum_{k=1}^{t} \frac{\eta_k^2}{1+\left(\sum_{j=k+1}^{t}\eta_j\right)^{\nu}} &\leq \eta_0^2 t^{-2\theta} + \frac{3^{2\theta}\eta_0^2}{\min\left\{1, \left(\frac{\eta_0}{1-\theta}\right)^{\nu}\right\}} \int_2^{t+1} \frac{u^{-2\theta}du}{1+\left[(t+1)^{1-\theta}-u^{1-\theta}\right]^{\nu}} \\
&\leq \eta_0^2 2^{2\theta}(t+1)^{-2\theta} + \frac{3^{2\theta}\eta_0^2 C_0^{\mathsf{OL}}}{\min\left\{1, \left(\frac{\eta_0}{1-\theta}\right)^{\nu}\right\}} \times \begin{cases} (t+1)^{\omega}\log(t+1), & (\nu,\theta) \in \Omega, \\ (t+1)^{\omega}, & (\nu,\theta) \notin \Omega. \end{cases}
\end{aligned}$$

Now we verify that $-2\theta \leq \omega$ on the whole domain $(0,\infty)\times(0,1)$ of parameters. When $0 < \theta \leq 1/2$ and $0 < \nu \leq 1$, $\omega = -2\theta + (1-\nu) + \nu\theta \geq -2\theta$. When $1/2 < \theta < 1$ and $0 < \nu \leq 1$, $\omega = -\nu(1-\theta) \geq -1+\theta > -\theta > -2\theta$. When $0 < \theta < \nu/(\nu+1)$ and $\nu > 1$, $\omega = -\theta > -2\theta$. When $\nu/(\nu+1) \leq \theta < 1$ and $\nu > 1$, $\theta > \nu/(\nu+2)$, so $\omega = -\nu(1-\theta) > -2\theta$.

We complete the proof by letting

$$C^{\mathsf{OL}} = \frac{\eta_0^2 2^{2\theta}}{\log 2} + \frac{3^{2\theta}\eta_0^2 C_0^{\mathsf{OL}}}{\min\left\{1, \left(\frac{\eta_0}{1-\theta}\right)^{\nu}\right\}}. \tag{5.15}$$

□

Proof of Theorem 1. First, we shall apply Proposition 11. To verify the assumptions in Proposition 11, we need only to determine the constant C_1^{S} to guarantee (5.8), i.e., for $k = 1, \ldots, t-1$,

$$c_{\mathsf{M}}\mathrm{Tr}(\mathscr{L}_K^s)\left[\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s}\right] \sum_{l=1}^{k} \frac{\eta_l^2}{1+\left(\sum_{j=l+1}^{k}\eta_j\right)^{2-s}} \leq \frac{1}{2}. \tag{5.16}$$

Recall $r > 0$ and $0 < s \leq 1$. We apply Lemma 12 with

$$\nu = 2 - s \geq 1, \quad \text{and,} \quad 0 < \theta = \frac{\min\{2r, 2-s\}}{1+\min\{2r, 2-s\}} \leq \frac{\nu}{\nu+1}, \tag{5.17}$$

so $\omega = -\theta < 0$, and for $k = 1, \ldots, t-1$,

$$\sum_{l=1}^{k} \frac{\eta_l^2}{1 + \left(\sum_{j=l+1}^{k} \eta_j\right)^{2-s}} \leq C^{\mathsf{OL}} \begin{cases} (k+1)^{-\theta} \log(k+1), & s = 1, \\ (k+1)^{-\theta}, & 0 < s < 1. \end{cases} \tag{5.18}$$

Recall $\eta_0 \leq 1$. The above inequality is continued by

$$\begin{aligned} C^{\mathsf{OL}} &= \frac{\eta_0^2 2^{2\theta}}{\log 2} + \frac{3^{2\theta} \eta_0^2 C_0^{\mathsf{OL}}}{\min\left\{1, \left(\frac{\eta_0}{1-\theta}\right)^{2-s}\right\}} \leq \frac{\eta_0^s 4^\theta}{\log 2} + \frac{9^\theta \eta_0^2 C_0^{\mathsf{OL}}}{\eta_0^{2-s} \min\{1, (1-\theta)^{-2+s}\}} \\ &\leq \eta_0^s \left(\frac{4^\theta}{\log 2} + 9^\theta C_0^{\mathsf{OL}}\right). \end{aligned} \tag{5.19}$$

On the other hand, $(k+1)^{-\theta} \leq 1$, and $(k+1)^{-\theta} \log(k+1) \leq \frac{1}{e\theta}$ (see (A.6)). Therefore, to achieve (5.16) (which is just (5.8) for Proposition 11), we need simply to let

$$C_1^{\mathsf{S}} = \left\{ 2c_{\mathsf{M}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s} \right] \left(\frac{4^\theta}{\log 2} + 9^\theta C_0^{\mathsf{OL}}\right) \left(1 + \frac{1}{e\theta}\right) \right\}^{-1/s}. \tag{5.20}$$

Second, we apply Theorem 10, of which the conditions are now all satisfied. We plug (5.9) of Proposition 11, into (5.4) of Theorem 10, to obtain

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})] \leq & \frac{\|g^*\|_2^2 ((r/e)^{2r} + \kappa^{4r})}{\left(\sum_{k=1}^{t} \eta_k\right)^{2r}} \\ & + \sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s} \right] \left(2\|\beta^*\|_2^2 \sqrt{c_{\mathsf{M}}} + 2\sigma^2\right) \\ & \times \sum_{k=1}^{t} \frac{\eta_k^2}{1 + \left(\sum_{j=k+1}^{t} \eta_j\right)^{2-s}}. \end{aligned} \tag{5.21}$$

For the first term in the right-hand side of (5.21), we apply Lemma 12 with $\nu = 2r$. The last sum in (5.21) is bounded above in (5.18). We have

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})] \leq & \|g^*\|_2^2 ((r/e)^{2r} + \kappa^{4r}) \left(\frac{\eta_0 (1 - 2^{\theta-1})}{1-\theta}\right)^{-2r} (t+1)^{-2r(1-\theta)} \\ & + 2(\sigma^2 + \sqrt{c_{\mathsf{M}}} \|\beta^*\|_2^2) \sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s} \right] C^{\mathsf{OL}} \\ & \times \begin{cases} (t+1)^{-\theta} \log(t+1), & s = 1, \\ (t+1)^{-\theta}, & 0 < s < 1. \end{cases} \end{aligned}$$

From $\theta = \frac{\min\{2r,\nu\}}{1+\min\{2r,\nu\}} \le \frac{2r}{1+2r}$, we have $-2r(1-\theta) \le -\theta$. Therefore,

$$\mathbb{E}[\mathcal{E}(\hat{\varphi}_{t+1})] \le C_1 \begin{cases} (t+1)^{-\theta}\log(t+1), & s=1, \\ (t+1)^{-\theta}, & 0<s<1. \end{cases}$$

with

$$\begin{aligned} C_1 =& \frac{\|g^*\|_2^2\left((r/e)^{2r}+\kappa^{4r}\right)}{\log 2}\left(\frac{\eta_0(1-2^{\theta-1})}{1-\theta}\right)^{-2r} \\ &+ 2(\sigma^2+\sqrt{c_{\mathsf{M}}}\|\beta^*\|_2^2)\sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_K^s)\left[\left(\frac{2-s}{2e}\right)^{2-s}+\kappa^{4-2s}\right]C^{\mathsf{OL}}. \end{aligned} \tag{5.22}$$

□

Proof of Theorem 2. First, for applying Proposition 11, we need only to find an upper bound C_2^{S} of step-sizes to guarantee (5.8), i.e., for $k = 1, \ldots, T-1$,

$$c_{\mathsf{M}}\mathrm{Tr}(\mathscr{L}_K^s)\left[\left(\frac{2-s}{2e}\right)^{2-s}+\kappa^{4-2s}\right]\sum_{l=1}^{k}\frac{\eta_l^2}{1+\left(\sum_{j=l+1}^{k}\eta_j\right)^{2-s}} \le \frac{1}{2}. \tag{5.23}$$

Recall $\eta_t = \eta_0 T^{-2r/(2r+1)}$. We write $\eta = \eta_t$. For any $k \le T-1$, we bound the sum in (5.23) as

$$\begin{aligned} \sum_{l=1}^{k}\frac{\eta_l^2}{1+\left(\sum_{j=l+1}^{k}\eta_j\right)^{2-s}} &= \eta^2 + \sum_{t=1}^{k-1}\frac{\eta^2}{1+(t\eta)^{2-s}} \le \eta^2 + \eta\int_0^{k-1}\frac{\eta du}{1+(\eta u)^{2-s}} \\ &\le \eta^2 + \eta\int_0^{\eta T}\frac{du}{1+u^{2-s}} \le \eta^2 + \eta\begin{cases} \frac{2-s}{1-s}, & 0<s<1, \\ \log(\eta T+1), & s=1, \end{cases} \end{aligned} \tag{5.24}$$

where in the last inequality we used (A.4) for $0 < s < 1$. Recall that $\eta_0 \le 1$.

$$\eta\log(\eta T+1) \le \eta\log(T^{\frac{1}{2r+1}}+1) \le C_2^{\mathsf{S}} T^{\frac{-2r}{2r+1}}\left(\log 2 + \frac{1}{2r+1}\log T\right) \le C_2^{\mathsf{S}}\frac{2er+1}{2er},$$

where in the last inequality we used (A.6). We have a coarse estimate for $k \le T-1$,

$$\sum_{l=1}^{k}\frac{\eta_l^2}{1+\left(\sum_{j=l+1}^{k}\eta_j\right)^{2-s}} \le C_2^{\mathsf{S}}C_{2*}^{\mathsf{S}}, \quad C_{2*}^{\mathsf{S}} := 2 + \begin{cases} 1/(1-s), & 0<s<1, \\ 1/(2er), & s=1. \end{cases}$$

Therefore, to guarantee (5.23) (which is just (5.8) for Proposition 11), we just need

$$C_2^{\mathsf{S}} = \left\{ 2c_{\mathsf{M}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e} \right)^{2-s} + \kappa^{4-2s} \right] C_{2*}^{\mathsf{S}} \right\}^{-1}. \tag{5.25}$$

We plug (5.9) of Proposition 11, into (5.4) of Theorem 10 to obtain

$$\begin{aligned} \mathbb{E}[\mathcal{E}(\hat{\varphi}_{T+1})] \leq & \frac{\|g^*\|_2^2((r/e)^{2r} + \kappa^{4r})}{\left(\sum_{k=1}^T \eta_k\right)^{2r}} + \sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e} \right)^{2-s} + \kappa^{4-2s} \right] \\ & \times (2\|\beta^*\|_2^2 \sqrt{c_{\mathsf{M}}} + 2\sigma^2) \sum_{k=1}^T \frac{\eta_k^2}{1 + \left(\sum_{j=k+1}^T \eta_j\right)^{2-s}}. \end{aligned} \tag{5.26}$$

At the right-hand side, the first term is bounded with

$$\left(\sum_{k=1}^T \eta_k \right)^{-2r} \leq \eta_0^{-2r} T^{-2r/(2r+1)},$$

and the second term is bounded in (5.24),

$$\sum_{k=1}^T \frac{\eta_k^2}{1 + \left(\sum_{j=k+1}^T \eta_j\right)^{2-s}} \leq \eta_0^2 T^{-\frac{2r}{2r+1}} + \eta_0 \begin{cases} \frac{2-s}{1-s} T^{-\frac{2r}{2r+1}}, & 0 < s < 1, \\ \frac{2r+2}{2r+1} T^{-\frac{2r}{2r+1}} \log(T+1), & s = 1, \end{cases}$$

where in the case $s = 1$ we have used

$$\log(T^a + 1) \leq \log 2 + a \log T \leq (a+1) \log(T+1), \quad \text{for any } T \geq 1, a > 0.$$

The proof is therefore completed by letting

$$\begin{aligned} C_2 = & \frac{\eta_0^{-2r} \|g^*\|_2^2}{\log 2} ((r/e)^{2r} + \kappa^{4r}) + 2\sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e} \right)^{2-s} + \kappa^{4-2s} \right] \\ & \times (\|\beta^*\|_2^2 \sqrt{c_{\mathsf{M}}} + \sigma^2) \times \begin{cases} \eta_0^2 + \eta_0 \frac{2-s}{1-s}, & 0 < s < 0, \\ \eta_0^2 + \eta_0 \frac{2r+2}{2r+1}, & s = 1. \end{cases} \end{aligned} \tag{5.27}$$

□

6 Bounding the Estimation Error

In this section, we bound the estimation error in $\mathcal{H}_K$ metric and prove Theorems 3 and 4. The analysis parallels Section 5. We continue the Bound (4.9) in Proposition 7 to derive Bound (6.2) in Theorem 13 for general step-sizes. The coarse estimation in Proposition 11 is used again. Then the settings on step-sizes are applied to derive Theorems 3 and 4 respectively. We see similar analysis in [11] for randomized Kaczmarz algorithms, in [14] for kernel-based online algorithms, and in [26] for capacity independent analysis of online algorithms.

Theorem 13. *Let $t \geq 0$ be an integer. Let $\{\hat{\beta}_k : 1 \leq k \leq t+1\}$ be defined by (1.3) with step-sizes $\{\eta_k\} \subset (0, \kappa^{-2}]$. Suppose that Assumptions 2 (with $0 < s \leq 1$), 3, and 4 (with $r > 0$) hold. In particular, when $t \geq 2$ we assume for any $k \leq t-1$ that*

$$c_{\mathsf{M}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e} \right)^{2-s} + \kappa^{4-2s} \right] \sum_{l=1}^{k} \frac{\eta_l^2}{1 + \left(\sum_{j=l+1}^{k} \eta_j \right)^{2-s}} \leq \frac{1}{2}. \tag{6.1}$$

Then,

$$\mathbb{E}[\|\hat{\beta}_{t+1} - \beta^*\|_K^2] \leq C^K \left[\left(\sum_{k=1}^{t} \eta_k \right)^{-2r} + \sum_{k=1}^{t} \frac{\eta_k^2}{1 + \left(\sum_{j=k+1}^{t} \eta_j \right)^{1-s}} \right], \tag{6.2}$$

where C^K is a constant independent of t and will be specified in the proof, and when $s = 1$, $(\sum_{j=k+1}^{t} \eta_j)^{1-s} := 1$ even when $k = t$ that vanishes the sum.

Proof. Assumption 4 that $\beta^* = L_K^{1/2} \mathscr{L}_C^r g^\dagger$ (for some $g^\dagger \in L^2(\mathcal{T})$ and $r > 0$) guarantees $\beta^* \in \mathcal{H}_K$. So we start from Proposition 7 by bounding $\mathbb{E}[\|\hat{\beta}_{t+1} - \beta^*\|_K^2] \leq \Upsilon_1^K + \Upsilon_2^K$, where

$$\Upsilon_1^K = \left\| \left[\prod_{k=1}^{t} (I - \eta_k L_K L_C) \right] \beta^* \right\|_K^2, \quad \text{and}$$

$$\Upsilon_2^K = \sum_{k=1}^{t} \eta_k^2 (\sigma^2 + \mathbb{E}\sqrt{\mathbb{E}_{x_k} \langle \beta^* - \hat{\beta}_k, x_k \rangle_2^4}) \left(\mathbb{E} \left\| \left[\prod_{j=k+1}^{t} (I - \eta_j L_K L_C) \right] L_K x_k \right\|_K^4 \right)^{1/2}.$$

Recall that $\mathscr{L}_C = L_K^{1/2} L_C L_K^{1/2}$. We use Lemma 8 to bound the operator norm,

$$\begin{aligned}\Upsilon_1^K &\leq \left\| L_K^{1/2} \left[\prod_{k=1}^{t} (I - \eta_k \mathscr{L}_C) \right] \mathscr{L}_C^r g^\dagger \right\|_K^2 \leq \left\| \mathscr{L}_C^r \prod_{k=1}^{t} (I - \eta_k \mathscr{L}_C) \right\|_{\mathsf{op}(L^2)}^2 \|g^\dagger\|_2^2 \\ &\leq \frac{\|g^\dagger\|_2^2 ((r/e)^{2r} + \|\mathscr{L}_C\|_{\mathsf{op}(L^2)}^{2r})}{1 + (\sum_{k=1}^{t} \eta_k)^{2r}}.\end{aligned}$$

For Υ_2^K, we consider its different factors separately. First, recall that $\hat{\beta}_k$ is independent of x_k. Assumption 3 (moment condition) guarantees that

$$\mathbb{E} \sqrt{\mathbb{E}_{x_k} \langle \beta^* - \hat{\beta}_k, x_k \rangle_2^4} \leq \sqrt{c_{\mathsf{M}}} \mathbb{E}[\langle \beta^* - \hat{\beta}_k, x_k \rangle_2^2] = \sqrt{c_{\mathsf{M}}} \mathbb{E}[\mathcal{E}(\hat{\varphi}_k)].$$

With Proposition 11, our assumption on step-sizes guarantees that

$$\mathbb{E}[\mathcal{E}(\hat{\varphi}_k)] \leq 2\|\beta^*\|_2^2 + \frac{\sigma^2}{\sqrt{c_{\mathsf{M}}}}, \text{ for all } k = 1, \ldots, t.$$

Second, we use Lemma 9 and recall that $\|L_K^{1/2} f\|_K = \|f\|_2$ for any $f \in L^2(\mathcal{T})$ to obtain

$$\begin{aligned}\Upsilon_{2*}^K &:= \left(\mathbb{E} \left\| \left[\prod_{j=k+1}^{t} (I - \eta_j L_K L_C) \right] L_K x_k \right\|_K^4 \right)^{1/2} \\ &= \left(\mathbb{E} \left\| \left[\prod_{j=k+1}^{t} (I - \eta_j \mathscr{L}_C) \right] L_K^{1/2} x_k \right\|_2^4 \right)^{1/2} \\ &\leq \sqrt{c_{\mathsf{M}}} \mathbb{E} \left\| \left[\prod_{j=k+1}^{t} (I - \eta_j \mathscr{L}_C) \right] L_K^{1/2} x_k \right\|_2^2.\end{aligned}$$

Recall that $\mathbb{E}[\|Ax_t\|_2^2] = \mathbb{E}\mathrm{Tr}(Ax_t \otimes x_t A') = \mathrm{Tr}(AL_C A')$ for any bounded linear operator A on $L^2(\mathcal{T})$.

$$\begin{aligned}\Upsilon_{2*}^K &\leq \sqrt{c_{\mathsf{M}}} \mathrm{Tr} \left(\mathscr{L}_C \prod_{j=k+1}^{t} (I - \eta_j \mathscr{L}_C)^2 \right) \\ &\leq \sqrt{c_{\mathsf{M}}} \mathrm{Tr}(\mathscr{L}_C^s) \left\| \mathscr{L}_C^{\frac{1-s}{2}} \prod_{j=k+1}^{t} (I - \eta_j \mathscr{L}_C) \right\|_{\mathsf{op}(L^2)}^2,\end{aligned}$$

where we abuse the notation a little and let $\mathscr{L}_C^{(1-s)/2}$ denote the identity operator when $s = 1$. Thanks to Lemma 8,

$$\Upsilon_{2*}^K \leq \sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_C^s)\frac{\left(\frac{1-s}{2e}\right)^{1-s} + \|\mathscr{L}_C\|_{\mathsf{op}(L^2)}^{1-s}}{1 + (\sum_{j=k+1}^t \eta_j)^{1-s}}, \text{ when } 0 < s < 1,$$

and $\Upsilon_{2*}^K \leq \sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_C)$ when $s = 1$.

To summarize, for any $t \geq 1$, when $0 < s < 1$,

$$\mathbb{E}[\|\hat{\beta}_{t+1} - \beta^*\|_K^2] \leq C^K \left[\left(\sum_{k=1}^t \eta_k\right)^{-2r} + \sum_{k=1}^t \frac{\eta_k^2}{1 + (\sum_{j=k+1}^t \eta_j)^{1-s}}\right], \tag{6.3}$$

where

$$C^K = \max\left\{\|g^\dagger\|_2^2\left(\left(\frac{r}{e}\right)^{2r} + \|\mathscr{L}_C\|_{\mathsf{op}(L^2)}^{2r}\right),\right.$$
$$\left.(2\sqrt{c_{\mathsf{M}}}\|\beta^*\|_2^2 + 2\sigma^2)\sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_C^s)\left[\left(\frac{1-s}{2e}\right)^{1-s} + \|\mathscr{L}_c\|_{\mathsf{op}(L^2)}^{1-s}\right]\right\},$$

and when $s = 1$,

$$\mathbb{E}[\|\hat{\beta}_{t+1} - \beta^*\|_K^2] \leq C^K \left[\left(\sum_{k=1}^t \eta_k\right)^{-2r} + \frac{1}{2}\sum_{k=1}^t \eta_k^2\right], \tag{6.4}$$

where $C^K = \max\left\{\|g^\dagger\|_2^2((r/e)^{2r} + \|\mathscr{L}_C\|_{\mathsf{op}(L^2)}^{2r}), 4(\sqrt{c_{\mathsf{M}}}\|\beta^*\|_2^2 + \sigma^2)\sqrt{c_{\mathsf{M}}}\mathrm{Tr}(\mathscr{L}_C)\right\}$. Bounds (6.3) and (6.4) are unified by abusing the notation and denoting $0^0 = 1$ (so as to make $(\sum_{j=k+1}^t \eta_j)^0 = 1$ even when the sum is zero). The proof is then completed. □

We are at the position of proving Theorems 3 and 4 as corollaries of Theorem 13.

Proof of Theorem 3. To apply Theorem 13, we need only to select a proper bound C_3^{S} of step-sizes, to guarantee (6.1), i.e., for $k = 1, \ldots, t-1$,

$$c_{\mathsf{M}}\mathrm{Tr}(\mathscr{L}_K^s)\left[\left(\frac{2-s}{2e}\right)^{2-s} + \kappa^{4-2s}\right]\sum_{l=1}^k \frac{\eta_l^2}{1 + \left(\sum_{j=l+1}^k \eta_j\right)^{2-s}} \leq \frac{1}{2}. \tag{6.5}$$

To bound the sum in (6.5), we apply Lemma 12 with $\nu = 2 - s > 1$ and note that $0 < \theta \leq 1/2$, so $\omega = -\theta < 0$. We have

$$\sum_{l=1}^k \frac{\eta_l^2}{1 + \left(\sum_{j=l+1}^k \eta_j\right)^{2-s}} \leq C^{\mathsf{OL}}(\nu = 2 - s, \theta)(k+1)^{-\theta} \leq \eta_0^s\left(\frac{4^\theta}{\log 2} + 9^\theta C_0^{\mathsf{OL}}\right),$$

where the last inequality is just (5.19). Therefore, to achieve (6.5) (or equivalently, (6.1) for Theorem 13), we simply need to let

$$C_3^{\mathsf{S}} = \left\{ 2c_{\mathsf{M}} \mathrm{Tr}(\mathscr{L}_K^s) \left[\left(\frac{2-s}{2e} \right)^{2-s} + \kappa^{4-2s} \right] \left(\frac{4^\theta}{\log 2} + 9^\theta C_0^{\mathsf{OL}} \right) \right\}^{-1/s}. \tag{6.6}$$

By Theorem 13,

$$\mathbb{E}[\|\hat{\beta}_{t+1} - \beta^*\|_K^2] \le C^K \left[\left(\sum_{k=1}^t \eta_k \right)^{-2r} + \sum_{k=1}^t \frac{\eta_k^2}{1 + \left(\sum_{j=k+1}^t \eta_j \right)^{1-s}} \right]. \tag{6.7}$$

We bound the first term in the right-hand side of (6.7) by Lemma 12 with $\nu = 2r$,

$$\left(\sum_{k=1}^t \eta_k \right)^{-2r} \le \left(\frac{\eta_0(1-2^{\theta-1})}{1-\theta} \right)^{-2r} (t+1)^{-2r(1-\theta)}. \tag{6.8}$$

We bound the last sum in (6.7) by Lemma 12. Note that now $\nu = 1 - s \in (0,1)$ and $0 < \theta \le 1/2$, so for Lemma 12,

$$\omega = 1 - 2\theta - \nu + \nu\theta = s(1-\theta) - \theta = \begin{cases} -\dfrac{2r}{2r+s+1}, & 2r < 1-s, \\ -(1-s)/2, & 2r \ge 1-s. \end{cases}$$

From the definition of θ, we see that $(\nu, \theta) \in \Omega$ if and only if $\theta = 1/2$, which is equivalent to $2r \ge 1 - s$. So,

$$\sum_{k=1}^t \frac{\eta_k^2}{1 + \left(\sum_{j=k+1}^t \eta_j \right)^{1-s}} \le C^{\mathsf{OL}} \begin{cases} (t+1)^{-2r/(2r+s+1)}, & 2r < 1-s, \\ (t+1)^{-(1-s)/2} \log(t+1), & 2r \ge 1-s. \end{cases} \tag{6.9}$$

We now show that the rates of (6.8) is no slower than that of (6.9). In fact, when $2r < 1-s$, $-2r(1-\theta) = -\frac{2r}{2r+s+1}$. When $2r \ge 1-s$, $-2r(1-\theta) = -r \le -(1-s)/2$. We have proved that

$$\mathbb{E}\left[\left\| \hat{\beta}_{t+1} - \beta^* \right\|_K^2 \right] \le C_3 \begin{cases} (t+1)^{-2r/(2r+s+1)}, & 2r < 1-s, \\ (t+1)^{-(1-s)/2} \log(t+1), & 2r \ge 1-s, \end{cases}$$

where

$$C_3 = \frac{C^K}{\log 2} \left(\frac{\eta_0(1-2^{\theta-1})}{1-\theta} \right)^{-2r} + C^K C^{\mathsf{OL}}. \tag{6.10}$$

□

Proof of Theorem 4. First, we specify that

$$C_4^{\mathsf{S}} = \left[2c_{\mathsf{M}} \mathrm{Tr}(\mathscr{L}_K^s) \left(\left(\frac{2-s}{2e} \right)^{2-s} + \kappa^{4-2s} \right) C_{4*}^{\mathsf{S}} \right]^{-1}, \tag{6.11}$$

where C_{4*}^{S} is specified in (6.12) below. Then, we verify bound (6.1) of Theorem 13. To this end, we substitute $\eta_t = \eta_0 T^{-\theta}$ with $\theta = (s+2r)/(1+s+2r)$. Note that $2-s \geq 1$, $\eta_0 \leq 1$, and $k \leq T-1$.

$$\begin{aligned}
&\sum_{l=1}^{k} \frac{\eta_l^2}{1+(\sum_{j=l+1}^{k} \eta_j)^{2-s}} = \eta_0^2 T^{-2\theta} + \sum_{l=1}^{k-1} \frac{\eta_0^2 T^{-2\theta}}{1+(\eta_0 l T^{-\theta})^{2-s}} \\
\leq & \eta_0^2 T^{-2\theta} + \eta_0 T^{-\theta} \int_0^{T-2} \frac{\eta_0 T^{-\theta} du}{1+(\eta_0 T^{-\theta} u)^{2-s}} \leq \eta_0^2 T^{-2\theta} + \eta_0 T^{-\theta} \int_0^{T^{1-\theta}} \frac{du}{1+u^{2-s}} \\
\leq & \eta_0^2 T^{-2\theta} + \eta_0 T^{-\theta} \begin{cases} \frac{2-s}{1-s}, & \text{when } 0 < s < 1, \\ \log(T^{1-\theta}+1), & \text{when } s = 1, \end{cases} \\
\leq & \eta_0 C_{4*}^{\mathsf{S}} := \eta_0 \begin{cases} 2 + \frac{1}{1-s}, & \text{when } 0 < s < 1, \\ 1 + \log 2 + \frac{1-\theta}{e\theta}, & \text{when } s = 1, \end{cases}
\end{aligned} \tag{6.12}$$

where in the last inequality we used (A.6) and $T^{-\theta}\log(T^{1-\theta}+1) \leq T^{-\theta}\log(2T^{1-\theta}) \leq T^{-\theta}\log 2 + T^{-\theta}(1-\theta)\log T \leq \log 2 + \frac{1-\theta}{e\theta}$. So (6.1) is verified. Then by Theorem 13,

$$\mathbb{E}[\|\hat{\beta}_{T+1} - \beta^*\|_K^2] \leq C^K \left[(\eta_0 T^{1-\theta})^{-2r} + \sum_{k=1}^{T} \frac{\eta_0^2 T^{-2\theta}}{1+(\eta_0 T^{-\theta}(T-k))^{1-s}} \right]. \tag{6.13}$$

We now estimate the last sum in (6.13). When $0 < s < 1$,

$$\begin{aligned}
&\sum_{k=1}^{T} \frac{\eta_0^2 T^{-2\theta}}{1+(\eta_0 T^{-\theta}(T-k))^{1-s}} \leq \eta_0^2 T^{-2\theta} + \eta_0 T^{-\theta} \int_0^{T-1} \frac{\eta_0 T^{-\theta} du}{1+(\eta_0 T^{-\theta} u)^{1-s}} \\
\leq & \eta_0 T^{-\theta} + \eta_0 T^{-\theta} \int_0^{\eta_0 T^{1-\theta}} \frac{du}{1+u^{1-s}} \leq 2\eta_0 T^{-\theta} + \eta_0 T^{-\theta} \int_1^{T^{1-\theta}} u^{s-1} du \\
\leq & 2\eta_0 T^{-\theta} + \frac{\eta_0}{s} T^{-\theta+s(1-\theta)} \leq \eta_0 (2+s^{-1}) T^{-2r/(1+s+2r)},
\end{aligned} \tag{6.14}$$

where in the last step we used the definition $\theta = (s+2r)/(1+s+2r)$ we made above. When $s = 1$, $\theta = (1+2r)/(2+2r)$. Recall the definition $(\sum \eta_j)^0 := 1$ in Theorem 13 even when the sum vanishes. We have

$$\sum_{k=1}^{T} \frac{\eta_0^2 T^{-2\theta}}{1+(\eta_0 T^{-\theta}(T-k))^{1-s}} = \frac{\eta_0^2}{2} T^{1-2\theta} = \frac{\eta_0^2}{2} T^{-2r/(1+s+2r)}. \tag{6.15}$$

We summarize the above analysis to obtain

$$\mathbb{E}\left[\left\|\hat{\beta}_{T+1}-\beta^*\right\|_K^2\right] \le C_4 T^{-2r/(1+s+2r)},$$

where

$$C_4 = C^K \eta_0^{-2r} + C^K \eta_0 \begin{cases} 2+s^{-1}, & \text{when } 0<s<1, \\ \eta_0/2, & \text{when } s=1. \end{cases} \tag{6.16}$$

The proof of Theorem 4 is complete. □

7 A Numerical Experiment

In this section, we provide a simple numerical example with artificial data, to verify the feasibility of our assumptions, and the empirical performance of the algorithm (1.3). Settings of the example are taken from [5].

We set $\mathcal{T} = [0,1]$. The functional linear model (1.1) is specified by the Gaussian process

$$X(u) = \sum_{k=1}^{N} \frac{\sqrt{2}Z_k}{k^\alpha}\cos(k\pi u), \quad 0 \le u \le 1, \tag{7.1}$$

where N is a large integer to be specified later, $Z_1, \ldots, Z_N \sim \mathcal{N}(0,1)$ are independent standard normal random variables, and $\alpha > 0$. We employ the reproducing kernel used in [5],

$$\begin{aligned} K(u,v) &= -\frac{1}{3}B_4\left(\frac{|u-v|}{2}\right) - \frac{1}{3}B_4\left(\frac{u+v}{2}\right) \\ &= \sum_{k=1}^{\infty} \frac{2}{(k\pi)^4}\cos(k\pi u)\cos(k\pi v), \quad u,v \in [0,1], \end{aligned}$$

where $B_4(x) = x^4 - 2x^3 + x^2 - \frac{1}{30}$ is the 4-th Bernoulli polynomial. Write $\mathcal{H}_K$ the associated RKHS. It is easy to verify that $\{\sqrt{2}\cos(k\pi x)\}_{k=1}^{\infty}$ is an orthogonal set in $\mathcal{H}_K$, and is also an orthonormal set of $L^2(\mathcal{T})$. We define the coefficient isometry

$$\mathcal{M} : \mathbb{R}^N \to L^2(\mathcal{T}), \quad \boldsymbol{u} = (u_1, \ldots, u_N)^T \mapsto \sum_{k=1}^{N} \sqrt{2}u_k \cos(k\pi x).$$

Let $\langle\cdot,\cdot\rangle_{\mathrm{E}}$ and $\|\cdot\|_{\mathrm{E}}$ denote the Euclidean inner product and norm in $\mathbb{R}^N$, respectively. It is obvious that $\langle\mathcal{M}(\boldsymbol{u}),\mathcal{M}(\boldsymbol{v})\rangle_2 = \langle\boldsymbol{u},\boldsymbol{v}\rangle_{\mathrm{E}}$. Write $\boldsymbol{K} = \pi^{-4}(1^{-4},2^{-4},\ldots,N^{-4})^T$. Then,

$$L_K\mathcal{M}(\boldsymbol{u}) = \mathcal{M}(\boldsymbol{K}\circ\boldsymbol{u}), \quad \text{for any } \boldsymbol{u}\in\mathbb{R}^N,$$

where for any vectors $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^T$, $\boldsymbol{u}\circ\boldsymbol{v} = (u_1v_1,\ldots,u_Nv_N)^T$ is the component-wise vector product.

From (7.1), the covariance function is

$$C(u,v) = \mathbb{E}[X(u)X(v)] = \sum_{k=1}^{N}\frac{2}{k^{2\alpha}}\cos(k\pi u)\cos(k\pi v).$$

Write $\boldsymbol{C}^r = (1^{-2\alpha r},2^{-2\alpha r},\ldots,N^{-2\alpha r})^T$ for $r>0$. We see that L_C and L_K are commutable and $L_C\mathcal{M}(\boldsymbol{u}) = \mathcal{M}(\boldsymbol{C}\circ\boldsymbol{u})$. Define the oracle slope function $\beta^* = \mathcal{M}(\boldsymbol{\beta}^*)$. Let $\boldsymbol{g}\in\mathbb{R}^N$ with $\|\boldsymbol{g}\|_{\mathrm{E}} = 1$. We set $\boldsymbol{\beta}^* = \boldsymbol{K}^r\circ\boldsymbol{C}^{r-\frac{1}{2}}\circ\boldsymbol{g}$ to meet Assumption 1. Obviously, $\mathrm{Tr}(\mathscr{L}_K^s) = \pi^{-4s}\sum_{k=1}^{N}k^{-2\alpha s-4s}$. From the setting that N is large, it is reasonable to interpret the assumption $\mathrm{Tr}(\mathscr{L}_K^s)<\infty$ as $-2\alpha s-4s<-1$. So, Assumption 2 is satisfied when $\alpha>\max\{0,(1-4s)/(2s)\}$. Assumption 3 with $c_{\mathsf{M}}=3$ is guaranteed thanks to the Gaussian process setting (7.1).

Now we construct the artificial data. For $t\geq 1$, let $\boldsymbol{x}_t$ be an independent copy of the coefficient vector $(1^{-\alpha}Z_1,\ldots,N^{-\alpha}Z_N)^T$ of X. Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be drawn independently from $\mathcal{N}(0,\sigma^2)$. Then $y_i = \langle x_t,\beta^*\rangle_2+\varepsilon_t = \langle\boldsymbol{x}_t,\boldsymbol{\beta}^*\rangle_{\mathrm{E}}+\varepsilon_t$. The iterative algorithm (1.5) can be formulated in $\mathbb{R}^N$, and one just needs to map the output back to $\mathcal{H}_K$ through $\mathcal{M}$. We use $\hat{\beta}_t = \mathcal{M}(\boldsymbol{\beta}_t)$ and rewrite (1.5) as

$$\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^* = (I-\eta_tA_t)(\boldsymbol{\beta}_t-\boldsymbol{\beta}^*)+\eta_t\varepsilon_t\boldsymbol{K}\circ\boldsymbol{x}_t,$$

where I is the identity matrix in $\mathbb{R}^N$, and $A_t = (\boldsymbol{K}\circ\boldsymbol{x}_t)\boldsymbol{x}_t^T$ is a rank-one positive semi-definite matrix. Since the oracle function β^* is available, the excess generalization error is estimated by

$$\mathcal{E}(\hat{\varphi}_{t+1}) = \left\|L_C^{1/2}(\hat{\beta}_{t+1}-\beta^*)\right\|_2^2 = \left\|\boldsymbol{C}^{1/2}\circ(\boldsymbol{\beta}_{t+1}-\boldsymbol{\beta}^*)\right\|_{\mathrm{E}}^2.$$

The simulation results are visualized in Figure 4.

Acknowledgments

Part of the work of Xin Guo was done when he worked at The Hong Kong Polytechnic University and supported partially by the Research Grants Council of Hong Kong [Project

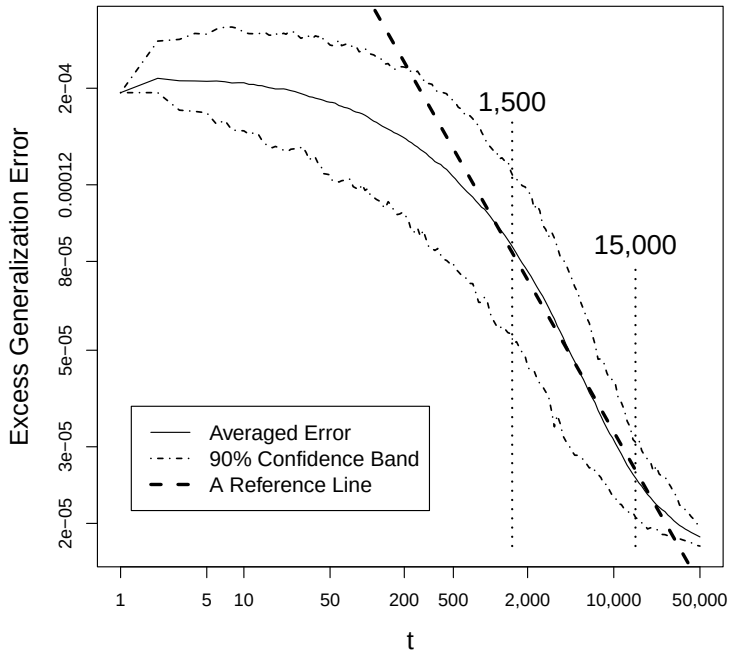

Figure 4: Excess generalization error of Algorithm (1.3) on artificial data. We set $N = 20$, $s = 0.7$, $r = 0.5$, and $\alpha = 0.01$. The experiment is repeated 100 times, and the averaged excess generalization error $\mathcal{E}[\hat{\varphi}_t]$ is plotted as the solid line as a function of t. We further visualize the 90% empirical confidence band by dash-dotted lines (i.e., the error curves of 90% experiments fall onto the band). To better demonstrate the polynomial convergence rate, logarithmic scale is used for both axes. A reference dashed line with slope -0.5 is employed (corresponding to the theoretical convergence rate $O(t^{-0.5})$ predicted by Theorem 1). This reference line demonstrates that right after about 1,500 burn-in iterations, the observed rate of error decay matches the theoretical estimation. The error decay slows down and drifts away from the theoretical estimation later after $t > 15,000$, because of numerical error.

No. PolyU 15305018]. The work of Zheng-Chu Guo is supported by Zhejiang Provincial Natural Science Foundation of China [Project No. LR20A010001], National Natural Science Foundation of China [Project No. U21A20426, No. 12271473], and Fundamental Research Funds for the Central Universities [Project No. 2021XZZX001]. The work of Lei Shi is supported partially by Shanghai Science and Technology Program [Project No. 21JC1400600 and Project No. 20JC1412700] and National Natural Science Foundation of China [Grant No. 12171093]. All authors contributed equally to this work and are listed

alphabetically. The corresponding author is Lei Shi.

References

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

[2] Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

[3] Gilles Blanchard and Nicole Krämer. Optimal learning rates for kernel conjugate gradient regression. *Advances in neural information processing systems*, 23, 2010.

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

[5] T. 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] Xiaming Chen, Bohao Tang, Jun Fan, and Xin Guo. Online gradient descent algorithms for functional data learning. *J. Complexity*, 70(101635):1–14, 2022.

[8] 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.

[9] Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 2016.

[10] Jun Fan, Fusheng Lv, and Lei Shi. An RKHS approach to estimate individualized treatment rules based on functional predictors. *Math. Found. Comput.*, 2(2):169–181, 2019.

[11] Xin Guo, Junhong Lin, and Ding-Xuan Zhou. Rates of convergence of randomized Kaczmarz algorithms in Hilbert spaces. *Appl. Comput. Harmon. Anal.*, 61:288–318, 2022.

[12] Zheng-Chu Guo, Andreas Christmann, and Lei Shi. Optimality of robust online learning. *arXiv preprint arXiv:2304.10060*, 2023.

[13] Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Probl.*, 33(7):074009, 29, 2017.

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

[15] Zheng-Chu Guo and Lei Shi. Optimal rates for coefficient-based regularized regression. *Appl. Comput. Harmon. Anal.*, 47(3):662–701, 2019.

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

[17] Xuqing He and Hongwei Sun. Error analysis of classification learning algorithms based on lums loss. *Math. Found. Comput.*, published online first, 2022.

[18] Junhong Lin and Lorenzo Rosasco. Generalization properties of doubly stochastic learning algorithms. *J. Complexity*, 47:42–61, 2018.

[19] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18(1):3202–3232, 2017.

[20] Jiading Liu and Lei Shi. Statistical optimality of divide and conquer kernel-based functional linear regression. *arXiv preprint arXiv:2211.10968*, 2022.

[21] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] James O. Ramsay and Bernard W. Silverman. *Fitting differential equations to functional data: Principal differential analysis*. Springer, 2005.

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

[24] Shuhua Wang and Baohuai Sheng. Error analysis of kernel regularized pairwise learning with a strongly convex loss. *Math. Found. Comput.*, published online first, 2022.

[25] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

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

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

[28] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098, 2005.

A Appendix: A Technical Lemma

In this section of Appendix, we include the following Lemma 14, which is commonly used in the literature [14, 7, 26, 11] with smaller domains of parameters. Lemma 14 covers the whole domain $(\nu,\theta)\in(0,\infty)\times(0,1)$, and the proof is not elsewhere available to our best knowledge. We use Figure 5 to elucidate the rates in Lemma 14. Figure 5 is also helpful for understanding the rates in Lemma 12, and Theorems 1 and 3.

Lemma 14. *For $b\geq 2$, $0<\theta<1$, and $\nu>0$,*

$$\int_1^b \frac{u^{-2\theta}du}{1+(b^{1-\theta}-u^{1-\theta})^\nu} \leq C_0^{\mathsf{OL}} \begin{cases} b^\omega \log b, & \theta=\frac{1}{2} \text{ and } \nu\leq 1, \text{ or, } \theta\leq\frac{1}{2} \text{ and } \nu=1, \\ b^\omega, & \text{otherwise,} \end{cases} \tag{A.1}$$

where C_0^{OL} is independent of b, and

$$\omega=\omega(\nu,\theta)=\begin{cases} 1-2\theta-\nu+\nu\theta, & 0<\nu\leq 1 \text{ and } 0<\theta\leq 1/2, \\ -\theta, & \nu\geq 1 \text{ and } 0<\theta\leq \nu/(\nu+1), \\ -\nu(1-\theta), & 1/2\leq\theta<1 \text{ and } \theta\geq\nu/(\nu+1). \end{cases} \tag{A.2}$$

In particular, when $\nu\geq 1$, $\omega=-\min\{\theta,\nu(1-\theta)\}$. For different combinations of parameters ν and θ, the constant $C_0^{\mathsf{OL}}=C_0^{\mathsf{OL}}(\nu,\theta)$ will be specified below in (A.7), (A.8), (A.9), (A.10), and (A.11), respectively.

We purposely allow the domains in (A.2) to overlap, to facilitate the applications. In spite of the piecewise definition, $\omega(\nu,\theta)$ is a continuous function on $(0,\infty)\times(0,1)$. We demonstrate the structure of ω in Figure 5. The estimate in Lemma 14 is tight, that is, we can reverse the order of the inequality (A.1) by replacing C_0^{OL} with a smaller positive constant independent of b. We skip the discussion of tightness.

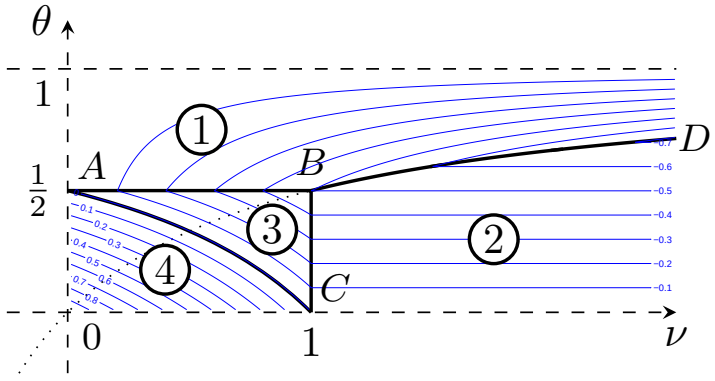

Figure 5: Summary of the convergence rates $\omega(\nu,\theta)$. The domain $(0,\infty)\times(0,1)$ is divided into four regimes by the solid black lines. Contours of ω are given in blue lines. In Regime 1, $\omega=-\nu(1-\theta)$. In Regime 2, $\omega=-\theta$. In Regimes 3 and 4, $\omega=1-2\theta-\nu+\nu\theta=-\nu(1-\theta)+(1-2\theta)=-\theta+(1-\theta)(1-\nu)$. The values of ω continuously extend to the boundaries between regimes. Arc BD is from the hyperbola $\theta=\nu/(\nu+1)$ which is extended by the dotted line. ω approaches its infimum along the ridge $B\to D$. Logarithm factor in (A.1) only appears on the line segments AB and BC including point B. In Regime 4 including Arc AC (which is $\theta=(1-\nu)/(2-\nu)$), $\omega\geq 0$ and the integral in (A.1) does not converge to zero as $b\to\infty$.

Proof of Lemma 14. To verify the estimate (A.1), we divide the integral interval into $[1,b/2]$ and $[b/2,b]$, and denote Υ_1^{OL} and Υ_2^{OL} the associated parts of the integral in (A.1),

respectively. First,

$$\begin{aligned}\Upsilon_1^{\mathsf{OL}} \leq & \frac{1}{1+(b^{1-\theta}-(b/2)^{1-\theta})^{\nu}} \int_1^{b/2} u^{-2\theta} du \\ \leq & \frac{b^{-\nu(1-\theta)}}{(1-2^{\theta-1})^{\nu}} \times \begin{cases} \frac{(b/2)^{1-2\theta}}{1-2\theta}, & \text{when } 0<\theta<1/2, \\ \log \frac{b}{2}, & \text{when } \theta = 1/2, \\ \frac{1}{2\theta-1}, & \text{when } 1/2<\theta<1, \end{cases} \\ \leq & \frac{1}{(1-2^{\theta-1})^{\nu}} \times \begin{cases} \frac{2^{2\theta-1}}{1-2\theta} b^{1-2\theta-\nu+\nu\theta}, & 0<\theta<1/2, \\ b^{-\nu/2}\log b, & \theta = 1/2, \\ \frac{1}{2\theta-1} b^{-\nu(1-\theta)}, & 1/2<\theta<1. \end{cases} \end{aligned} \tag{A.3}$$

Second, to estimate Υ_2^{OL}, we change the variable as $\xi = b^{1-\theta} - u^{1-\theta}$ to give $d\xi = -(1-\theta)u^{-\theta}du$. Therefore,

$$\begin{aligned}\Upsilon_2^{\mathsf{OL}} = & \int_{b/2}^{b} \frac{u^{-2\theta}du}{1+(b^{1-\theta}-u^{1-\theta})^{\nu}} = \int_0^{b^{1-\theta}-(b/2)^{1-\theta}} \frac{u^{-\theta}d\xi}{(1+\xi^{\nu})(1-\theta)} \\ \leq & \frac{(b/2)^{-\theta}}{1-\theta} \int_0^{b^{1-\theta}} \frac{d\xi}{1+\xi^{\nu}}. \end{aligned}$$

Recall that for any $\nu > 0$ and $\tau \geq 1$,

$$\int_0^{\tau} \frac{d\xi}{1+\xi^{\nu}} \leq 1 + \int_1^{\tau} \xi^{-\nu} d\xi \leq \begin{cases} 1+\frac{\tau^{1-\nu}-1}{1-\nu} \leq \frac{1}{1-\nu}\tau^{1-\nu}, & 0<\nu<1, \\ 1+\log\tau, & \nu = 1, \\ 1+\frac{1-\tau^{1-\nu}}{\nu-1} \leq \frac{\nu}{\nu-1}, & \nu > 1. \end{cases} \tag{A.4}$$

Therefore,

$$\Upsilon_2^{\mathsf{OL}} \leq \begin{cases} \frac{2^{\theta}}{(1-\theta)(1-\nu)} b^{1-2\theta-\nu+\nu\theta}, & 0<\nu<1, \\ \frac{2^{\theta}}{1-\theta}\left(\frac{1}{\log 2}+1-\theta\right) b^{-\theta}\log b, & \nu = 1, \\ \frac{2^{\theta}\nu}{(1-\theta)(\nu-1)} b^{-\theta}, & \nu > 1. \end{cases} \tag{A.5}$$

Now we merge (A.3) and (A.5) to derive (A.1). Note that the bounds of Υ_1^{OL} are divided according to θ, while the bounds of Υ_2^{OL} are divided according to ν, so the merging appears complicated. Figure 5 provides a clear picture.

- *Case 1:* $\nu/(\nu+1) \leq \theta < 1$ and $\theta > 1/2$. This corresponds to Regime 1 in Figure 5, including the boundary BD but excluding line segment AB and point B. Below

we show that $\omega = -\nu(1-\theta)$. In fact, now $\Upsilon_1^{\mathsf{OL}} \le (1-2^{\theta-1})^{-\nu}(2\theta-1)^{-1}b^{-\nu(1-\theta)}$. When $0<\nu<1$, $\theta > 1/2$ implies $1-2\theta-\nu+\nu\theta < -\nu(1-\theta)$, so $\Upsilon_2^{\mathsf{OL}} \le 2^{\theta}(1-\theta)^{-1}(1-\nu)^{-1}b^{-\nu(1-\theta)}$. When $\nu = 1$, recall that

$$\max_{1\le u<\infty} u^{-a}\log u = \frac{1}{ea}, \quad \text{for any } a>0, \tag{A.6}$$

where the maximum is achieved at $u = e^{1/a}$. Since $\theta > 1/2$, $b^{-\theta+(1-\theta)}\log b \le \frac{1}{e(2\theta-1)}$, and we have

$$\Upsilon_2^{\mathsf{OL}} \le \frac{2^{\theta}}{1-\theta}\left(\frac{1}{\log 2}+1-\theta\right)\frac{b^{-(1-\theta)}}{e(2\theta-1)}.$$

When $\nu > 1$, the condition $\nu/(\nu+1) \le \theta$ implies $-\theta \le -\nu(1-\theta)$, so $\Upsilon_2^{\mathsf{OL}} \le 2^{\theta}\nu(1-\theta)^{-1}(\nu-1)^{-1}b^{-\nu(1-\theta)}$. We have proved that

$$\int_1^b \frac{u^{-2\theta}du}{1+(b^{1-\theta}-u^{1-\theta})^{\nu}} = \Upsilon_1^{\mathsf{OL}} + \Upsilon_2^{\mathsf{OL}} \le C_0^{\mathsf{OL}} b^{-\nu(1-\theta)},$$

with

$$C_0^{\mathsf{OL}} = \frac{(1-2^{\theta-1})^{-\nu}}{(2\theta-1)} + \begin{cases} \dfrac{2^{\theta}(\nu+1)}{(1-\theta)|1-\nu|}, & \nu>0 \text{ and } \nu\neq 1, \\ \dfrac{2^{\theta}}{(1-\theta)(e(2\theta-1))}\left(\dfrac{1}{\log 2}+1-\theta\right), & \nu = 1. \end{cases} \tag{A.7}$$

- *Case 2:* $\nu > 1$ and $0 < \theta < \nu/(\nu+1)$. This corresponds to Regime 2 in Figure 5, excluding the boundaries BC and BD. Below we show that $\omega = -\theta$. In fact, in this regime $\Upsilon_2^{\mathsf{OL}} \le 2^{\theta}\nu(1-\theta)^{-1}(\nu-1)^{-1}b^{-\theta}$. When $0<\theta<1/2$, $1-2\theta-\nu+\nu\theta = -\theta+(1-\theta)(1-\nu) < -\theta$, so $\Upsilon_1^{\mathsf{OL}} \le \frac{2^{2\theta-1}}{(1-2^{\theta-1})^{\nu}(1-2\theta)}b^{-\theta}$. When $\theta = 1/2$, we use (A.6) to see $b^{-\frac{\nu}{2}+\frac{1}{2}}\log b \le \frac{2}{e(\nu-1)}$, so $\Upsilon_1^{\mathsf{OL}} \le (1-2^{\theta-1})^{-\nu}\frac{2}{e(\nu-1)}b^{-\theta}$. When $1/2<\theta<1$, $\theta < \nu/(\nu+1)$ implies $-\nu(1-\theta) < -\theta$, so $\Upsilon_1^{\mathsf{OL}} \le (1-2^{\theta-1})^{-\nu}(2\theta-1)^{-1}b^{-\theta}$. We have proved that

$$\int_1^b \frac{u^{-2\theta}du}{1+(b^{1-\theta}-u^{1-\theta})^{\nu}} = \Upsilon_1^{\mathsf{OL}} + \Upsilon_2^{\mathsf{OL}} \le C_0^{\mathsf{OL}} b^{-\theta},$$

with

$$C_0^{\mathsf{OL}} = \frac{2^{\theta}\nu}{(1-\theta)(\nu-1)} + \frac{1}{(1-2^{\theta-1})^{\nu}} \times \begin{cases} 2^{2\theta-1}/(1-2\theta), & 0<\theta<1/2, \\ \frac{2}{e(\nu-1)}, & \theta = 1/2, \\ 1/(2\theta-1), & 1/2<\theta<1. \end{cases} \tag{A.8}$$

- *Case 3:* $0 < \theta < 1/2$ and $0 < \nu < 1$. This corresponds to Regimes 3 and 4 in Figure 5, including Arc AC but excluding boundaries AB, BC, and point B. Now it is obvious that $\Upsilon_1^{\mathsf{OL}} + \Upsilon_2^{\mathsf{OL}} \leq C_0^{\mathsf{OL}} b^{1-2\theta-\nu+\nu\theta}$, with

$$C_0^{\mathsf{OL}} = \frac{2^{2\theta-1}}{(1-2^{\theta-1})^{\nu}(1-2\theta)} + \frac{2^{\theta}}{(1-\theta)(1-\nu)}. \tag{A.9}$$

- *Case 4:* $\theta = 1/2$ and $0 < \nu \leq 1$. This corresponds to line segment AB, including point B. Now $-\nu(1-\theta) = -\nu/2$, $b^{-\nu/2}\log b \geq b^{-\theta}\log b \geq b^{-\theta}\log 2$, and $1-2\theta-\nu+\nu\theta = -\nu(1-\theta)$. So

$$\Upsilon_2^{\mathsf{OL}} \leq b^{-\nu(1-\theta)}(\log b) \times \begin{cases} \dfrac{2^{\theta}}{1-\theta}\left(\dfrac{1}{\log 2} + 1 - \theta\right), & \nu = 1, \\ \dfrac{2^{\theta}}{(1-\theta)(1-\nu)\log 2}, & 0 < \nu < 1, \end{cases}$$

and $\Upsilon_1^{\mathsf{OL}} \leq (1-2^{\theta-1})^{-\nu} b^{-\nu/2}\log b$. So, $\Upsilon_1^{\mathsf{OL}} + \Upsilon_2^{\mathsf{OL}} \leq C_0^{\mathsf{OL}} b^{-\nu(1-\theta)}\log b$ with

$$C_0^{\mathsf{OL}} \leq \frac{1}{(1-2^{\theta-1})^{\nu}} + \begin{cases} \dfrac{2^{\theta}}{1-\theta}\left(\dfrac{1}{\log 2} + 1 - \theta\right), & \nu = 1, \\ \dfrac{2^{\theta}}{(1-\theta)(1-\nu)\log 2}, & 0 < \nu < 1. \end{cases} \tag{A.10}$$

- *Case 5:* $\nu = 1$ and $0 < \theta < 1/2$. This corresponds to line segment BC, excluding point B. In this case, $1-2\theta-\nu+\nu\theta = -\theta$. So $\Upsilon_1^{\mathsf{OL}} + \Upsilon_2^{\mathsf{OL}} \leq C_0^{\mathsf{OL}} b^{-\theta}\log b$ with

$$C_0^{\mathsf{OL}} = \frac{2^{\theta}}{1-\theta}\left(\frac{1}{\log 2} + 1 - \theta\right) + \frac{2^{2\theta-1}}{(1-2^{\theta-1})^{\nu}(1-2\theta)\log 2}. \tag{A.11}$$

□