Online gradient descent learning algorithm†

Yiming Ying and Massimiliano Pontil

Department of Computer Science, University College London

Gower Street, London, WC1E 6BT, England, UK

{y.ying, m.pontil}@cs.ucl.ac.uk

Abstract

This paper considers the least-square online gradient descent algorithm in a reproducing kernel Hilbert space (RKHS) without an explicit regularization term. We present a novel capacity independent approach to derive error bounds and convergence results for this algorithm. The essential element in our analysis is the interplay between the generalization error and a weighted cumulative error which we define in the paper. We show that, although the algorithm does not involve an explicit RKHS regularization term, choosing the step sizes appropriately can yield competitive error rates with those in the literature.

Short Title: Online gradient descent learning

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

AMS Subject Classification Numbers: 68Q32, 68T05, 62J02, 62L20.

† Contact author: Yiming Ying, Telephone: +44 (0)20 7679 0374, Fax: +44 (0)20 7387 1397

1 Introduction

Let X be a compact subset of Euclidean space $\mathbb{R}^d$, Y a subset in $\mathbb{R}$ and $\mathbb{N}_\ell := \{1, \ldots, \ell\}$ for any $\ell \in \mathbb{N}$. Let ρ be a fixed but unknown probability measure on $Z := X \times Y$. The *generalization error* for a function $f : X \to Y$ is defined as

$$\mathcal{E}(f) := \int_Z \big(f(x) - y\big)^2 d\rho(z). \tag{1.1}$$

The function minimizing the above error is called the *regression function* and is given by

$$f_\rho(x) = \int_Y y d\rho(y|x), \qquad x \in X, \tag{1.2}$$

where $\rho(\cdot|x)$ is the conditional probability measure at x induced by ρ.

We consider the problem of approximating the regression function from a set of training data drawn from ρ. In this paper, we restrict our attention to online learning algorithms for computing an approximator of f_ρ in a *reproducing kernel Hilbert space* (RKHS).

Let $K : X \times X \to \mathbb{R}$ be a *Mercer kernel*, that is, a continuous, symmetric and positive semi-definite function, see e.g. [12]. The RKHS $\mathcal{H}_K$ associated with K is defined [1] to be the completion of the linear span of the set of functions $\{K_x(\cdot) := K(x, \cdot) : x \in X\}$ with inner product satisfying, for any $x \in X$ and $g \in \mathcal{H}_K$, the *reproducing property*

$$\langle K_x, g \rangle_K = g(x). \tag{1.3}$$

Let $\{z_t = (x_t, y_t) : t \in \mathbb{N}\}$ be a sequence of random samples independently distributed according to ρ. *The online gradient descent algorithm* [10, 19, 22, 28] is defined as

$$\begin{cases} f_1 = 0, \text{ and for } t \in \mathbb{N} \\ f_{t+1} = f_t - \eta_t\big((f_t(x_t) - y_t)K_{x_t} + \lambda f_t\big), \end{cases} \tag{1.4}$$

where $\lambda \geq 0$ is called the *regularization parameter*. We call the sequence $\{\eta_t : t \in \mathbb{N}\}$ the *step sizes* or *learning rates* and $\{f_t : t \in \mathbb{N}\}$ the *learning sequence*. Clearly, each output f_{t+1} only depends only on $\{z_j : j \in \mathbb{N}_t\}$.

The class of learning algorithms displayed above is also referred to as stochastic approximation algorithms in the setting of RKHS's. Such a stochastic approximation procedure dates back to [21]. One can see [22, 28, 32, 35] and references therein for more background material.

When the parameter $\lambda > 0$, we call (1.4) the *online regularized algorithm*. This algorithm has been well studied in the recent literature, see e.g. [20, 22, 33]. In this paper, we are mainly concerned with the online gradient descent algorithm *without* an explicit RKHS regularization term, which is given by (1.4) with $\lambda = 0$, that is,

$$\begin{cases} f_1 = 0, \text{ and for } t \in \mathbb{N} \\ f_{t+1} = f_t - \eta_t (f_t(x_t) - y_t) K_{x_t}. \end{cases} \tag{1.5}$$

We study the approximation of f_{T+1}, the output of iteration step (sample size) T, to the regression function f_ρ. The nature of the least-square loss leads to the measurement in the metric of $\mathcal{L}^2_{\rho_X}$ defined as $\|f\|_\rho = \|f\|_{\mathcal{L}^2_{\rho_X}} := \left(\int_X |f(x)|^2 d\rho_X\right)^{\frac{1}{2}}$, where ρ_X is the marginal distribution of ρ on X. A direct computation yields that

$$\|f_{T+1} - f_\rho\|^2_\rho = \mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho), \tag{1.6}$$

see e.g. [12] for a proof.

Our primary goal is to estimate the error (1.6) for the least-square online algorithm (1.5) by means of properties of ρ and K. We shall show how the choice of the step sizes in the algorithm affects its generalization error. Moreover, we shall derive error rates for algorithm (1.5) which are competitive with those in the literature. We mainly focus on two different types of step sizes. The first one is a *universal* polynomially decaying sequence of the form $\{\eta_t = O(t^{-\theta}),\ t \in \mathbb{N}\}$ with $\theta \in (0, 1)$. The second type of step sizes is of the form $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ with $\eta = \eta(T)$ depending on the iteration number (sample size) T. As we shall discuss in Section 2, the function $\eta(T)$ provides a trade-off between the learning rate η and the iteration steps T and its choice ensures convergence of the algorithm. This is in the spirit of early stopping implicit regularization, see e.g. [32, 36].

The rest of the paper is organized as follows. The next section summarizes our main results and Section 3 collects discussions on related work. To formulate our basic ideas, in Section 4 we provide a novel approach to the error analysis of the online algorithm (1.4) under some conditions on the step sizes. The essential element in our analysis is an appealing relation between the generalization error and a weighted form of cumulative sample error (see Proposition 1) in online learning literature (e.g. [8, 10]). In Section 5, we develop error bounds and convergence results for the online algorithm (1.5). Finally, Section 6 establishes explicit error rates for

algorithm (1.5) and, as a byproduct, improves the preceding rates [22, 33] for the online regularized algorithm given by (1.4) with $\lambda > 0$.

2 Main results

We begin with some background material and notations. Firstly, let $\mathcal{C}(X)$ be the space of continuous functions on X with the norm $\|\cdot\|_\infty$, $\kappa := \sup_{x\in X}\sqrt{K(x,x)}$ and note, by the reproducing property (1.3), for every $f \in \mathcal{H}_K$, that

$$\|f\|_\infty \leq \kappa \|f\|_K. \tag{2.1}$$

In addition, for every $t \in \mathbb{N}$ we denote the expectation $\mathbb{E}_{z_1,\ldots,z_t}$ as $\mathbb{E}_{Z^t}$ and use the convention $\mathbb{E}_{Z^0}\big[\xi\big] = \xi$ for any random variable ξ. Secondly, we introduce the notion of $\mathcal{K}$-functional [5] which plays a central role in approximation theory, namely

$$\mathcal{K}(s, f_\rho) := \inf_{f\in\mathcal{H}_K}\{\|f - f_\rho\|_\rho + s\|f\|_K\}, \quad s > 0. \tag{2.2}$$

Next, we define, for any $\theta \in (0,1)$, the quantity

$$\mu(\theta) := \begin{cases} 1248(1+\kappa)^4 & \text{if } \theta = \frac{1}{2}, \\ \frac{208(1+\kappa)^4}{(1-2^{\theta-1})|2\theta-1|}\Big(\ln\big(\frac{8}{1-\theta}\big) + \frac{1}{\min\{\theta,1-\theta\}}\Big) & \text{otherwise.} \end{cases} \tag{2.3}$$

Finally, we require that $\int_Z y^2 d\rho(z) < \infty$ which implies that the quantity $\mathcal{E}(f_\rho) + \|f_\rho\|_\rho^2$ is finite.

We are now ready to state our main results. The first one deals with error bounds for the online gradient descent algorithm (1.5).

Theorem 1. *Let $\theta \in (0,1)$ and $\{\eta_t = \frac{1}{\mu}t^{-\theta} : t \in \mathbb{N}\}$ with some constant $\mu \geq \mu(\theta)$. Define $\{f_t : t \in \mathbb{N}\}$ by (1.5). Then, for any $T \in \mathbb{N}$, $\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big]$ is bounded by*

$$\Big[\mathcal{K}(b_\theta\sqrt{\mu}\; T^{-(1-\theta)/2}, f_\rho)\Big]^2 + \frac{c_\rho c_\theta}{\mu}\; T^{-\min\{\theta,1-\theta\}}\ln\big(\frac{8T}{1-\theta}\big), \tag{2.4}$$

where $c_\rho := 4(1+\kappa)^4\big(20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)\big)$, $b_\theta := 2(1+\kappa)\sqrt{\frac{1-\theta}{1-2^{\theta-1}}}$ and $c_\theta = 13$ if $\theta = \frac{1}{2}$ and $\frac{13}{(1-2^{\theta-1})|1-2\theta|}$ otherwise.

The $\mathcal{K}$-functional term of our upper bound in Theorem 1 concerns the approximation of the function f_ρ in the space $\mathcal{L}^2_{\rho_X}$ by functions in the

space $\mathcal{H}_K$. Its polynomial decay can be characterized by requiring that f_ρ lies in the interpolation space $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$ defined as

$$(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty} := \big\{ f \in \mathcal{L}^2_{\rho_X} : \|f\|_{\beta,\infty} = \sup_{s>0} s^{-\beta} \mathcal{K}(s, f) < \infty \big\}, \qquad (2.5)$$

where $\beta \in [0, 1]$. Indeed, it is easy to see that, for some $c > 0$, $\mathcal{K}(s, f_\rho) \leq cs^\beta$ for any $s > 0$ if and only if $f_\rho \in (\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$. Note [5] that $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{0,\infty} = \mathcal{L}^2_{\rho_X}$ and $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{1,\infty} = \mathcal{H}_K$. Therefore, we can intuitively regard the interpolation space $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$ as an intermediate space between the metric space $\mathcal{L}^2_{\rho_X}$ and the much smaller approximation space $\mathcal{H}_K$. Similar ideas of using the $\mathcal{K}$-functional to characterize the approximation property of the space $\mathcal{H}_K$ can be found in [11, 12, 23].

In general, we have [5] that $\lim_{s\to 0+} \mathcal{K}(s, f_\rho) = \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho$. This gives rise to convergence of the online gradient descent algorithm (1.5).

Theorem 2. *Let $\theta \in (0, 1)$ and $\mu(\theta)$ be given by (2.3). Define $\{f_t : t \in \mathbb{N}\}$ by (1.5). If the step sizes satisfy $\eta_t = \frac{1}{\mu} t^{-\theta}$ for $t \in \mathbb{N}$ with some constant $\mu \geq \mu(\theta)$ then we have that*

$$\lim_{T\to\infty} \mathbb{E}_{Z^T} \big[\|f_{T+1} - f_\rho\|^2_\rho \big] = \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|^2_\rho.$$

We will give the proofs of Theorems 1 and 2 in Section 5.

If the space $\mathcal{H}_K$ has a good approximation property then the limit in Theorem 2 equals zero. To see this, we say that $\mathcal{H}_K$ is dense in $\mathcal{L}^2_{\rho_X}$ if

$$\inf_{f\in\mathcal{H}_K} \|f - g\|_\rho = 0, \quad \text{for all} \quad g \in \mathcal{L}^2_{\rho_X}. \qquad (2.6)$$

We note that, in [26], a kernel is called *universal* if $\inf_{f\in\mathcal{H}_K} \|f - g\|_\infty = 0$, for all $g \in \mathcal{C}(X)$. For example, the Gaussian kernel $K_\sigma(x, x') = e^{-\frac{|x-x'|^2}{2\sigma^2}}$ is universal for every $\sigma > 0$. Universal kernels are sufficient to ensure our density condition (2.6), since $\|f\|_\rho \leq \|f\|_\infty$ and $\mathcal{C}(X)$ is dense in $\mathcal{L}^2_{\rho_X}$. Note that $f_\rho \in \mathcal{L}^2_{\rho_X}$. Therefore, an immediate consequence of Theorem 2 is the following corollary.

Corollary 1. *Suppose the assumptions in Theorem 2 hold true. Moreover, if $\mathcal{H}_K$ is dense in $\mathcal{L}^2_{\rho_X}$ then we have that*

$$\lim_{T\to\infty} \mathbb{E}_{Z^T} \big[\|f_{T+1} - f_\rho\|^2_\rho \big] = 0.$$

By choosing θ appropriately, we can get explicit error rates if the regression function f_ρ lies in the regularity space $L_K^\beta(\mathcal{L}^2_{\rho_X})$. To define this space, we introduce the integral operator $L_K : \mathcal{L}^2_{\rho_X} \to \mathcal{L}^2_{\rho_X}$ defined as

$$L_K f(x) := \int_X K(x,x')f(x')d\rho_X(x'), \quad \forall\, x \in X \text{ and } f \in \mathcal{L}^2_{\rho_X}.$$

Since K is a Mercer kernel, L_K is compact and positive. Therefore, the fractional power operator L_K^β is well-defined for any $\beta > 0$. We indicate its range space by

$$L_K^\beta(\mathcal{L}^2_{\rho_X}) := \Big\{ f = \sum_{j=1}^\infty \lambda_j^\beta a_j \phi_j : \|L_K^{-\beta} f\|_\rho := \sum_{j=1}^\infty a_j^2 < \infty \Big\}, \tag{2.7}$$

where $\{\lambda_j : j \in \mathbb{N}\}$ are the positive eigenvalues of the operator L_K and $\{\phi_j : j \in \mathbb{N}\}$ are the corresponding orthonormal eigenfunctions. Thus, the smaller β is, the bigger the range space $L_K^\beta(\mathcal{L}^2_{\rho_X})$ will be. In particular, we know [25] that $L_K^\beta(\mathcal{L}^2_{\rho_X}) \subseteq \mathcal{H}_K$ for $\beta > \frac{1}{2}$ and $L_K^{\frac{1}{2}}(\mathcal{L}^2_{\rho_X}) = \mathcal{H}_K$ with the norm satisfying

$$\|g\|_K = \|L_K^{-\frac{1}{2}} g\|_\rho, \quad \forall g \in \mathcal{H}_K. \tag{2.8}$$

One can find more details in [12, 25].

Of course, one can assume that $f_\rho \in (\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$. This is a natural assumption since $L_K^\beta(\mathcal{L}^2_{\rho_X}) \subseteq (\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{2\beta,\infty}$ for $\beta \in [0, \frac{1}{2}]$, see the Appendix. However, in this paper we concentrate on the hypothesis space $L_K^\beta(\mathcal{L}^2_{\rho_X})$ since this facilitates a comparison of our results with those in the related literature (e.g. [7, 25, 27, 32, 33, 34]).

We are now in a position to state our error rates for algorithm (1.5). Hereafter, the expression $a_T = O(b_T)$ means that there exists an absolute constant c such that $a_T \le c b_T$ for all $T \in \mathbb{N}$.

Theorem 3. *Let $\theta \in (0,1)$ and $\mu(\theta)$ be given by (2.3). Define $\{f_t : t \in \mathbb{N}\}$ by (1.5). If $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with some $0 < \beta \le \frac{1}{2}$ then, by selecting $\eta_t = \frac{1}{\mu(\frac{2\beta}{2\beta+1})} t^{-\frac{2\beta}{2\beta+1}}$ for $t \in \mathbb{N}$, for any $T \in \mathbb{N}$ there holds*

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] = O\big(T^{-\frac{2\beta}{2\beta+1}} \ln T\big). \tag{2.9}$$

In Section 6.1, we will derive the error rate (2.9) from some modified error bounds similar to (2.4). Since the best rate of the second term on the right-hand side of (2.4) is $O(T^{-\frac{1}{2}} \ln T)$, the associated error rate

(2.9) is never faster than the rate $O(T^{-\frac{1}{2}}\ln T)$ achieved for $\beta = \frac{1}{2}$. In contrast, it is known that regularization algorithms [4, 7, 25, 31] in *batch learning*[1] can achieve better rates than $O(T^{-\frac{1}{2}})$ if $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > \frac{1}{2}$. Unfortunately, for the online algorithm (1.5) with polynomially decaying step sizes, we do not know how to get better rates beyond the range $\beta \in (0, \frac{1}{2}]$.

We turn our attention to the case that the step sizes of (1.5) are in the form of $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ while the learning rate η depends on the sample number T. In this scenario, the error bounds for algorithm (1.5) read as follows.

Theorem 4. *Let $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be produced by algorithm (1.5). If $16(\kappa+1)^4 \ln\big(8T\big)\eta \le 1$, then $\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big]$ is bounded by*

$$\Big[\mathcal{K}(2(1+\kappa)\big(\eta T\big)^{-\frac{1}{2}}, f_\rho)\Big]^2 + c_\rho \eta \ln\big(8T\big). \tag{2.10}$$

Note that the function $\mathcal{K}(\cdot, f_\rho)$ is non-decreasing and $\lim_{s\to 0+}\mathcal{K}(s, f_\rho) = \inf_{f\in\mathcal{H}_K}\|f - f_\rho\|_\rho$. Thus, from the error bound (2.10) we see that large number of iteration T reduces the $\mathcal{K}$-functional, but enlarge the last term in (2.10); on the other hand, small number of iteration reduces the last term in (2.10), but enlarges the $\mathcal{K}$-functional. Hence, trading off the two terms in the error bound (2.10) suggests an early stopping rule $T = T(\eta)$ which guarantees convergence of algorithm (1.5) as $\eta \to 0+$.

However, our purpose in this paper is to derive error rates with respect to the iteration steps (sample number) T by choosing η appropriately, and thus we take the inverse function's description $\eta = \eta(T)$.

In addition to the error $\|f_{T+1} - f_\rho\|_\rho$, there are other interesting relevant quantities such as the error $\|f_{T+1} - f_\rho\|_K$ (if $f_\rho \in \mathcal{H}_K$). Observe that the global error $\|f_{T+1} - f_\rho\|_\rho$ can not wholly describe the local properties of f_{T+1} which is usually measured by $|f_{T+1}(x_0) - f_\rho(x_0)|$ for any $x_0 \in X$. However, if $f_\rho \in \mathcal{H}_K$, for every $x_0 \in X$ we have that $|f_{T+1}(x_0) - f_\rho(x_0)| \le \kappa\|f_{T+1} - f_\rho\|_K$. In this case, the convergence and the error rate in $\mathcal{H}_K$ reflect the local performance of the predictor f_{T+1}. Indeed, it was further pointed out in [25] that the convergence in $\mathcal{H}_K$ yields convergence in $\mathcal{C}^k(X)$ under some conditions on K, where $\mathcal{C}^k$ denotes the space of all functions whose derivatives up to order k are continuous.

[1] In this learning setting, we usually use the whole batch of examples at one time.

When the step sizes of (1.5) are of the form $\{\eta_t = \eta(T) : t \in \mathbb{N}_T\}$, we can get convergence results in $\mathcal{L}^2_{\rho_X}$ as well as in $\mathcal{H}_K$.

Theorem 5. *Let $\{\eta_t = \eta(T) : t \in \mathbb{N}_T\}$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be produced by (1.5). Then the following statements hold true.*

(1) If the step size satisfies

$$\lim_{T\to\infty} T\eta(T) = \infty, \quad \lim_{T\to\infty} \eta(T)\ln T = 0 \tag{2.11}$$

then there holds

$$\lim_{T\to\infty} \mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|^2_\rho\right] = \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|^2_\rho. \tag{2.12}$$

(2) If $f_\rho \in \mathcal{H}_K$ and the step size satisfies

$$\lim_{T\to\infty} T\eta(T) = \infty, \quad \lim_{T\to\infty} \eta^2(T)T = 0 \tag{2.13}$$

then we have that

$$\lim_{T\to\infty} \mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|^2_K\right] = 0. \tag{2.14}$$

Note that, if the step size decays in the form of $\eta(T) = O(T^{-\theta}),\ T \to \infty$ with $\theta > 0$ then the hypothesis (2.11) allows the choice $\theta \in (0, 1)$ while (2.13) requires that $\theta \in (\frac{1}{2}, 1)$.

We will prove Theorems 4 and 5 in Section 5. In Section 6, we shall establish the following error rates in $\mathcal{L}^2_{\rho_X}$ as well as in $\mathcal{H}_K$ when the step sizes are of the form $\{\eta_t = \eta(T) : t \in \mathbb{N}_T\}$.

Theorem 6. *Let $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be produced by (1.5). If $f_\rho \in L^\beta_K(\mathcal{L}^2_{\rho_X})$ for some $\beta > 0$ then, by choosing $\eta := \frac{\beta}{64(1+\kappa)^4(2\beta+1)} T^{-\frac{2\beta}{2\beta+1}}$, we have that*

$$\mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|^2_\rho\right] = O\Big(T^{-\frac{2\beta}{2\beta+1}} \ln T\Big). \tag{2.15}$$

Moreover, if $\beta > \frac{1}{2}$ then there holds

$$\mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|^2_K\right] = O\Big(T^{-\frac{2\beta-1}{2\beta+1}}\Big). \tag{2.16}$$

As the last contribution of this paper, we further improve the preceding error rate [33] for the online regularized algorithm (1.4) with $\lambda > 0$. The proof will be given in Section 6.2.

Theorem 7. *Let $\lambda > 0$, $\theta \in (0,1)$ and $\mu(\theta)$ be given by (2.3). Define $\{f_t : t \in \mathbb{N}_{T+1}\}$ by (1.4). Assume that $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with some $0 < \beta \le 1$. For any $0 < \varepsilon < \frac{2\beta}{2\beta+1}$, choose $\eta_t = \frac{1}{\mu(\frac{2\beta}{2\beta+1})+1} t^{-\frac{2\beta}{2\beta+1}}$ and $\lambda = T^{-\frac{1}{2\beta+1}+\frac{\varepsilon}{2\beta}}$. Then there holds*

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] = O\Big(T^{-\frac{2\beta}{2\beta+1}+\varepsilon}\Big). \tag{2.17}$$

The preceding error rates of the online regularized algorithm (1.4) with $\lambda > 0$ were established in [33] under the assumption that $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with some $\beta \in (0,1]$. Namely, for any arbitrarily small $\varepsilon > 0$, by choosing $\lambda := \lambda(T)$ appropriately, there holds

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho\big] = O\Big(T^{-\frac{\beta}{2(\beta+1)}+\varepsilon}\Big) \tag{2.18}$$

and, for $\frac{1}{2} < \beta \le 1$, we further have that

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_K\big] = O\big(T^{-\frac{2\beta-1}{4\beta+2}+\varepsilon}\big). \tag{2.19}$$

By Cauchy-Schwarz inequality, we see from (2.17) that, for any arbitrarily small $\varepsilon > 0$, there holds

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho\big] \le \Big(\mathbb{E}_{\mathbf{z}\in Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big]\Big)^{\frac{1}{2}} = O\Big(T^{-\frac{\beta}{2\beta+1}+\varepsilon}\Big).$$

Therefore, the rate (2.17) is much better than (2.18).

We end this section with some remarks. Firstly, the above error rates for algorithm (1.5) are *capacity independent* except the prior requirement that $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with some $\beta > 0$. It is an open problem to improve these rates when some additional information is known such as the regularity of the kernel K or some polynomial decay of the eigenvalues of L_K [7, 12, 31]. Secondly, all the rates are proved with respect to expectation norm and we do not know how to convert them into similar probabilistic bounds such as those in [25, 32]. Finally, we wish to emphasize that when the step sizes are a polynomially decaying sequence like that in Theorem 1, the learning scheme (1.5) is a truly online algorithm. Instead, when we use a constant step size $\eta_t = \eta(T)$, algorithm (1.5) is only an informal online algorithm since the knowledge of the sample size is required in advance.

3 Related work

In this section, we discuss and compare our results with related work. A main conclusion of this discussion is that the online algorithm (1.5) is competitive with the commonly used least-square learning schemes.

3.1 Cumulative loss analysis

The mistake (regret) bounds for the cumulative loss $\sum_{t=1}^{T}(y_t - f_t(x_t))^2$ for general online algorithms have been well studied in the literature. See e.g. [2, 8, 9, 10, 17, 18, 19, 29, 35] and references therein. Specifically, mistake bounds were derived for the online density estimation in [2]. Section 6 in [17] derived, for a learning algorithm different from (1.5), upper bounds on the relative expected instantaneous loss, measuring the predicting ability of the last output in the linear regression problem. The online algorithm studied in [10] in the linear regression setting is closest to our algorithm (1.5). This paper discussed how the choice of the learning rate affects the bound for the cumulative loss. Related mistake bounds in this setting can also be found in [35]. In [19], the authors proposed general online regularized algorithms with kernels and presented their cumulative loss bounds.

For a more detailed review of mistake bounds in this direction, one can refer to [29, Section 5]. There, generalization bounds for the average prediction were also derived from the cumulative loss bounds for certain algorithms in RKHS's. More precisely, for each $t \in \mathbb{N}$, let $H_t : X \to \mathbb{R}$ be the function produced by a prediction algorithm when fed with the independent data $\{(x_j, y_j) : j \in \mathbb{N}_{t-1}\}$. Consider the average prediction $\overline{H}_{T+1} = \frac{1}{T+1}\sum_{t=1}^{T+1} H_t$ and let $D \in \mathcal{H}_K$ with $D(x) \in Y := [-M, M]$ for all $x \in X$. It was shown ([29, Corollary 2]) that there exists a prediction algorithm such that, for any $\delta > 0$ and $T \in \mathbb{N}$, with probability at least $1 - \delta$, there holds

$$\mathcal{E}(\overline{H}_{T+1}) \le \mathcal{E}(D) + \frac{2M}{\sqrt{T+1}}\Big(\sqrt{\kappa^2 + M^2}\big(\|D\|_K + 1\big) + 2M\sqrt{2\ln\frac{2}{\delta}}\Big).$$

The main difference of our generalization bounds (2.4) and (2.10) from the above one is that we have no assumption on the functions in the given upper bounds.

Soon after Proposition 1 in Section 4.1, we shall discuss an appealing relationship between statistical generalization error of f_{T+1} and a weighted modification of cumulative loss $\sum_{t=1}^{T}(y_t - f_t(x_t))^2$. However, it remains to be understood whether the standard cumulative loss bounds for algorithm (1.5) can be directly applied to the study of its generalization error.

3.2 Comparison with other regularization schemes

Although deriving cumulative loss bounds of the online gradient descent algorithm (1.4) is very useful, it is also important to further understand the statistical behavior of f_{T+1}. In [22], the authors studied the performance of f_{T+1} in the $\mathcal{H}_K$ norm when f_{T+1} is given by the online regularized algorithm (1.4) with $\lambda > 0$. More general online regularized schemes (1.4) involving commonly used loss functions in classification and regression were discussed in [20, 33]. Using quite different methods from ours, [35] obtained similar generalization bounds as (2.10) for the online gradient descent algorithm associated with uniformly Lipschitz loss functions, linear kernels and constant step sizes (learning rates).

In particular, in what follows we compare our rates for algorithm (1.5) with the state-of-art ones for other least-square learning algorithms under the same hypothesis that $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > 0$.

Firstly, we begin by the comparison with the rates for the online regularized algorithms. In this discussion, we note that the rates (2.9), (2.15) and (2.16) of algorithm (1.5) are comparable to the rates (2.17), (2.19) of the online regularized algorithm (1.4).

Secondly, we present the comparison with the rates of *averaged stochastic gradient descent algorithms* introduced in [35]. This class of online algorithms and the derived error bounds are stated in the linear kernel setting. However, we can easily extend them to the general kernel setting as follows. Assume $2\kappa^2\eta_t < 1$ for $t \in \mathbb{N}_T$, and let $r_1 = 0$, $\hat{f}_1 = 0$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be defined by (1.5). For any $t \in \mathbb{N}_T$, we update $r_{t+1} = r_t + \eta_t - 2\kappa^2\eta_t^2$ and $\hat{f}_{t+1} = \frac{r_t}{r_{t+1}}\hat{f}_t + \frac{r_{t+1}-r_t}{r_{t+1}}f_t$. Then, the generalization bound (see [35, Theorem 5.2]) for the average output $\hat{f}_{T+1}$ in the kernel setting can be cast as

$$\mathbb{E}_{Z^T}\big[\mathcal{E}(\hat{f}_{T+1})\big] \le \inf_{f\in\mathcal{H}_K}\Big\{\big(1 + \frac{2\kappa^2\sigma^2_{T+1}}{r_{T+1}}\big)\mathcal{E}(f) + \frac{1}{2r_{T+1}}\|f\|_K^2\Big\}, \qquad (3.1)$$

where $\sigma_{T+1}^2 = \sum_{t\in\mathbb{N}_T} \eta_t^2$. Note that $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2$ for any $f \in \mathcal{L}_{\rho_X}^2$. In view of the *regularization error*

$$\mathcal{D}(\lambda) := \inf_{f\in\mathcal{H}_K} \big\{\|f - f_\rho\|_\rho^2 + \lambda\|f\|_K^2\big\}, \tag{3.2}$$

the bound (3.1) implies that $\mathbb{E}_{Z^T}[\|\hat{f}_{T+1} - f_\rho\|_\rho^2]$ is bounded by

$$\Big(1 + \frac{2\kappa^2\sigma_{T+1}^2}{r_{T+1}}\Big)\mathcal{D}\Big(\frac{1}{2(r_{T+1} + 2\kappa^2\sigma_{T+1}^2)}\Big) + \frac{2\kappa^2\sigma_{T+1}^2}{r_{T+1}}\mathcal{E}(f_\rho). \tag{3.3}$$

When the step sizes are of the form $\eta_t = \frac{1}{4(1+\kappa^2)}t^{-\theta}$ with $\theta \in (0,1)$, it is not hard to observe that $r_T = O(T^{1-\theta})$, $\frac{\sigma_{T+1}^2}{r_{T+1}} = O(T^{-\min\{\theta,1-\theta\}})$ for $\theta \neq \frac{1}{2}$ and $\frac{\sigma_{T+1}^2}{r_{T+1}} = O(T^{-\frac{1}{2}}\ln T)$ for $\theta = \frac{1}{2}$. Furthermore, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $\beta \in (0, \frac{1}{2}]$ then we know from [25, Lemma 3] that

$$\mathcal{D}(\lambda) \le \lambda^{2\beta}\|L_K^{-\beta} f_\rho\|_\rho^2, \tag{3.4}$$

which implies that

$$\mathcal{D}\Big(\frac{1}{2(r_{T+1} + 2\kappa^2\sigma_{T+1}^2)}\Big) = O(T^{-2\beta(1-\theta)}). \tag{3.5}$$

Therefore, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (0, \frac{1}{2}]$, choosing $\theta = \frac{2\beta}{2\beta+1}$ and putting the above observations for r_{T+1}, $\frac{\sigma_{T+1}^2}{r_{T+1}}$ and (3.5) into (3.3) yields that

$$\mathbb{E}_{Z^T}[\|\hat{f}_{T+1} - f_\rho\|_\rho^2] = \begin{cases} O(T^{-\frac{1}{2}}\ln T) & \text{for } \beta = \frac{1}{2}, \\ O(T^{-\frac{2\beta}{2\beta+1}}) & \text{for } \beta \in (0, \frac{1}{2}). \end{cases} \tag{3.6}$$

Similarly, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (0, \frac{1}{2}]$ and the step sizes have the form $\{\eta_t \equiv \eta : t \in \mathbb{N}_T\}$ then, by choosing $\eta = \frac{1}{4(1+\kappa^2)}T^{-\frac{2\beta}{2\beta+1}}$, we see from (3.3) and (3.4) that

$$\mathbb{E}_{Z^T}[\|\hat{f}_{T+1} - f_\rho\|_\rho^2] = O(T^{-\frac{2\beta}{2\beta+1}}). \tag{3.7}$$

Based on the above discussion, we see that the rates (2.9) and (2.15) for algorithm (1.5) are the same, up to a logarithmic term, as the corresponding rates (3.6) and (3.7) for the averaged stochastic gradient descent algorithm.

Thirdly, we move on to the comparison with the Tikhonov regularization algorithm (see e.g. [16]) in batch learning which is defined by

$$f_{\mathbf{z},\lambda} := \arg\inf_{f\in\mathcal{H}_K}\Big\{\frac{1}{T}\sum_{t\in\mathbb{N}_T}(f(x_t)-y_t)^2+\lambda\|f\|_K^2\Big\}. \tag{3.8}$$

The capacity independent generalization bounds for this algorithm in [34] can be expressed as

$$\mathbb{E}_{Z^T}\big[\mathcal{E}(f_{\mathbf{z},\lambda})\big] \le \left(1+\frac{2\kappa^2}{T\lambda}\right)^2 \inf_{f\in\mathcal{H}_K}\Big\{\mathcal{E}(f)+\lambda\|f\|_K^2\Big\}.$$

In terms of the regularization error $\mathcal{D}(\lambda)$, it can be equivalently stated as

$$\mathbb{E}_{Z^T}\big[\big\|f_{\mathbf{z},\lambda}-f_\rho\big\|_\rho^2\big] \le \mathcal{D}\big(\lambda\big)+\big(\mathcal{E}(f_\rho)+\mathcal{D}\big(\lambda\big)\big)\Big\{\frac{4\kappa^2}{T\lambda}+\left(\frac{2\kappa^2}{T\lambda}\right)^2\Big\}. \tag{3.9}$$

But, if $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta\in(0,\frac{1}{2}]$ we know from (3.4) that $\mathcal{D}(\lambda)\le \lambda^{2\beta}\|L_K^{-\beta}f_\rho\|_\rho^2$. Putting this into (3.9) and trading off T and λ, for $f_\rho\in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with some $\beta\in(0,\frac{1}{2}]$, the choice $\lambda = T^{-\frac{1}{2\beta+1}}$ gives the rate

$$\mathbb{E}_{Z^T}\big[\big\|f_{\mathbf{z},\lambda}-f_\rho\big\|_\rho^2\big] = O\Big(T^{-\frac{2\beta}{2\beta+1}}\Big). \tag{3.10}$$

When $f_\rho\in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta\in(\frac{1}{2},1]$, it was further improved in [25], by choosing $\lambda=\lambda(T)$ appropriately, that[2]

$$\mathbb{E}_{Z^T}\big[\|f_{\mathbf{z},\lambda}-f_\rho\|_\rho^2\big] = O\big(T^{-\frac{2\beta}{2\beta+1}}\big), \tag{3.11}$$

and

$$\mathbb{E}_{Z^T}\big[\|f_{\mathbf{z},\lambda}-f_\rho\|_K^2\big] = O\big(T^{-\frac{2\beta-1}{2\beta+1}}\big). \tag{3.12}$$

Under the same assumptions, the rates (2.9), (2.15) and (2.16) of the online algorithm (1.5) are almost the same as the corresponding rates (3.10), (3.11) and (3.12) in batch learning.

Finally, we end this subsection with the remark that generalization bounds associated with $\mathcal{D}(\lambda)$ such as (3.9) for the regularization algorithms are quite neat and interesting, but they tend to suffer a saturation phenomenon.

[2]The rates are originally stated in the form of probability inequalities. We employ their expectation versions for consistency with the other bounds presented in the paper.

To explain this phenomenon, we observe that

$$\mathcal{D}(\lambda) = \inf_{f\in\mathcal{H}_K}\{\|f-f_\rho\|_\rho^2+\lambda\|f\|_K^2\} \geq \inf_{f\in\mathcal{H}_K}\Big\{\|f-f_\rho\|_\rho^2+\lambda\frac{\|f\|_\rho^2}{\kappa^2}\Big\}.$$

Since $\|f-f_\rho\|_\rho \geq |\|f\|-\|f_\rho\|_\rho|$, by letting $t=\|f\|_\rho$, we have that

$$\mathcal{D}(\lambda) \geq \inf_{t\in\mathbb{R}}\Big\{(t-\|f_\rho\|_\rho)^2+\frac{\lambda t^2}{\kappa^2}\Big\} = \frac{\lambda}{\kappa^2+\lambda}\|f_\rho\|_\rho^2. \tag{3.13}$$

If f_ρ is not identically zero (i.e., $\|f_\rho\|_\rho > 0$), the inequality (3.13) implies that the optimal decay of $\mathcal{D}(\lambda)$ is $O(\lambda)$, $\lambda\to 0+$ which, by (3.4), is achieved when $f_\rho \in L_K^{\frac{1}{2}}(\mathcal{L}^2_{\rho_X}) = \mathcal{H}_K$. Equivalently, the rate of $\mathcal{D}(\lambda)$ is never faster than $O(\lambda)$, $\lambda\to 0+$ even if f_ρ lies in any smaller space $L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > \frac{1}{2}$ than $\mathcal{H}_K$. Therefore, in this case the consequent trade-off rate (3.10) derived from (3.9) is at most $O(T^{-\frac{1}{2}})$, $T\to\infty$, which is achieved at $\beta=\frac{1}{2}$ and can not be improved beyond the range $\beta\in(0,\frac{1}{2}]$. This is the so-called saturation phenomenon in the context of inverse problems [15].

For similar reasons, the rates (3.6) and (3.7) derived from (3.3) for the averaged stochastic gradient descent algorithm have the same problem. In contrast, our capacity independent rate (2.15) for algorithm (1.5) does not suffer this drawback and is arbitrarily close to T^{-1} as $\beta\to\infty$.

3.3 Capacity independent optimality

In this subsection, we explain in two folds that the error rates for algorithm (1.5) are optimal, up to a logarithmic term, in the capacity independent sense. We initially illustrate this by citing the lower bounds in [7] when $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta\in(\frac{1}{2},1]$. Then, we demonstrate a specific example to contrast our rates with the best possible rate in the non-parametric statistical literature (see e.g. [6, 27]).

We begin with the citation of the lower bound in [7]. Consider $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta\in(\frac{1}{2},1]$ and the eigenvalues (arranged in decreasing order) of L_K have the decay $\lambda_i = O(i^{-b})$ with $b>1$. It was proved in [7] that the rate $T^{-\frac{2\beta b}{2\beta b+1}}$ in $\mathcal{L}^2_{\rho_X}$ is optimal, see [7] for a precise definition of optimal rates. Since $K(x,x') = \sum_{i\in\mathbb{N}}\lambda_i\phi_i(x)\phi_i(x')$ for any $x,x'\in X$ and $\int_X\phi_i^2(x)d\rho_X(x)=1$ for any $i\in\mathbb{N}$, it follows that $\sum_{i=1}^\infty\lambda_i = \int_X K(x,x)d\rho_X(x)\leq\kappa^2$, see e.g. [12]. Therefore, for any $i\in\mathbb{N}$ we have

that $i\lambda_i \leq \sum_{j\in\mathbb{N}_i}\lambda_j \leq \kappa^2$ which implies that $\lambda_i = O(i^{-1})$ for all kernels. Since our capacity independent rates are independent of the eigenvalues of L_K, taking $b \to 1$ in the rate $T^{-\frac{2\beta b}{2\beta b+1}}$ leads to the eigenvalue-independent optimal rate $T^{-\frac{2\beta}{2\beta+1}}$ (see also [32] for a discussion). In this sense, our rates (2.15) and (2.17) are almost optimal for the capacity independent case under the condition that $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta \in (\frac{1}{2}, 1]$.

Now, we illustrate by a specific example that our rates for algorithm (1.5) are comparable to the optimal rate in non-parametric statistics. For simplicity, we only consider smooth splines [30] in one dimension.

Let $X = [0,1]$, $d\rho_X = dx$. For smooth splines, the reproducing kernel space can be regarded as a fractional Sobolev space $H^s[0,1]$ with $s > \frac{1}{2}$ which is defined by Fourier coefficients

$$H^s[0,1] := \Big\{ f = a_0 + \sum_{k\in\mathbb{N}} \sqrt{2}k^{-s}(a_k \sin(2\pi kx) + b_k \cos(2\pi kx)) : \\ \|f\|_s^2 := a_0^2 + \sum_{k\in\mathbb{N}}(a_k^2 + b_k^2) < \infty \Big\}.$$

It is easy to see, for any $x, x' \in X$, that the reproducing kernel of the space $H^s[0,1]$ can be represented by

$$K_s(x,x') = 1 + \sum_{k\in\mathbb{N}} 2k^{-2s}\cos(2\pi k(x-x')).$$

In addition, $\lambda_1 = 1$, $\lambda_{2j} = \lambda_{2j+1} = j^{-2s}$ for $j \in \mathbb{N}$ and $\phi_1(x) = 1$, $\phi_{2j}(x) = \sqrt{2}\sin(2\pi jx)$ and $\phi_{2j+1}(x) = \sqrt{2}\cos(2\pi jx)$ for $j \in \mathbb{N}$. Therefore, in this case, the range space $L^\beta_{K_s}(\mathcal{L}^2_{dx})$ with $\beta > 0$ defined by (2.7) is identical to the Sobolev space $H^{2\beta s}[0,1]$.

It is well known (e.g. [27]) in non-parametric statistics that the rate $O(T^{-\frac{4\beta s}{4\beta s+1}})$ is optimal if $f_\rho \in L^\beta_{K_s}(\mathcal{L}^2_{dx}) = H^{2\beta s}[0,1]$ with $\beta > \frac{1}{2}$. Therefore, if $d\rho_X = dx$, $f_\rho \in L^\beta_{K_s}(\mathcal{L}^2_{dx})$, and $\mathcal{H}_{K_s} = H^s[0,1]$ with $s > \frac{1}{2}$ with $\beta > \frac{1}{2}$ then our rate $O(T^{-\frac{2\beta}{2\beta+1}}\ln T)$ of algorithm (1.5) given by (2.15) is suboptimal. But it is arbitrarily close to the best one $O(T^{-\frac{4\beta s}{4\beta s+1}})$ as $s \to \frac{1}{2}+$.

4 Error decomposition and basic estimates

In this section, we formulate our basic ideas by providing a novel approach to the error analysis of online gradient descent algorithms in $\mathcal{L}^2_{\rho_X}$. As

mentioned in the introduction, the main purpose of this paper is to analyze the online algorithm (1.5). However, we would like to state a unified approach for the general online algorithm (1.4) for any $\lambda \geq 0$ since this will, as a byproduct, allow us to improve the preceding error rate in [33] for the algorithm (1.4) with $\lambda > 0$ as proved in Section 6.2 below.

We first establish some useful observations used later. For $\lambda > 0$, define the *regularizing function* by

$$f_\lambda = \arg \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) + \lambda \|f\|_K^2 \right\}. \tag{4.1}$$

Lemma 1. *Let $\lambda > 0$ and f_λ be defined as above. Then we have that*

$$L_K(f_\lambda - f_\rho) + \lambda f_\lambda = 0 \tag{4.2}$$

and

$$\|f_\lambda - f_\rho\|_\rho \leq \|f_\rho\|_\rho. \tag{4.3}$$

Proof. The first equality is well-known, see e.g. [12].

For the inequality (4.3), we know from the equality (1.6) that (4.1) is equivalent to $f_\lambda = \arg\inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\}$. By taking $f = 0$, the definition of f_λ yields the desired estimate $\|f_\lambda - f_\rho\|_\rho^2 \leq \|f_\rho\|_\rho^2$. □

With (4.2) at hand, we can interpret the online algorithm (1.4) with $\lambda \geq 0$ as the following useful form. With a slight abuse of notation, we indicate f_ρ by f_0.

Lemma 2. *Let $\lambda \geq 0$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be defined as (1.4). Then, for any $t \in \mathbb{N}_T$ we have that*

$$f_{t+1} - f_\lambda = (I - \eta_t(L_K + \lambda I))(f_t - f_\lambda) + \eta_t \mathcal{B}(f_t, z_t), \tag{4.4}$$

where I is the identity operator and the vector-valued random variable $\mathcal{B}(f_t, z_t)$ is defined by

$$\mathcal{B}(f_t, z_t) := L_K(f_t - f_\rho) + (y_t - f_t(x_t))K_{x_t}. \tag{4.5}$$

Proof. The equality (4.4) for the case $\lambda = 0$ is easily verified since we have set $f_0 = f_\rho$.

For the case $\lambda > 0$, we use the property (4.2) of the regularizing function f_λ in Lemma 1. By the definition of f_{t+1} given by (1.4) with $\lambda > 0$, we know that

$$\begin{aligned} f_{t+1} - f_\lambda &= f_t - f_\lambda - \eta_t(f_t(x_t) - y_t)K_{x_t} - \eta_t\lambda f_t \\ &= (I - \eta_t(L_K + \lambda I))(f_t - f_\lambda) + \eta_t L_K(f_t - f_\lambda) \\ &\quad - \eta_t\lambda f_\lambda + \eta_t(y_t - f_t(x_t))K_{x_t}. \end{aligned}$$

But the equality (4.2) tells us that $-\lambda f_\lambda = L_K(f_\lambda - f_\rho)$. Hence, putting this back into the above equation and arranging it yield the desired equality (4.4). □

With the help of the formula (4.4), we can describe our approach by three steps in which the first one is referred to as *error decomposition*.

4.1 Error decomposition

For $\lambda \geq 0$ and $t \in \mathbb{N}$, set the operator $\omega_k^t(L_K + \lambda I) := \prod_{j=k}^t (I - \eta_j(L_K + \lambda I))$ for $k \in \mathbb{N}_t$ and $\omega_{t+1}^t(L_K + \lambda I) := I$. Applying induction to the equality (4.4) and noting that $f_1 = 0$, for any $t \in \mathbb{N}_T$ we have that

$$f_{t+1} - f_\lambda = -\omega_1^t(L_K + \lambda I)f_\lambda + \sum_{j\in\mathbb{N}_t} \eta_j\omega_{j+1}^t(L_K + \lambda I)\mathcal{B}(f_j, z_j). \quad (4.6)$$

Applying (4.6) with $t = T$, we get the following error decomposition

$$f_{T+1} - f_\lambda = -\omega_1^T(L_K + \lambda I)f_\lambda + \sum_{t\in\mathbb{N}_T} \eta_t\omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t). \quad (4.7)$$

The above error decomposition technique is well-known in statistical learning theory in order to realize the error analysis for least-square related learning algorithms, see e.g. [22, 32, 33]. One can find similar ideas used for different learning schemes, see e.g. [3, 12, 14, 24, 25] and references therein.

We will use the error decomposition (4.7) to estimate the expectation of $\|f_{T+1} - f_\rho\|_\rho^2$. To this end, we need some useful preparations. Let L be a linear operator from $\mathcal{L}_{\rho_X}^2$ to itself, we use $\|L\|$ to denote its operator norm, that is, $\|L\| := \sup_{\|f\|_\rho=1} \|Lf\|_\rho$. We also need the following technical lemma which forms the essential estimates of our approach.

Lemma 3. *Let $\lambda \geq 0, \beta > 0$ and $\eta_t(\kappa^2 + \lambda) \leq 1$ for any integer $t \in [j, k]$. Then there holds*

$$\|\omega_j^k (L_K + \lambda I) L_K^\beta\|^2 \leq \frac{2\Big(\big(\frac{\beta}{e}\big)^\beta + \kappa^{2\beta}\Big)^2}{\exp\big\{\lambda \sum_{t=j}^k \eta_t\big\}\big(1 + \big(\sum_{t=j}^k \eta_t\big)^{2\beta}\big)}.$$

Proof. Note that, for any $x, x' \in X$

$$K(x, x') = \langle K_x, K_{x'} \rangle_K \leq \|K_x\|_K \|K_{x'}\|_K \leq \kappa^2. \tag{4.8}$$

Combing this with the fact $\rho_X(X) = 1$, we see from Cauchy-Schwarz inequality that, for any $f \in \mathcal{L}^2_{\rho_X}$

$$\begin{aligned}
\|L_K f\|_\rho^2 &= \int_X \Big| \int_X K(x, x') f(x') d\rho_X(x') \Big|^2 d\rho_X(x) \\
&\leq \int_X \Big(\int_X \big|K(x, x')\big|^2 d\rho_X(x') \int_X \big|f(x')\big|^2 d\rho_X(x') \Big) d\rho_X(x) \\
&= \|f\|_\rho^2 \int_X \int_X \big|K(x, x')\big|^2 d\rho_X(x') d\rho_X(x) \leq \kappa^4 \|f\|_\rho^2.
\end{aligned} \tag{4.9}$$

Hence, $\|L_K\| \leq \kappa^2$. Since $\{\lambda_\ell : \ell \in \mathbb{N}\}$ are the eigenvalues of L_K, it follows that $\sup_\ell \lambda_\ell \leq \kappa^2$ and

$$\|L_K^\beta\| := \sup_{\ell \in \mathbb{N}} \lambda_\ell^\beta \leq \kappa^{2\beta}. \tag{4.10}$$

Moreover, for any $x \leq (\kappa^2 + \lambda)^{-1}$,

$$\|I - x(L_K + \lambda I)\| := \sup_{\ell \in \mathbb{N}} (1 - x(\lambda_\ell + \lambda)) \leq 1 - x\lambda. \tag{4.11}$$

Applying (4.11) with $x = \eta_t$ iteratively for $t \in [j, k]$ and (4.10), we have that

$$\begin{aligned}
\|\omega_j^k (L_K + \lambda I) L_K^\beta\| &\leq \prod_{t=j}^k \|I - \eta_t (L_K + \lambda I)\| \|L_K^\beta\| \\
&\leq \kappa^{2\beta} \prod_{t=j}^k (1 - \eta_t \lambda) \leq \exp\Big\{ -\lambda \sum_{t=j}^k \eta_t \Big\} \kappa^{2\beta},
\end{aligned} \tag{4.12}$$

where the elementary inequality $1 - \eta_t \lambda \leq e^{-\eta_t \lambda}$ for any $t \in [j, k]$ is used in the last inequality.

Also,

$$\begin{aligned}\|\omega_j^k(L_K+\lambda I)L_K^\beta\| &:= \sup_{\ell\in\mathbb{N}}\prod_{t=j}^{k}(1-\eta_t(\lambda_\ell+\lambda))\lambda_\ell^\beta \\ &\le \sup_{\ell\in\mathbb{N}}\exp\big\{-(\lambda+\lambda_\ell)\sum_{t=j}^{k}\eta_t\big\}\lambda_\ell^\beta \\ &\le \exp\Big\{-\lambda\sum_{t=j}^{k}\eta_t\Big\}\sup_{x\ge 0}\exp\Big\{-x\sum_{t=j}^{k}\eta_t\Big\}x^\beta \\ &= \exp\Big\{-\lambda\sum_{t=j}^{k}\eta_t\Big\}\Big(\frac{\beta}{e}\Big)^\beta\Big(\sum_{t=j}^{k}\eta_t\Big)^{-\beta}.\end{aligned}$$

Consequently, the above inequality and (4.12) implies that $\|\omega_t^T(L_K+\lambda I)L_K^\beta\|^2$ is bounded by

$$\exp\Big\{-\lambda\sum_{t=j}^{k}\eta_t\Big\}\bigg(\Big(\frac{\beta}{e}\Big)^\beta+\kappa^{2\beta}\bigg)^2\min\Big\{1,\Big(\sum_{t=j}^{k}\eta_t\Big)^{-2\beta}\Big\}.$$

Combining this with the elementary inequality that $\min\{a^{-1},b^{-1}\}\le\frac{2}{a+b}$ for any $a>0,b>0$ finishes the lemma. □

Now we present the upper bound for $\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\rho\|_\rho^2\big]$ which is the foundation of our approach introduced in this paper. Hereafter, we adopt the convention that $\sum_{j=\ell+1}^{\ell}\eta_j=0$ for any $\ell\in\mathbb{N}$.

Proposition 1. *Let $\lambda\ge 0$ and $\{f_t: t\in\mathbb{N}\}$ be defined as (1.4). Then, for any $T\in\mathbb{N}$, $\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\rho\|_\rho^2\big]$ is bounded by*

$$\begin{aligned}&\|f_\lambda-f_\rho\|_\rho^2+\|\omega_1^T(L_K+\lambda I)f_\lambda\|_\rho^2+2\|\omega_1^T(L_K+\lambda I)f_\lambda\|_\rho\|f_\lambda-f_\rho\|_\rho \\ &\quad+2(1+\kappa)^4\sum_{t\in\mathbb{N}_T}\frac{\eta_t^2\exp\big\{-\lambda\sum_{j=t+1}^{T}\eta_j\big\}}{1+\sum_{j=t+1}^{T}\eta_j}\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big].\end{aligned}\tag{4.13}$$

Proof. Since $f_{T+1}-f_\rho=f_{T+1}-f_\lambda+f_\lambda-f_\rho$, there holds

$$\begin{aligned}\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\rho\|_\rho^2\big]&=\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\lambda\|_\rho^2\big]+\|f_\lambda-f_\rho\|_\rho^2 \\ &\quad+2\mathbb{E}_{Z^T}\big[\langle f_{T+1}-f_\lambda,f_\lambda-f_\rho\rangle_\rho\big].\end{aligned}\tag{4.14}$$

In order to use (4.14) to prove (4.13), we need to estimate the first term and the last term on the right-hand side of (4.14).

For the last term on the right-hand side of (4.14), observe that the data is i.i.d. and f_t is only dependent on $\{z_1, \ldots, z_{t-1}\}$, not on z_t. Moreover, recall the definition of the regression function: $f_\rho(x) = \int_X y d\rho(y|x)$. Thus, $\mathcal{B}(f_t, z_t)$ has a nice vanishing property

$$\mathbb{E}_{z_t}\big[\mathcal{B}(f_t, z_t)\big] = 0, \quad \forall t \in \mathbb{N}. \tag{4.15}$$

Applying (4.15) to the equality (4.7) implies that $\mathbb{E}_{Z^T}\big[f_{T+1} - f_\lambda\big] = -\omega_1^T(L_K + \lambda I)f_\lambda$. Therefore,

$$\begin{aligned}\mathbb{E}_{Z^T}\langle f_{T+1} - f_\lambda, f_\lambda - f_\rho\rangle_\rho &= \langle \mathbb{E}_{Z^T}\big[f_{T+1} - f_\lambda\big], f_\lambda - f_\rho\rangle_\rho \\ &\le \|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho \|f_\lambda - f_\rho\|_\rho. \end{aligned} \tag{4.16}$$

For the first term on the right-hand side of (4.14), we first use the equality (4.7) to get that

$$\begin{aligned}&\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\lambda\|_\rho^2\big] = \|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho^2 \\ &+ \mathbb{E}_{Z^T}\big[\|\sum_{t\in\mathbb{N}_T} \eta_t \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t)\|_\rho^2\big] \\ &- 2\mathbb{E}_{Z^T}\big[\langle \sum_{t\in\mathbb{N}_T} \eta_t \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t), \omega_1^T(L_K + \lambda I)f_\lambda\rangle_\rho. \end{aligned} \tag{4.17}$$

Below, we estimate the last two terms on the right-hand side of (4.17) separately.

For the second term on the right-hand side of (4.17), we write it as

$$\sum_{t\in\mathbb{N}_T}\sum_{t'\in\mathbb{N}_T} \eta_t\eta_{t'}\mathbb{E}_{Z^T}\Big[\langle \omega_{t'+1}^T(L_K + \lambda I)\mathcal{B}(f_{t'}, z_{t'}), \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t)\rangle_\rho\Big].$$

Therefore, by (4.5), for $t > t'$

$$\begin{aligned}&\mathbb{E}_{Z^t}\langle \omega_{t'+1}^T(L_K + \lambda I)\mathcal{B}(f_{t'}, z_{t'}), \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t)\rangle_\rho \\ &= \mathbb{E}_{Z^{t-1}}\mathbb{E}_{z_t}\langle \omega_{t+1}^T(L_K + \lambda I)\omega_{t'+1}^T(L_K + \lambda I)\mathcal{B}(f_{t'}, z_{t'}), \mathcal{B}(f_t, z_t)\rangle_\rho \\ &= \mathbb{E}_{Z^{t-1}}\langle \omega_{t+1}^T(L_K + \lambda I)\omega_{t'+1}^T(L_K + \lambda I)\mathcal{B}(f_{t'}, z_{t'}), \mathbb{E}_{z_t}\big[\mathcal{B}(f_t, z_t)\big]\rangle_\rho = 0.\end{aligned}$$

By the symmetry of t, t', the above equality also holds true for $t' > t$.

Consequently, it follows that

$$\begin{aligned}
&\mathbb{E}_{Z^T}\big[\|\sum_{t\in\mathbb{N}_T}\eta_t\omega_{t+1}^T(L_K+\lambda I)\mathcal{B}(f_t,z_t)\|_\rho^2\big]\\
&\quad=\sum_{t\in\mathbb{N}_T}\eta_t^2\mathbb{E}_{Z^t}\big[\|\omega_{t+1}^T(L_K+\lambda I)\mathcal{B}(f_t,z_t)\|_\rho^2\big]\\
&\quad\le\sum_{t\in\mathbb{N}_T}\eta_t^2\|\omega_{t+1}^T(L_K+\lambda I)L_K^{\frac{1}{2}}\|^2\mathbb{E}_{Z^t}\big[\|L_K^{-\frac{1}{2}}\mathcal{B}(f_t,z_t)\|_\rho^2\big]\\
&\quad=\sum_{t\in\mathbb{N}_T}\eta_t^2\|\omega_{t+1}^T(L_K+\lambda I)L_K^{\frac{1}{2}}\|^2\mathbb{E}_{Z^t}\big[\|\mathcal{B}(f_t,z_t)\|_K^2\big],
\end{aligned}\tag{4.18}$$

where we have used equation (2.8) in the last equality.

To estimate $\mathbb{E}_{Z^t}\big[\|\mathcal{B}(f_t,z_t)\|_K^2\big]$, we note from (4.15) that $\mathbb{E}_{z_t}\big[(f_t(x_t)-y_t)K_{x_t}\big]=L_K(f_t-f_\rho)$, and thus we can rewrite $\mathbb{E}_{z_t}\big[\big\|\mathcal{B}(f_t,z_t)\big\|_K^2\big]$ as

$$\mathbb{E}_{z_t}\big[\big\|(f_t(x_t)-y_t)K_{x_t}\big\|_K^2\big]-\big\|\mathbb{E}_{z_t}\big[(f_t(x_t)-y_t)K_{x_t}\big]\big\|_K^2.\tag{4.19}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{Z^t}\big[\big\|\mathcal{B}(f_t,z_t)\big\|_K^2\big]\le\mathbb{E}_{Z^{t-1}}\mathbb{E}_{z_t}\big[\big\|(f_t(x_t)-y_t)K_{x_t}\big\|_K^2\big]\\
&\le\kappa^2\mathbb{E}_{Z^{t-1}}\big[\mathbb{E}_{z_t}\big[\big(f_t(x_t)-y_t\big)^2\big]\big]=\kappa^2\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big].
\end{aligned}\tag{4.20}$$

In addition, applying Lemma 3 with $\beta=\frac{1}{2}$ and $j=t+1,k=T$ implies that $\|\omega_{t+1}^T(L_K+\lambda I)L_K^{\frac{1}{2}}\|^2\le\frac{2(1+\kappa)^2\exp\{-\lambda\sum_{j=t+1}^T\eta_j\}}{1+\sum_{j=t+1}^T\eta_j}$. Consequently, plugging this and (4.20) into (4.18) yields that

$$\begin{aligned}
&\mathbb{E}_{Z^T}\big[\|\sum_{t\in\mathbb{N}_T}\eta_t\omega_{t+1}^T(L_K+\lambda I)\mathcal{B}(f_t,z_t)\|_\rho^2\big]\\
&\le 2(1+\kappa)^4\sum_{t\in\mathbb{N}_T}\frac{\exp\big\{-\lambda\sum_{j=t+1}^T\eta_j\big\}}{1+\sum_{j=t+1}^T\eta_j}\,\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big].
\end{aligned}\tag{4.21}$$

For the last term on the right-hand side of (4.17), we use (4.15) to get that

$$\begin{aligned}
&\mathbb{E}_{Z^T}\big[\langle\sum_{t\in\mathbb{N}_T}\eta_t\omega_{t+1}^T(L_K+\lambda I)\mathcal{B}(f_t,z_t),\omega_1^T(L_K+\lambda I)f_\lambda\rangle_\rho\\
&=\sum_{t\in\mathbb{N}_T}\langle\eta_t\omega_{t+1}^T(L_K+\lambda I)\mathbb{E}_{Z^{t-1}}\mathbb{E}_{z^t}\big[\mathcal{B}(f_t,z_t)\big],\omega_1^T(L_K+\lambda I)f_\lambda\rangle_\rho=0.
\end{aligned}$$

Substituting this and (4.21) into (4.17) yields that

$$\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\lambda\|_\rho^2\big] \le \|\omega_1^T(L_K+\lambda I)f_\lambda\|_\rho^2 + 2(1+\kappa)^4\sum_{t\in\mathbb{N}_T}\frac{\exp\{-\lambda\sum_{j=t+1}^T\eta_j\}}{1+\sum_{j=t+1}^T\eta_j}\,\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]. \tag{4.22}$$

Combining this with (4.16), (4.14) completes the upper bound (4.13). □

Now the error decomposition and Proposition 1 help us reduce the goal of our error analysis to the estimation of the four terms in (4.13). The first three terms of (4.13) are frequently referred to as the *approximation error* [24, 25]. Since the data is i.i.d. and f_t does not depend on z_t, the last term of (4.13) can be rewritten as

$$\mathbb{E}_{Z^T}\Big[\sum_{t\in\mathbb{N}_T}\frac{\eta_t^2\exp\{-\lambda\sum_{j=t+1}^T\eta_j\}}{1+\sum_{j=t+1}^T\eta_j}(y_t-f_t(x_t))^2\Big].$$

It can be regarded as the expectation of a weighted form of the cumulative loss $\sum_{t\in\mathbb{N}_T}(y_t-f_t(x_t))^2$ in the online learning literature [8, 10, 19]. Thus, we call the second term of (4.13) the *cumulative sample error*. We estimate these two errors separately in the following two subsections which constitute the second and last steps of our approach.

4.2 Estimates for the approximation error

Here, we establish some basic estimates for the deterministic approximation errors involving $\|f_\lambda-f_\rho\|_\rho$ and $\|\omega_1^T(L_K+\lambda I)f_\lambda\|$ which is the second step of our approach.

To estimate the term $\|f_\lambda-f_\rho\|_\rho$, we only need to consider the case $\lambda>0$ since $f_0=f_\rho$. For this purpose, recall Lemma 3 in [25].

Lemma 4. *Let $\lambda>0$ and f_λ be defined by (4.1). If $f_\rho\in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $0<\beta\le 1$ then there holds*

(a) $\|f_\lambda-f_\rho\|_\rho\le\lambda^\beta\|L_K^{-\beta}f_\rho\|_\rho$;

(b) If moreover, $\frac{1}{2}<\beta\le 1$ *then* $\|f_\lambda-f_\rho\|_K\le\lambda^{\beta-\frac{1}{2}}\|L_K^{-\beta}f_\rho\|_\rho$.

Using the above lemma, we can estimate the quantity $\|\omega_1^T(L_K+\lambda I)f_\lambda\|_\rho$ as follows.

Lemma 5. *Let $\lambda \geq 0$ and $\eta_t(\kappa^2 + \lambda) \leq 1$ for each $t \in \mathbb{N}_T$. Then the following statements hold true.*

(a) If $\lambda = 0$ then we have that

$$\|\omega_1^T(L_K) f_\rho\|_\rho \leq \mathcal{K}\big(2(1+\kappa)\big(\sum_{t\in\mathbb{N}_T} \eta_t\big)^{-\frac{1}{2}}, f_\rho\big);$$

(b) If $\lambda > 0$ then we have that

$$\|\omega_1^T(L_K + \lambda I) f_\lambda\|_\rho \leq 2\exp\Big\{-\lambda \sum_{t\in\mathbb{N}_T} \eta_t\Big\}\|f_\rho\|_\rho.$$

Proof. We first prove property (a). Since $\eta_t\kappa^2 \leq 1$ for $t \in \mathbb{N}_T$, using (4.11) with $\lambda = 0$ and $x = \eta_t$ for $t \in \mathbb{N}_T$ iteratively implies that $\|\omega_1^T(L_K)\| \leq \prod_{t\in\mathbb{N}_T} \|I - \eta_t L_K\| \leq 1$. Thus, for any $f \in \mathcal{H}_K$ there holds

$$\begin{aligned}\|\omega_1^T(L_K) f_\rho\|_\rho &\leq \|\omega_1^T(L_K)(f - f_\rho)\|_\rho + \|\omega_1^T(L_K) f\|_\rho \\ &\leq \|f - f_\rho\|_\rho + \|\omega_1^T(L_K) L_K^{\frac{1}{2}}\|\ \|L_K^{-\frac{1}{2}} f\|_\rho. \\ &= \|f - f_\rho\|_\rho + \|\omega_1^T(L_K) L_K^{\frac{1}{2}}\| \|f\|_K, \end{aligned} \tag{4.23}$$

where (2.8) is used in the last equality. Applying Lemma 3 with $\lambda = 0, \beta = \frac{1}{2}, j = 1$, and $k = T$ implies that $\|\omega_1^T(L_K) L_K^{\frac{1}{2}}\| \leq 2(1+\kappa)\Big(\sum_{t\in\mathbb{N}_T} \eta_t\Big)^{-\frac{1}{2}}$. Then, substituting this into the right-hand side of (4.23) yields that

$$\|\omega_1^T(L_K) f_\rho\|_\rho \leq \inf_{f\in\mathcal{H}_K} \big\{\|f - f_\rho\|_\rho + 2(1+\kappa)\Big(\sum_{t\in\mathbb{N}_T} \eta_t\Big)^{-\frac{1}{2}} \|f\|_K\big\}. \tag{4.24}$$

Turn to the proof of property (b), note that $\eta_t(\kappa^2 + \lambda) \leq 1$ for any $t \in \mathbb{N}_T$. Thus, applying (4.11) again with $x = \eta_t$ for $t \in \mathbb{N}_T$ implies that

$$\|\omega_1^T(L_K + \lambda I) f_\lambda\|_\rho \leq \prod_{t\in\mathbb{N}_T} (1 - \eta_t\lambda)\|f_\lambda\|_\rho \leq \exp\big\{-\lambda \sum_{t\in\mathbb{N}_T} \eta_t\big\}\|f_\lambda\|_\rho.$$

But, by (4.3), $\|f_\lambda\|_\rho \leq 2\|f_\rho\|_\rho$ which finishes property (b). □

The last step of our approach is to estimate the cumulative sample error.

4.3 Estimates for the cumulative sample error

In this subsection, we describe basic estimates for the cumulative sample error.

Observe that the cumulative sample error (the last term in (4.13)) can be bounded by

$$\Big[\sup_{t\in\mathbb{N}_T}\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]\Big]\Big[\sum_{t\in\mathbb{N}_T}\frac{\eta_t^2\exp\big\{-\lambda\sum_{j=t+1}^{T}\eta_j\big\}}{1+\sum_{j=t+1}^{T}\eta_j}\Big]. \tag{4.25}$$

Therefore, it remains to estimate the above summand involving the step sizes and the uniform bounds for the learning sequence separately. Let us first deal with the summand.

Lemma 6. *Let $\lambda\geq 0$, $\theta\in[0,1)$ and $\mu\geq\max\{\lambda,1\}+\kappa^2$. If the step sizes are in the form of $\{\eta_t=\frac{1}{\mu}t^{-\theta}:t\in\mathbb{N}_T\}$ then, for any $\ell\in\mathbb{N}_T$, the summation $\sum_{j\in\mathbb{N}_\ell}\frac{\eta_j^2\exp\{-\lambda\sum_{k=j+1}^{\ell}\eta_k\}}{1+\sum_{k=j+1}^{\ell}\eta_k}$ is bounded by*

$$\frac{c_\theta}{\mu}\Big(\exp\big\{-\lambda d_\theta\ell^{1-\theta}/\mu\big\}\ell^{-\min\{\theta,1-\theta\}}+\ell^{-\theta}\Big)\ln\big(\frac{8\ell}{1-\theta}\big), \tag{4.26}$$

where $d_\theta=\frac{1-2^{\theta-1}}{1-\theta}$ and $c_\theta=1$ for $\theta=0$, 13 for $\theta=\frac{1}{2}$, and $\frac{13}{(1-2^{\theta-1})|2\theta-1|}$ otherwise.

The technical proof is given in the Appendix.

Now, we can see that if the step sizes are of the form $\{\eta_t=\frac{1}{\mu}t^{-\theta}:t\in\mathbb{N}_T\}$ with $\theta\in(0,1)$ and $\mu\geq\max\{\lambda,1\}+\kappa^2$, the summation part involving the step sizes in (4.25) is bounded by (4.26) with $\ell=T$, which tends to zero as $T\to\infty$.

Hence, in order to estimate the whole term (4.25) it suffices to establish the uniform bounds for $\sup_{t\in\mathbb{N}_{T+1}}\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]$. This is intuitively reasonable since we expect that the learning sequence $\{f_t:t\in\mathbb{N}_{T+1}\}$ approximates the regression function f_ρ. To this end, we need the following direct result from Lemma 6.

Corollary 2. *Let $\lambda\geq 0$, $\theta\in[0,1)$ and $\eta_t=\frac{1}{\mu}t^{-\theta}$ for $t\in\mathbb{N}_T$ with $\mu\geq\kappa^2+\lambda$. For any $T\in\mathbb{N}$, if $\mu-\lambda$ is larger than the quantity*

$$\mu(\theta):=\begin{cases}16(1+\kappa)^4\ln\big(8T\big) & \text{if } \theta=0,\\ 1248(1+\kappa)^4 & \text{if } \theta=\frac{1}{2},\\ \frac{208(1+\kappa)^4}{(1-2^{\theta-1})|2\theta-1|}\Big(\ln\big(\frac{8}{1-\theta}\big)+\frac{1}{\min\{\theta,1-\theta\}}\Big) & \text{otherwise,}\end{cases} \tag{4.27}$$

then there holds

$$\sum_{j\in\mathbb{N}_t}\frac{\eta_j^2\exp\{-\lambda\sum_{k=j+1}^t\eta_k\}}{1+\sum_{k=j+1}^t\eta_k}\leq\frac{1}{8(1+\kappa)^4},\quad\forall t\in\mathbb{N}_T. \tag{4.28}$$

Proof. For any $\ell\in\mathbb{N}_T$ and $\lambda\geq 0$, the quality (4.26) is bounded by $\frac{2c_\theta}{\mu}\ell^{-\min\{\theta,1-\theta\}}\ln(\frac{8\ell}{1-\theta})$. Thus, the hypothesis (4.28) holds true as long as $\mu\geq\kappa^2+\max\{\lambda,1\}$ and $\frac{16(1+\kappa)^4c_\theta}{\mu}t^{-\min\{\theta,1-\theta\}}\ln\left(\frac{8t}{1-\theta}\right)\leq 1$ for any $t\in\mathbb{N}_T$.

When $\theta=0$, by the definition of c_0, the above inequality is virtually identical to $\mu\geq 16(1+\kappa)^4\ln\big(8T\big)$.

For $\theta\in(0,1)$, note that $\ln t=\frac{1}{\min\{\theta,1-\theta\}}\ln\big(t^{\min\{\theta,1-\theta\}}\big)\leq\frac{t^{\min\{\theta,1-\theta\}}}{\min\{\theta,1-\theta\}}$. Therefore, the inequality

$$\begin{aligned}&\frac{16(1+\kappa)^4c_\theta}{\mu}t^{-\min\{\theta,1-\theta\}}\ln\left(\frac{8t}{1-\theta}\right)\\&\leq\frac{16(1+\kappa)^4}{\mu}c_\theta\Big(\ln(\frac{8}{1-\theta})+\frac{1}{\min\{\theta,1-\theta\}}\Big)\leq 1\end{aligned}$$

and the definition of c_θ given in Lemma 6 yield the desired result. □

We are ready to establish the uniform bounds for the learning sequence.

Proposition 2. *Let $\lambda\geq 0$ and $\{f_t:t\in\mathbb{N}_{T+1}\}$ be produced by algorithm (1.4). Moreover, if the step sizes satisfy $\eta_t(\kappa^2+\lambda)\leq 1$ for $t\in\mathbb{N}_T$ and the inequality (4.28) then we have that*

$$\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]\leq 20\|f_\rho\|_\rho^2+3\mathcal{E}(f_\rho),\qquad\forall\ t\in\mathbb{N}_{T+1}. \tag{4.29}$$

Proof. Since $f_1=0$, from (1.6) it follows that

$$\mathcal{E}(f_1)=\|f_1-f_\rho\|_\rho^2+\mathcal{E}(f_\rho)=\|f_\rho\|_\rho^2+\mathcal{E}(f_\rho),$$

which implies that (4.29) holds true for $t=1$. As the induction assumption, suppose (4.29) holds true for all $k\in\mathbb{N}_t$. To advance the induction, we need to estimate $\mathbb{E}_{Z^t}\big[\mathcal{E}(f_{t+1})\big]$.

Applying (4.13) with $T=t\in\mathbb{N}$, $\mathbb{E}_{Z^t}\big[\|f_{t+1}-f_\rho\|_\rho^2\big]$ is bounded by

$$\begin{aligned}&\|f_\lambda-f_\rho\|_\rho^2+\|\omega_1^t(L_K+\lambda I)f_\lambda\|_\rho^2+2\|\omega_1^t(L_K+\lambda I)f_\lambda\|_\rho\|f_\lambda-f_\rho\|_\rho\\&+2(1+\kappa)^4\sum_{k\in\mathbb{N}_t}\frac{\eta_k^2\exp\{-\lambda\sum_{j=k+1}^t\eta_j\}}{1+\sum_{j=k+1}^t\eta_j}\mathbb{E}_{Z^{k-1}}\big[\mathcal{E}(f_k)\big].\end{aligned} \tag{4.30}$$

Note that $\|f_\lambda - f_\rho\|_\rho \le \|f_\rho\|_\rho$ by (4.11). Since $\eta_t(\kappa^2+\lambda) \le 1$ for $t \in \mathbb{N}$, $\|\omega_1^t(L_K+\lambda I)\| \le \prod_{j\in\mathbb{N}_t} \|I - \eta_j(L_K+\lambda I)\| \le 1$, Hence,

$$\|\omega_1^t(L_K+\lambda I)f_\lambda\|_\rho^2 \le \|\omega_1^t(L_K+\lambda I)\|^2\|f_\lambda\|_\rho^2 \le \|f_\lambda\|_\rho^2 \le 4\|f_\rho\|_\rho^2.$$

Similarly, by (4.3), there holds that

$$\|\omega_1^t(L_K+\lambda I)f_\lambda\|_\rho\|f_\lambda - f_\rho\|_\rho \le 2\|f_\rho\|_\rho^2.$$

Putting the above estimates, the induction assumption, and hypothesis (4.28) into (4.30) implies that

$$\mathbb{E}_{Z^t}\big[\|f_{t+1} - f_\rho\|_\rho^2\big] \le 9\|f_\rho\|_\rho^2 + \Big(20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)\Big)/4.$$

Note that $\mathcal{E}(f_{t+1}) = \mathcal{E}(f_\rho) + \|f_{t+1} - f_\rho\|_\rho^2$ by (1.6), and thus

$$\mathbb{E}_{Z^T}\big[\mathcal{E}(f_{t+1})\big] \le 20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho),$$

which advances the induction and completes the proof. □

In the following sections, we apply Proposition 1 and the basic estimates for the approximation error and cumulative sample error to prove our main results stated in Section 2.

5 Error bounds and convergence

In this section, we mainly develop error bounds and convergence results for algorithm (1.5) based on the Proposition 1.

In this case, we have that $\lambda = 0$ and $f_0 = f_\rho$. Therefore, the error decomposition formula (4.7) is reduced to

$$f_{T+1} - f_\rho = -\omega_1^T(L_K)f_\rho + \sum_{t\in\mathbb{N}_T} \eta_t\omega_{t+1}^T(L_K)\mathcal{B}(f_t, z_t). \tag{5.1}$$

By Proposition 1, $\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big]$ is bounded by

$$\|\omega_1^T(L_K)f_\rho\|_\rho^2 + 2(1+\kappa)^4 \sum_{k\in\mathbb{N}_T} \frac{\eta_t^2}{1+\sum_{j=t+1}^T \eta_j}\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]. \tag{5.2}$$

5.1 Proofs of error bounds

This subsection proves error bounds stated in Theorems 1 and 4. The essential estimates will also be used to derive error rates in Section 6.

We start with the following useful lemma. One can find its modified form in [32].

Lemma 7. *If $f \in L_K^{\beta}(\mathcal{L}^2_{\rho_X})$ with some $\beta > 0$ then*

$$\|\omega_1^T(L_K)f\|_\rho \le 2\Big(\Big(\frac{\beta}{e}\Big)^{\beta} + \kappa^{2\beta}\Big)\|L_K^{-\beta}f\|_\rho\Big(\sum_{t\in\mathbb{N}_T}\eta_t\Big)^{-\beta}. \tag{5.3}$$

In addition, if $\beta > \frac{1}{2}$ then $\|\omega_1^T(L_K)f\|_K$ is bounded by

$$2\Big(\Big(\frac{2\beta-1}{2e}\Big)^{\beta-\frac{1}{2}} + \kappa^{2\beta-1}\Big)\|L_K^{-\beta}f\|_\rho\Big(\sum_{t\in\mathbb{N}_T}\eta_t\Big)^{-(\beta-\frac{1}{2})}. \tag{5.4}$$

Proof. Applying Lemma 3 with $\lambda = 0, j = 1$, and $k = T$, the desired estimates (5.3) and (5.4) follow immediately from the following observations

$$\|\omega_1^T(L_K)f\|_\rho \le \|\omega_1^T(L_K)L_K^{\beta}\|\ \|L_K^{-\beta}f\|_\rho$$

and

$$\|\omega_1^T(L_K)f\|_K = \|\omega_1^T(L_K)L_K^{-\frac{1}{2}}f\|_\rho \le \|\omega_1^T(L_K)L_K^{\beta-\frac{1}{2}}\|\ \|L_K^{-\beta}f\|_\rho.$$

This completes our lemma. □

Recall the definitions of $\mu(\theta)$ in Corollary 2 and c_θ in Lemma 6, our error bounds read as follows.

Proposition 3. *Let $\lambda = 0,\ \theta \in [0,1)$ and $\eta_t = \frac{1}{\mu}t^{-\theta}$ for $t \in \mathbb{N}$ with some constant $\mu \ge \mu(\theta)$. Define $\{f_t : t \in \mathbb{N}\}$ by (1.5). Then, for any $T \in \mathbb{N}$ we have that*

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] \le \|\omega_1^T(L_K)f_\rho\|_\rho^2 + \frac{c_\rho c_\theta}{\mu}\ T^{-\min\{\theta,1-\theta\}}\ln\Big(\frac{8T}{1-\theta}\Big)$$

where $c_\rho := 4(1+\kappa)^4\big(20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)\big)$.

Proof. Since $\eta_t = \frac{1}{\mu}t^{-\theta}$ with $\mu \ge \mu(\theta)$ for any $t \in \mathbb{N}_T$, by Corollary 2 we know that the inequality (4.28) on the step sizes holds true for any

$t \in \mathbb{N}_T$. Hence, the bound (4.29) in Proposition 2 holds true. Now, the desired result follows by applying (4.26) with $\lambda = 0$ and $\ell = T$ to the right-hand side of (5.2). □

To apply Proposition 3 to prove Theorem 1, we need an easy observation, that is, if $0 < \theta < 1$ then, for any $t \in \mathbb{N}_T$,

$$\sum_{j=t}^{T} j^{-\theta} \geq \frac{(T+1)^{1-\theta} - t^{1-\theta}}{1-\theta}. \tag{5.5}$$

We are in a position to establish Theorem 1 stated in Section 2.

Proof of Theorem 1. Recall that $\eta_t = \frac{1}{\mu} t^{-\theta}$ for $\theta \in (0,1)$. Therefore, by (5.5) we have that $\displaystyle\sum_{t=1}^{T} \eta_t \geq \frac{1 - 2^{\theta-1}}{(1-\theta)\mu} T^{1-\theta}$. Putting this and property (a) in Lemma 5 together, the bound (2.4) immediately follows from Proposition 3 with $\theta \in (0,1)$. □

It still remains a question whether we can estimate the stronger error $\|f_{T+1} - f_\rho\|_K$ when the step sizes have the form $\{\eta_t = O(t^{-\theta}) : t \in \mathbb{N}\}$ with $\theta \in (0,1)$ and $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > \frac{1}{2}$.

We turn our attention to the case when the step sizes are in the form of $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ with $\eta = \eta(T)$ depending on T. In this case, we can deal with the expectation of the error $\|f_{T+1} - f_\rho\|_K^2$ if $f_\rho \in \mathcal{H}_K$.

Lemma 8. *Let $f_\rho \in \mathcal{H}_K$, $\eta_t = \eta$ for $t \in \mathbb{N}_T$ with $\eta\kappa^2 \leq 1$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be defined as (1.5). Then we have that*

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_K^2\big] \leq \|\omega_1^T(L_K) f_\rho\|_K^2 + \kappa^2\eta^2 \sum_{t \in \mathbb{N}_T} \mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big].$$

Proof. Since the argument here is essentially similar to the proof of Proposition 1 except the norm $\|\cdot\|_\rho$ is replaced by $\|\cdot\|_K$ in $\mathcal{H}_K$, we only sketch its proof.

From the error decomposition (5.1), we know that

$$\begin{aligned}\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_K^2\big] &= \|\omega_1^T(L_K) f_\rho\|_K^2 + \eta^2 \mathbb{E}_{Z^T}\Big[\Big\| \sum_{t \in \mathbb{N}_T} \omega_{t+1}^T(L_K) \mathcal{B}(f_t, z_t)\Big\|_K^2\Big] \\ &\quad - 2\eta \sum_{t \in \mathbb{N}_T} \big\langle \omega_1^T(L_K) f_\rho, \mathbb{E}_{Z^T}\big[\omega_{t+1}^T(L_K)\mathcal{B}(f_t, z_t)\big]\big\rangle_K. \end{aligned} \tag{5.6}$$

By the vanishing property (4.15) of $\mathcal{B}(f_t, z_t)$, the second term on right-hand side of the above inequality is estimated as follows:

$$\mathbb{E}_{Z^T}\big[\| \sum_{t\in\mathbb{N}_T} \omega_{t+1}^T(L_K)\mathcal{B}(f_t,z_t)\|_K^2\big] = \sum_{t\in\mathbb{N}_T} \mathbb{E}_{Z^t}\big[\|\omega_{t+1}^T(L_K)\mathcal{B}(f_t,z_t)\|_K^2\big]$$

Consequently,

$$\begin{aligned}\mathbb{E}_{Z^T}\big[\| \sum_{t\in\mathbb{N}_T} \omega_{t+1}^T(L_K)\mathcal{B}(f_t,z_t)\|_K^2\big] &= \sum_{t\in\mathbb{N}_T} \mathbb{E}_{Z^t}\big[\|\omega_{t+1}^T(L_K)L_K^{-\frac{1}{2}}\mathcal{B}(f_t,z_t)\|_\rho^2\big] \\ &\le \sum_{t\in\mathbb{N}_T} \|\omega_{t+1}^T(L_K)\|^2 \mathbb{E}_{Z^t}\big[\|L_K^{-\frac{1}{2}}\mathcal{B}(f_t,z_t)\|_\rho^2\big] \\ &= \sum_{t\in\mathbb{N}_T} \|\omega_{t+1}^T(L_K)\|^2 \mathbb{E}_{Z^t}\big[\|\mathcal{B}(f_t,z_t)\|_K^2\big]. \end{aligned} \tag{5.7}$$

By (4.20), $\mathbb{E}_{Z^t}\big[\big\|\mathcal{B}(f_t,z_t)\big\|_K^2\big] \le \kappa^2 \mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]$. Also, since $\eta\kappa^2 \le 1$, the estimate (4.11) yields that, for any $t \in \mathbb{N}_T$ there holds $\|\omega_{t+1}^T(L_K)\| \le 1$. Putting these estimates back into (5.7), we have that

$$\mathbb{E}_{Z^T}\big[\| \sum_{t\in\mathbb{N}_T} \omega_{t+1}^T(L_K)\mathcal{B}(f_t,z_t)\|_K^2\big] \le \kappa^2 \sum_{t=1}^T \mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big]. \tag{5.8}$$

Using (4.15) again, the last term on the right-hand side of (5.6) follows that $\sum_{t\in\mathbb{N}_T} \langle \omega_1^T(L_K)f_\rho, \mathbb{E}_{Z^T}\big[\omega_{t+1}^T(L_K)\mathcal{B}(f_t,z_t)\big]\rangle_K = 0$. Cascading this equality, (5.6) and (5.8) yields the desired estimate. □

We are ready to present error bounds when the step sizes are of the form $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ with $\eta = \eta(T)$ depending on T.

Proposition 4. *Let $\eta_t = \eta$ for $t \in \mathbb{N}_T$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be defined as (1.5). If $16(\kappa+1)^4 \ln\big(8T\big)\eta \le 1$ then there holds*

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] \le \|\omega_1^T(L_K)f_\rho\|_\rho^2 + c_\rho \eta \ln\big(8T\big). \tag{5.9}$$

In addition, if $f_\rho \in \mathcal{H}_K$ then we have that

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_K^2\big] \le \|\omega_1^T(L_K)f_\rho\|_K^2 + c_\rho \eta^2 T. \tag{5.10}$$

Proof. We regard η as $\frac{1}{\mu}$. Then, the inequality $16(\kappa+1)^4 \ln\big(8T\big)\eta \le 1$ means that $\mu \ge 16(1+\kappa)^4 \ln\big(8T\big)$. By Corollary 2, the inequality (4.28)

on the step sizes is satisfied. Hence, the estimate (4.29) in Proposition 2 holds true.

The first estimate (5.9) follows from Proposition 3 in the case of $\theta = 0$. Combining Lemma 8 and (4.29) yields (5.10). □

From the above proposition, we can prove Theorem 4 stated in the Section 2.

Proof of Theorem 4. The error bound (2.10) is derived from (5.9) and property (a) in Lemma 5. □

5.2 Proofs of convergence results

We are in a position to apply the error bounds in Theorems 1 and 4 to prove Theorems 2 and 5, while Propositions 3 and 4 will be used to derive the explicit rates in the next section.

To prove Theorem 2, we introduce the following well-known property of $\mathcal{K}$-functional, see [5]. We include its proof for completeness.

Lemma 9. *Let $s > 0$ and $\mathcal{K}(s, f_\rho)$ be defined by (2.2). Then there holds*

$$\lim_{s\to 0+} \mathcal{K}(s, f_\rho) = \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho.$$

Proof. By the definition (2.2) of $\mathcal{K}$-functional, it is straightforward that $\underline{\lim}_{s\to 0+}\mathcal{K}(s, f_\rho) \geq \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho$. Hence, we only need to show that

$$\overline{\lim}_{s\to 0+}\mathcal{K}(s, f_\rho) \leq \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho. \tag{5.11}$$

To this end, for any $\varepsilon > 0$ we know that there exists $f_\varepsilon \in \mathcal{H}_K$ such that $\|f_\varepsilon - f_\rho\|_\rho \leq \varepsilon + \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho$. Therefore, by letting $f = f_\varepsilon$, we know, for any $s > 0$, that $\mathcal{K}(s, f_\rho)$ is bounded by $\varepsilon + \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho + s\|f_\varepsilon\|_K$. Letting $s \to 0+$ yields (5.11), and hence completes the lemma. □

We are ready to prove Theorem 2 based on Lemma 9 and Theorem 1.

Proof of Theorem 2. Since $\theta \in (0, 1)$ and $\eta_t = \frac{t^{-\theta}}{\mu}$ for any $t \in \mathbb{N}$ with $\mu \geq \mu(\theta)$, Theorem 1 holds true. Applying Lemma 9 with $s = b_\theta\sqrt{\mu}\ T^{-(1-\theta)/2}$ to (2.4) yields that $\overline{\lim}_{T\to\infty}\mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|_\rho^2\right] \leq \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho^2$.

On the other hand, since $f_{T+1} \in \mathcal{H}_K$, for any $T \in \mathbb{N}$ and any sample $\mathbf{z} = \{z_t : t \in \mathbb{N}_T\}$, we know that $\inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho^2 \le \|f_{T+1} - f_\rho\|_\rho^2$, which leads to $\underline{\lim}_{T\to\infty} \mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|_\rho^2\right] \ge \inf_{f\in\mathcal{H}_K} \|f - f_\rho\|_\rho^2$. This completes Theorem 2. □

To prove Theorem 5, we further need the following observation.

Lemma 10. *Let $f_\rho \in \mathcal{H}_K$, the step sizes $\{\eta_t = \eta(T) : t \in \mathbb{N}_T\}$ satisfy $\eta(T)\kappa^2 \le 1$ for any $T \in \mathbb{N}$ and $\lim_{T\to\infty} \eta(T)T = \infty$. Then we have that*

$$\lim_{T\to\infty} \|\omega_1^T(L_K) f_\rho\|_K = 0.$$

Proof. Recall the definition of $L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > 0$ and the fact $\mathcal{H}_K = L_K^{\frac{1}{2}}(\mathcal{L}^2_{\rho_X})$. Hence, if $f_\rho \in \mathcal{H}_K$ then

$$f_\rho = \sum_{j=1}^\infty \lambda_j^{\frac{1}{2}} a_j \phi_j, \quad \text{for some} \quad \{a_j : j \in \mathbb{N}\} \in \ell^2.$$

Before we move to the next step, it will be helpful to sketch the main ideas. In what follows we construct functions in $L_K^1(\mathcal{L}^2_{\rho_X})$ to approximate f_ρ arbitrarily in $\mathcal{H}_K$ while for functions in $L_K^1(\mathcal{L}^2_{\rho_X})$, we apply (5.4) in Lemma 7 with $\beta = 1$ to deal with it.

Now, write $f_\rho = \sum_{\lambda_j>0, j=1}^N \lambda_j^{\frac{1}{2}} a_j \phi_j + \sum_{\lambda_j>0, j=N+1}^\infty \lambda_j^{\frac{1}{2}} a_j \phi_j$. Then, $f_N = \sum_{\lambda_j>0, j=1}^N \lambda_j^1 (\lambda_j^{-\frac{1}{2}} a_j)\phi_j \in L_K^1(\mathcal{L}^2_{\rho_X})$ for any $N \in \mathbb{N}$ because $\|L_K^{-1} f_N\|_\rho^2 = \sum_{\lambda_j>0, j=1}^N (\lambda_j^{-\frac{1}{2}} a_j)^2 < \infty$. On the other hand, since $\sum_{j=1}^\infty a_j^2 < \infty$, then, for any $\varepsilon > 0$, there exists $N(\varepsilon) \in \mathbb{N}$ such that $\sum_{j=N(\varepsilon)+1}^\infty a_j^2 \le \varepsilon^2$. Therefore,

$$\|f_\rho - f_{N(\varepsilon)}\|_K^2 = \|\sum_{j=N(\varepsilon)+1}^\infty \lambda_j^{\frac{1}{2}} a_j \phi_j\|_K^2 = \sum_{j=N(\varepsilon)+1}^\infty a_j^2 \le \varepsilon^2.$$

By (2.8), we have that $\|(I - \eta(T)L_K) f\|_K = \|(I - \eta(T)L_K) L_K^{-\frac{1}{2}} f\|_\rho \le \|I - \eta(T)L_K\| \|L_K^{-\frac{1}{2}} f\|_\rho = \|I - \eta(T)L_K\| \, \|f\|_K$.

Also, since $\eta(T)\kappa^2 \le 1$, applying (4.11) with $x = \eta$ implies that $\|(I - \eta(T)L_K)\| \le 1$. Hence, for any $f \in \mathcal{H}_K$ there holds $\|(I - \eta(T)L_K) f\|_K \le \|f\|_K$. Consequently, applying the above inequality with $f = f_\rho - f_{N(\varepsilon)}$

tells us that

$$\begin{aligned}\|(I-\eta(T)L_K)^T f_\rho\|_K &\le \|(I-\eta(T)L_K)^T f_{N(\varepsilon)}\|_K \\ &\quad + \|(I-\eta(T)L_K)^T (f_\rho - f_{N(\varepsilon)})\|_K \\ &\le \|(I-\eta(T)L_K)^T f_{N(\varepsilon)}\|_K + \|f_\rho - f_{N(\varepsilon)}\|_K \\ &\le \|(I-\eta(T)L_K)^T f_{N(\varepsilon)}\|_K + \varepsilon. \end{aligned} \tag{5.12}$$

To estimate $\|(I-\eta(T)L_K)^T f_{N(\varepsilon)}\|_K$, applying (5.4) with $\eta_t = \eta(T)$ for any $t \in \mathbb{N}_T$, $f = f_{N(\varepsilon)}$, and $\beta = 1$ yields that

$$\|(I-\eta(T)L_K)^T f_{N(\varepsilon)}\|_K \le 2\big(1+\kappa\big)\|L_K^{-1} f_{N(\varepsilon)}\|_\rho \big(T\eta(T)\big)^{-\frac{1}{2}}.$$

Since $\lim_{T\to\infty} T\eta(T) = \infty$, the above estimation implies that there exists an integer $T(\varepsilon) \ge N(\varepsilon)$ such that

$$\|(I-\eta(T)L_K)^T f_{N(\varepsilon)}\|_K \le \varepsilon, \qquad \forall\ T \ge T(\varepsilon). \tag{5.13}$$

Putting (5.12) and (5.13) together, we know that

$$\|\omega_1^T(L_K) f_\rho\|_K = \|(I-\eta(T)L_K)^T f_\rho\|_K \le 2\varepsilon, \qquad \forall\ T \ge T(\varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, we obtain the desired claim. □

We now move on to the proof of Theorem 5.

Proof of Theorem 5. Observe whenever $\lim_{T\to\infty} \eta(T)\ln T = 0$ or $\lim_{T\to\infty} T\eta^2(T) = 0$, there exists $T_1(\varepsilon) \in \mathbb{N}$ such that $16(\kappa+1)^4\eta(T)\ln T \le 1$ for all $T \ge T_1(\varepsilon)$. The hypotheses in Lemma 8 and Theorem 4 are satisfied for any $T \ge T_1(\varepsilon)$. Therefore, (2.10) holds true for any $T \ge T_1(\varepsilon)$.

We first prove the convergence in $\mathcal{L}^2_{\rho_X}$. Applying Lemma 9 with $s = 2(1+\kappa)\big(\eta(T)T\big)^{-\frac{1}{2}}$ to (2.10), it follows from the hypothesis $\lim_{T\to\infty} T\eta(T) = \infty$ that $\overline{\lim}_{T\to\infty}\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\rho\|_\rho^2\big] \le \inf_{f\in\mathcal{H}_K}\|f-f_\rho\|_\rho^2$. On the other hand, since $f_{T+1} \in \mathcal{H}_K$, $\inf_{f\in\mathcal{H}_K}\|f-f_\rho\|_\rho^2 \le \|f_{T+1}-f_\rho\|_\rho^2$ for any data $\mathbf{z} = \{z_t : t \in \mathbb{N}_T\}$ which leads to $\inf_{f\in\mathcal{H}_K}\|f-f_\rho\|_\rho^2 \le \underline{\lim}_{T\to\infty}\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\rho\|_\rho^2\big]$. This verifies the first part of Theorem 5.

The proof for the second part is similar. Since $\lim_{T\to\infty} T\eta^2(T) = 0$, combining (5.10) in Proposition 4 with Lemma 10 yields that

$$\overline{\lim}_{T\to\infty}\mathbb{E}_{Z^T}\big[\|f_{T+1}-f_\rho\|_K^2\big] \le 0$$

which completes Theorem 5. □

6 Explicit error rates

In this section we use Propositions 3 and 4 to derive explicit error rates for $\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big]$ by choosing the step sizes appropriately, when the regression function f_ρ lies in $L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > 0$.

6.1 Rates for online learning with $\lambda = 0$

We start by establishing explicit error rates for algorithm (1.5) stated in Theorems 3 and 6.

Proof of Theorem 3. Recall that $\mu(\theta)$ is defined by (4.27) for $0 < \theta < 1$. Therefore, if $\eta_t = \frac{t^{-\theta}}{\mu(\theta)}$ then, by Proposition 3, we know that

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] \le \|\omega_1^T(L_K)f_\rho\|_\rho^2 + O\Big(T^{-\min\{\theta,1-\theta\}}\ \ln T\Big). \tag{6.1}$$

Since $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$, applying (5.3) with $f = f_\rho$ implies that $\|\omega_1^T(L_K)f_\rho\|_\rho = O\Big(\big(\sum_{t\in\mathbb{N}_T}\eta_t\big)^{-\beta}\Big)$. Therefore, we see from (5.5) that $\|\omega_1^T(L_K)f_\rho\|_\rho^2 = O\Big(T^{-2(1-\theta)\beta}\Big)$. Note that $O\Big(T^{-\min\{\theta,1-\theta\}}\ \ln T\Big) = O\Big(T^{-\theta}\ \ln T + T^{-(1-\theta)}\ \ln T\Big)$. Putting these estimates into (6.1) and noting the hypothesis $\beta \in (0, \frac{1}{2}]$, we know that

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] = O\Big(T^{-2(1-\theta)\beta}\ \ln T + T^{-\theta}\ \ln T\Big).$$

Selecting $\theta = \frac{2\beta}{2\beta+1}$ in the above inequality completes the theorem. □

By Proposition 4, we can establish Theorem 6.

Proof of Theorem 6. Observe that $\ln(8T) \le \frac{(8T)^\theta}{\theta} \le \frac{8T^\theta}{\theta}$ for any $\theta \in (0,1)$. Hence, we know that the choice of $\eta := \frac{\beta}{64(1+\kappa)^4(2\beta+1)}T^{-\frac{2\beta}{2\beta+1}}$ for $\beta > 0$ satisfies the hypothesis $16(\kappa+1)^4\ln\big(8T\big)\eta \le 1$ in Proposition 4.

Applying (5.3) with $f = f_\rho$, we know that $\|\omega_1^T(L_K)f_\rho\|_\rho = O((\eta T)^{-\beta})$. This in connection with (5.9) in Proposition 4 implies that

$$\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] \le O((\eta T)^{-2\beta}) + c_\rho\eta\ln\big(8T\big).$$

Substituting $\eta = \frac{\beta}{64(1+\kappa)^4(2\beta+1)}T^{-\frac{2\beta}{2\beta+1}}$ into the right-hand side of the above inequality yields the rate (2.15).

Similarly, if $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ with $\beta > \frac{1}{2}$ then, applying (5.4) with $f = f_\rho$ to (5.10) yields that

$$\begin{aligned}\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_K^2\big] &\le \|\omega_1^T(L_K)f_\rho\|_K^2 + c_\rho\eta^2 T \\ &\le O\Big(\big(T\eta\big)^{-2(\beta-\frac{1}{2})}\Big) + c_\rho\eta^2 T.\end{aligned}$$

Putting the choice $\eta = \frac{\beta}{64(1+\kappa)^4(2\beta+1)}T^{-\frac{2\beta}{2\beta+1}}$ back into the right-hand side of the above inequality directly implies the rate (2.16). □

6.2 Improved rates for online regularized learning

Our analysis can also yield improved rates for the online regularized algorithm given by (1.4) with $\lambda > 0$ which is stated as Theorem 7.

Proof of Theorem 7. Note that if $0 < \lambda \le 1$ then Corollary 2 tells us that the choice of the step sizes $\{\eta_t = \frac{1}{\mu(\theta)+1}t^{-\theta} : t \in \mathbb{N}_T\}$ with $\theta \in (0,1)$ satisfies the hypothesis (4.28). Consequently, by Proposition 2 we have that

$$\sup_{k\in\mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t)\big] \le 20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho).$$

Also, observe that $\mu := \mu(\theta) + 1 \ge \kappa^2 + \lambda$ for $\lambda \in (0,1]$ and denote $d_\theta = \frac{1-2^{\theta-1}}{1-\theta}$. Therefore, applying Lemma 6 with $\ell = T$ implies that $\displaystyle\sum_{t\in\mathbb{N}_T} \frac{\eta_t^2 \exp\big\{-\lambda\sum_{j=t+1}^T \eta_j\big\}}{1+\sum_{j=t+1}^T \eta_j}$ is bounded by

$$\frac{c_\theta}{\mu}\Big(\exp\big\{-\lambda d_\theta T^{1-\theta}/\mu\big\}T^{-\min\{\theta,1-\theta\}} + T^{-\theta}\Big)\ln\big(\frac{8T}{1-\theta}\big). \tag{6.2}$$

Combining this with Proposition 1 for the case $\lambda \in (0,1]$, $\mathbb{E}_{Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big]$ is bounded by

$$\begin{aligned}&\|f_\lambda - f_\rho\|_\rho^2 + \|\omega_1^T(L_K+\lambda I)f_\lambda\|_\rho^2 + 2\|f_\lambda - f_\rho\|_\rho\|\omega_1^T(L_K+\lambda I)f_\lambda\|_\rho \\ &+ \frac{c_\rho c_\theta}{\mu}\Big(\exp\big\{-\lambda d_\theta T^{1-\theta}/\mu\big\}T^{-\min\{\theta,1-\theta\}} + T^{-\theta}\Big)\ln(\frac{8T}{1-\theta}).\end{aligned} \tag{6.3}$$

To move to the next step, we note from Lemma 4 that

$$\|f_\lambda - f_\rho\|_\rho \le \lambda^\beta \|L_K^{-\beta} f_\rho\|_\rho, \qquad \forall \beta \in (0,1].$$

Also, property (b) in Lemma 5 and the inequality (5.5) with $t = 1$ implies that

$$\|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho \leq 2\exp\{-\lambda d_\theta T^{1-\theta}/\mu\}\|f_\rho\|_\rho.$$

Substituting these into (6.3), it follows that $\mathbb{E}_{Z^T}\left[\|f_{T+1} - f_\rho\|_\rho^2\right]$ is bounded by

$$\lambda^{2\beta}\|L_K^{-\beta}f_\rho\|_\rho^2 + \frac{c_\rho c_\theta}{\mu}T^{-\theta}\ln(\frac{8T}{1-\theta}) + 4\|L_K^{-\beta}f_\rho\|_\rho\|f_\rho\|_\rho\lambda^\beta\exp\{-\lambda d_\theta T^{1-\theta}/\mu\}$$
$$+ \Big(\frac{c_\rho c_\theta}{\mu} + 4\|f_\rho\|_\rho^2\Big)\exp\{-\lambda d_\theta T^{1-\theta}/\mu\}\ln\big(\frac{8T}{1-\theta}\big). \qquad (6.4)$$

Note that, for all $\epsilon > 0, s > 0$ and $c > 0$ the asymptotic behavior holds

$$\exp\{-cT^\epsilon\} = O(T^{-s}). \qquad (6.5)$$

Therefore, for any $0 < \varepsilon < \frac{\beta}{2\beta+1}$, choosing $\lambda = T^{-\frac{1}{2\beta+1}+\frac{\varepsilon}{2\beta}}$ and $\theta = \frac{2\beta}{2\beta+1}$ in (6.4) yields the desired error rate (2.17). □

Acknowledgements. The authors would like to thank Steve Smale, Mark Herbster, Yuan Yao, and Lorenzo Rosasco for many helpful discussions. They are also grateful to the anonymous referees for valuable comments and suggestions. This work was supported by EPSRC Grant GR/T18707/01 and by the IST Programme of the European Community, under the PASCAL Network of Excellence IST-2002-506778.

References

[1] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337–404.

[2] K. S. Azoury, and M. K. Warmuth, Relative loss bounds for on-line density estimation with the exponential family of distributions, *Machine Learning* **43** (2001), 211-246.

[3] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, Convexity, classification, and risk bounds, *J. Amer. Statist. Assoc.*, **101** (2005), 138-156.

[4] F. Bauer, S. Pereverzev and L. Rosasco, On regularization algorithms in learning theory, to appear in *J. Complexity*, 2006.

[5] J. Bergh and J. Löfström, *Interpolation spaces, an introduction,* Springer-Verlag, New York, 1976.

[6] T. Cai, Rates of convergence and adaption over Besov spaces under pointwise risk, *Statistica Sinica* **13** (2003), 881-902.

[7] A. Caponnetto and E. De Vito, Fast rates for regularized least squares algorithm, preprint, 2005. Available at http://cbcl.mit.edu/cbcl/publications/ai-publications/2005/AIM-2005-013.pdf

[8] N. Cesa-Bianchi, A. Conconi, and C. Gentile, On the generalization ability of on-line learning algorithms, *IEEE Trans. Inform. Theory* **50** (2004), 2050-2057.

[9] N. Cesa-Bianchi, A. Conconi and C. Gentile, A second-order perceptron algorithm, *SIAM J. Comput.* **34** (2005), 640-688.

[10] N. Cesa-Bianchi, P. Long, and M. K. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent, *IEEE Trans. Neural Networks* **7** (1996), 604–619.

[11] D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou, Support vector machine soft margin classifiers: error analysis, *J. Machine Learning Res.* **5** (2004), 1143–1175.

[12] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **39** (2001), 1–49.

[13] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory, *Found. Comput. Math.* **4** (2002), 413–428.

[14] E. De Vito, A. Caponnetto, and L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, *Found. Comput. Math.* **5** (2005), 59–85.

[15] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems,* Kluwer, Dordrecht, The Netherlands, 1996.

[16] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.

[17] J. Forster and M. K. Warmuth, Relative expected instantaneous loss bounds, *J. Computer and System Sciences* **64** (2002), 76-102.

[18] M. Herbster and M. K. Warmuth, Tracking the best expert, *Machine Learning* **32** (1998), 151-178.

[19] J. Kivinen, A. J. Smola, and R. C. Williamson, Online learning with kernels, *IEEE Trans. Signal Processing* **52** (2004), 2165–2176.

[20] M. Pontil, Y. Ying, and D. X. Zhou, Error analysis for online gradient descent algorithms in reproducing kernel Hilbert spaces, *Technical Report, Department of Computer Science, University College London,* December 2005.

[21] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Statis.,* **22** (1951), 400-407.

[22] S. Smale and Y. Yao, Online learning algorithms, *Found. Comp. Math.*, **6** (2006), 145–170.

[23] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.* **1** (2003), 17–41.

[24] S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.* **41** (2004), 279-305.

[25] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.*, to appear.

[26] I. Steinwart, Support vector machines are universally consistent, *J. Complexity* **18** (2002), 768–791.

[27] C. J. Stone, Optimal global rates of convergence for nonparametric regression, *Annals of Statistics* **10** (1982), 1040-1053.

[28] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.

[29] V. Vovk, On-line regression competitive with reproducing kernel Hilbert spaces, Technical report, 2006. Available at http://arxiv.org/abs/cs.LG/0511058.

[30] G. Wahba, *Spline models for observational data*, SIAM, 1990.

[31] Q. Wu, Y. Ying, and D. X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.*, **6** (2005), 171-192.

[32] Y. Yao, L. Rosasco, and A. Caponnetto, Early stopping in gradient descent boosting, to appear in *Constr. Approx.*, 2006.

[33] Y. Ying and D. X. Zhou, Online regularized classification algorithms, *IEEE Trans. Inform. Theory,* 11 (2006), 4775-4788.

[34] T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comp.* **15** (2003), 1397–1437.

[35] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, *Proceedings of 21st International Conference on Machine Learning,* (Carla E. Brodley, ed.), ACM 2004.

[36] T. Zhang and B. Yu, Boosting with early stopping: convergence and consistency, *Annals of Statistics* **33** (2005), 1538-1579.

Appendix

This appendix includes a detailed proof of Lemma 6 and the relationship between the functional space $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$ and $L^\beta_K(\mathcal{L}^2_{\rho_X})$.

Proof of Lemma 6: Recall the convention $\sum_{k=\ell+1}^{\ell} \eta_k = 0$ which gives rise to the equality

$$\sum_{j\in\mathbb{N}_\ell} \eta_j^2 \exp\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\}/\big(\sum_{k=j+1}^{\ell} \eta_k + 1\big)$$
$$= \sum_{j\in\mathbb{N}_{\ell-1}} \eta_j^2 \exp\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\}/\big(\sum_{k=j+1}^{\ell} \eta_k + 1\big) + \frac{1}{\mu^2 \ell^{2\theta}}. \tag{6.6}$$

Let us first prove the case $\theta \neq 0$. The last term on the right-hand side of (6.6) is easy: $\frac{1}{\mu^2\ell^{2\theta}} \leq \frac{1}{\mu\ell^{2\theta}}$ since $\mu \geq 1 + \kappa^2$.

To estimate the first term on the right-hand side of (6.6), we divide it into two terms:

$$\sum_{j\in\mathbb{N}_{\ell-1}} \frac{\eta_j^2 \exp\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\}}{\sum_{k=j+1}^{\ell} \eta_k + 1} = \sum_{\substack{j\in\mathbb{N} \\ j\leq \frac{\ell-1}{2}}} \frac{\eta_j^2 \exp\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\}}{\sum_{k=j+1}^{\ell} \eta_k + 1}$$
$$+ \sum_{j>\frac{\ell-1}{2}, j\in\mathbb{N}}^{\ell-1} \frac{\eta_j^2 \exp\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\}}{\sum_{k=j+1}^{\ell} \eta_k + 1}. \tag{6.7}$$

By (5.5), we know that $\sum_{t=j+1}^{\ell} t^{-\theta} \geq \frac{1}{1-\theta}\big((\ell+1)^{1-\theta} - (j+1)^{1-\theta}\big)$. Then, for any $j \leq \ell - 1$, there holds

$$\begin{aligned}
&\exp\big\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\big\}/\big(\sum_{k=j+1}^{\ell} \eta_k + 1\big) \\
&\leq \frac{\exp\big\{-\frac{\lambda}{(1-\theta)\mu}\big((\ell+1)^{1-\theta} - (j+1)^{1-\theta}\big)\big\}}{\frac{1}{\mu(1-\theta)}\big((\ell+1)^{1-\theta} - (j+1)^{1-\theta}\big) + 1}.
\end{aligned} \tag{6.8}$$

Below, we use (6.8) to estimate the two terms on the right-hand side of (6.7) respectively.

For the first term, since $j \leq \frac{\ell-1}{2}$, $(\ell+1)^{1-\theta} - (j+1)^{1-\theta} \geq (1-2^{\theta-1})(\ell+1)^{1-\theta}$. Thus, the inequality (6.8) implies that

$$\begin{aligned}
&\sum_{\substack{j\in\mathbb{N} \\ j\leq\frac{\ell-1}{2}}} \eta_j^2 \exp\big\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\big\}/\big(\sum_{k=j+1}^{\ell} \eta_k + 1\big) \\
&\leq \frac{1}{\mu}\Big(\frac{1-\theta}{1-2^{\theta-1}}\Big)(\ell+1)^{\theta-1} \exp\Big\{-\frac{\lambda(1-2^{\theta-1})}{(1-\theta)\mu}(\ell+1)^{1-\theta}\Big\} \sum_{j\leq\frac{\ell-1}{2}} j^{-2\theta}.
\end{aligned}$$

Also,

$$\sum_{j\leq\frac{\ell-1}{2}} j^{-2\theta} \leq 1 + \int_1^{\frac{\ell-1}{2}} x^{-2\theta} dx \leq \begin{cases} \frac{2}{1-2\theta}\ell^{1-2\theta} & \text{for } 0 < \theta < \frac{1}{2}, \\ \ln(\frac{e\ell}{2}) & \text{for } \theta = \frac{1}{2}, \\ \frac{2}{2\theta-1} & \text{for } \frac{1}{2} < \theta < 1. \end{cases}$$

Consequently, it follows that

$$\begin{aligned}
&\sum_{\substack{j\in\mathbb{N} \\ j\leq\frac{\ell-1}{2}}} \eta_j^2 \exp\big\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\big\}/\big(\sum_{k=j+1}^{\ell} \eta_k + 1\big) \\
&\leq \frac{\widetilde{c}_\theta}{\mu}\ell^{-\min\{\theta,1-\theta\}} \exp\big\{-\frac{\lambda(1-2^{\theta-1})}{(1-\theta)\mu}(\ell+1)^{1-\theta}\big\} \ln\big(\frac{e\ell}{2}\big)
\end{aligned} \tag{6.9}$$

where $\widetilde{c}_\theta = 2$ for $\theta = \frac{1}{2}$ and $\frac{2}{(1-2^{\theta-1})|2\theta-1|}$ for $\theta \neq 0, \frac{1}{2}$.

We now turn to the second term on the right-hand side of (6.7). Since

$\exp\{-\frac{\lambda}{(1-\theta)\mu}\big((\ell+1)^{1-\theta}-(j+1)^{1-\theta}\big)\}\le 1$, from (6.8), we have that

$$\sum_{j>\frac{\ell-1}{2},j\in\mathbb{N}}^{\ell-1}\eta_j^2\exp\{-\lambda\sum_{k=j+1}^{\ell}\eta_k\}/\big(\sum_{k=j+1}^{\ell}\eta_k+1\big)$$
$$\le\frac{4}{\mu^2}\ell^{-\theta}\sum_{j>\frac{\ell-1}{2},j\in\mathbb{N}}^{\ell-1}\frac{j^{-\theta}}{\frac{1}{\mu(1-\theta)}\big((\ell+1)^{1-\theta}-(j+1)^{1-\theta}\big)+1}.$$

Since $j^{-\theta}\le 3(1+x)^{-\theta}$ and $(\ell+1)^{1-\theta}-(j+1)^{1-\theta}\ge(\ell+1)^{1-\theta}-(x+1)^{1-\theta}$ for any $x\in[j,j+1]$ and $j\le\ell-1$, we have that

$$\sum_{j>\frac{\ell-1}{2}}^{\ell-1}\frac{j^{-\theta}}{\frac{1}{\mu(1-\theta)}\big((\ell+1)^{1-\theta}-(j+1)^{1-\theta}\big)+1}$$
$$\le 3\sum_{j>\frac{\ell-1}{2}}^{\ell-1}\int_j^{j+1}\frac{(x+1)^{-\theta}}{\frac{1}{\mu(1-\theta)}\big((\ell+1)^{1-\theta}-(x+1)^{1-\theta}\big)+1}dx$$
$$\le 3\int_{\frac{\ell-1}{2}}^{\ell}\frac{(x+1)^{-\theta}}{\frac{1}{\mu(1-\theta)}\big((\ell+1)^{1-\theta}-(x+1)^{1-\theta}\big)+1}dx$$
$$=3\mu\ln\Big(1+\Big(\frac{1-2^{\theta-1}}{(1-\theta)\mu}\Big)(\ell+1)^{1-\theta}\Big)\le 3\mu\ln\Big(\frac{2(\ell+1)}{1-\theta}\Big)$$

where $\frac{1}{\mu}\le\frac{1}{1+\kappa^2}\le 1$ is used in the last inequality. Therefore,

$$\sum_{j>\frac{\ell-1}{2},j\in\mathbb{N}}^{\ell-1}\eta_j^2\exp\{-\lambda\sum_{k=j+1}^{\ell}\eta_k\}/\big(\sum_{k=j+1}^{\ell}\eta_k+1\big)\le\frac{12}{\mu}\ell^{-\theta}\ln\Big(\frac{2(\ell+1)}{1-\theta}\Big).$$

Putting this and (6.9) together into (6.6), we have, for $\theta\in(0,1)$, that

$$\sum_{j\in\mathbb{N}_\ell}\eta_j^2\exp\{-\lambda\sum_{k=j+1}^{\ell}\eta_k\}/\big(\sum_{k=j+1}^{\ell}\eta_k+1\big)$$
$$\le\frac{c_\theta}{\mu}\Big(\exp\{-\frac{\lambda(1-2^{\theta-1})}{(1-\theta)\mu}(\ell+1)^{1-\theta}\}\ell^{-\min\{\theta,1-\theta\}}+\ell^{-\theta}\Big)\ln(\frac{8\ell}{1-\theta}).$$

where $c_\theta=13$ for $\theta=\frac12$, and $\frac{13}{(1-2^{\theta-1})|2\theta-1|}$ for $\theta\in(0,\frac12)\cup(\frac12,1)$.

For the case $\theta=0$, we know that

$$\sum_{j\in\mathbb{N}_\ell}\eta_j^2\exp\{-\lambda\sum_{k=j+1}^{\ell}\eta_k\}/\big(\sum_{k=j+1}^{\ell}\eta_k+1\big)=\frac{1}{\mu^2}\sum_{j=1}^{\ell}e^{-\frac{(\ell-j)\lambda}{\mu}}/\big(\frac{\ell-j}{\mu}+1\big)$$
$$\le\frac{1}{\mu^2}\sum_{j\in\mathbb{N}_\ell}\frac{1}{\frac{\ell-j}{\mu}+1}\le\frac{1}{\mu}\ln\big(e\ell\big).$$

This finishes the proof of Lemma 6. □

We end the appendix with the relationship between $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$ and $L^\beta_K(\mathcal{L}^2_{\rho_X})$ which is interesting in its own right.

Proposition 5. *Let the interpolation space $(\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{\beta,\infty}$ be defined by (2.5). Then, $L^\beta_K(\mathcal{L}^2_{\rho_X}) \subseteq (\mathcal{L}^2_{\rho_X}, \mathcal{H}_K)_{2\beta,\infty}$ for $\beta \in [0, \frac{1}{2}]$.*

Proof. We assume that the positive eigenvalues $\{\lambda_j : j \in \mathbb{N}\}$ of L_K are arranged in decreasing order. Then, $\|L_K\| = \lambda_1 \le \kappa^2$.

For any $f_1 \in L^\beta_K(\mathcal{L}^2_{\rho_X})$, there exists $\{a_j : j \in \mathbb{N}\}$ with $\sum_{j\in\mathbb{N}} a_j^2 < \infty$ such that $f_1 = \sum_{j\in\mathbb{N}} \lambda_j^\beta a_j \phi_j$. To estimate $\mathcal{K}(s, f_1)$ in the case of $s^2 \le \lambda_1$, we select $\ell \in \mathbb{N}$ such that $\lambda_{\ell+1} \le s^2 \le \lambda_\ell$ and let $f = \sum_{j\in\mathbb{N}_\ell} a_j \lambda_j^\beta \phi_j$. Therefore,

$$\begin{aligned}\|f - f_1\|_\rho + s\|f\|_K &= \lambda_{\ell+1}^\beta \Big\{ \sum_{j\ge\ell+1} a_j^2 (\frac{\lambda_j}{\lambda_{\ell+1}})^{2\beta} \Big\}^{\frac{1}{2}} + s\lambda_\ell^{\beta-\frac{1}{2}} \Big\{ \sum_{j\in\mathbb{N}_\ell} a_j^2 \big(\frac{\lambda_j}{\lambda_\ell}\big)^{2\beta-1} \Big\}^{\frac{1}{2}} \\ &\le 2s^{2\beta} \Big\{ \sum_{j\in\mathbb{N}} a_j^2 \Big\}^{\frac{1}{2}} = 2s^{2\beta} \|L_K^{-\beta} f_1\|_\rho.\end{aligned}$$

Consequently, this proves that $\mathcal{K}(s, f_1) \le 2s^{2\beta}\|L_K^{-\beta} f_1\|_\rho$ in the case $s^2 \le \lambda_1$. For $s^2 \ge \lambda_1$, we choose $f = 0$, then $\mathcal{K}(s, f_1) \le \|f_1\|_\rho \le s^{2\beta}\lambda_1^{-\beta}\|f_1\|_\rho \le s^{2\beta}\|L_K^{-\beta} f_1\|_\rho$. Above all, for any $s > 0$,

$$\mathcal{K}(s, f_1) \le 2s^{2\beta}\|L_K^{-\beta} f_1\|_\rho$$

which completes the proposition. □

We remark with a simple counterexample that the reverse conclusion in Proposition 5 between these spaces is not true. Let $d\rho_X$ be the Lebesgue measure dx on $[0, 1]$ and the RKHS be $H^s[0, 1]$ with $s > \frac{1}{2}$. By the argument at the end of Section 3, we know that $L^\beta_K(\mathcal{L}^2_{dx}) = H^{2\beta s}[0, 1]$. But it is well-known [5] that $H^{2\beta s}[0, 1]$ is not identical to the interpolation space $(\mathcal{L}^2_{dx}, H^s)_{2\beta,\infty}$. We conjecture that this reverse conclusion is also not true in the general case.