

Online gradient descent learning algorithms[†]

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* explicit regularization. We present a novel capacity independent approach to derive error bounds and convergence results for this algorithm. 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 for both offline and online regularization algorithms in the literature.

Short Title: Online gradient descent learning

Keywords and Phrases: 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 7387 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 $\mathbb{N}_\ell := \{1, \dots, \ell\}$ for any $\ell \in \mathbb{N}$. Let ρ be a fixed but unknown distribution on $Z := X \times Y$. The generalization (true) error for a function $f : X \rightarrow Y$ is defined as

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

The function minimizing the above error is called the *regression function* which can be specified by

$$f_\rho(x) = \int_Y y d\rho(y|x), \quad x \in X \quad (1.2)$$

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 finite set of training data drawn from ρ . In this paper, we restrict our attention to online learning algorithms for computing an approximator in a *reproducing kernel Hilbert space* (RKHS).

Let $K : X \times X \rightarrow \mathbb{R}$ be a *Mercer kernel*, that is, a continuous, symmetric and positive semi-definite kernel, see, for example, [11]. 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*

$$\langle K_x, g \rangle_K = g(x). \quad (1.3)$$

Let $\mathbf{z} := \{z_t = (x_t, y_t) : t \in \mathbb{N}_T\}$ be a set of random samples independently distributed according to ρ . *The online gradient descent algorithms* [9, 18, 21, 27] provides an advantageous way to deal with large training sets and is defined as $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t ((f_t(x_t) - y_t)K_{x_t} + \lambda f_t), \quad t \in \mathbb{N}_T \quad (1.4)$$

where $\lambda \geq 0$ is called the *regularization parameter*. We call the sequence $\{\eta_t : t \in \mathbb{N}_T\}$ the *step sizes* or *learning rates* and $\{f_t : t \in \mathbb{N}_{T+1}\}$ the *learning sequence*. We can use f_{T+1} , the last output of (1.4), to approximate (learn) the regression function f_ρ .

The class of learning algorithms displayed above is also referred to as stochastic approximation algorithms in the setting of reproducing kernel

Hilbert spaces. Such a stochastic approximation procedure dates back to [20]. One can see [21, 27, 31, 35] and references therein for more background material.

When the parameter $\lambda > 0$, we call (1.4) the *online regularized algorithm* which is well studied in the recent literature, see, for example, [19, 21, 32]. In this paper, we are mainly concerned with the online gradient descent algorithm *without* explicit regularization given by (1.4) with $\lambda = 0$, that is, $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t(f_t(x_t) - y_t)K_{x_t}, \quad t \in \mathbb{N}_T. \quad (1.5)$$

Since the last output f_{T+1} is used to approximate the regression function f_ρ , the efficiency of the algorithm (1.5) can be measured by the difference between f_{T+1} and f_ρ . The nature of the least-square loss leads to the measurement in the metric of $\mathcal{L}_{\rho_X}^2$ defined as $\|f\|_\rho = \|f\|_{\mathcal{L}_{\rho_X}^2} := (\int_X |f(x)|^2 d\rho_X)^{1/2}$, where ρ_X is the marginal distribution of ρ on X . A direct computation yields that

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

see, for example, [11] 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 the error rates. We mainly focus on two different types of step sizes: $\{\eta_t : t \in \mathbb{N}_T\}$ being a subset of a *universal sequence* $\{\eta_t : t \in \mathbb{N}\}$ which is *independent* of the sample number T or $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ with $\eta = \eta(T)$ depending on T . In particular, we shall show, by choosing the step sizes appropriately, that the error rates of (1.5) are competitive with those of offline and online regularized learning algorithms in the literature.

The rest of the paper is organized as follows. The next section summarizes our main results and presents some comparisons with previous work. To formulate our basic ideas, in Section 3 we provide a novel approach to the error analysis of the online algorithms (1.4) under some conditions on the step sizes which we verify in Section 4 in the case of commonly used step sizes. In Section 5 we develop error bounds and convergence results for the online algorithm (1.5) without regularization. Finally, Section 6 establishes explicit error rates for the online algorithm (1.5) and, as a byproduct, improves the preceding rates for the online regularized algorithm given by (1.4) with $\lambda > 0$.

2 Main results

We begin with some background material and notations for subsequent use. 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

$$\|f\|_\infty \leq \kappa \|f\|_K. \quad (2.1)$$

Throughout this paper, we always assume that $\int_Z y^2 d\rho(z) < \infty$ which implies that the quantity $\mathcal{E}(f_\rho) + \|f_\rho\|_\rho^2$ is finite. Finally, we define, for any $\theta \in (0, 1)$, the quantity

$$\mu(\theta) := \begin{cases} 1248(1 + \kappa)^4 & \text{if } \theta = 1/2, \\ \frac{208(1 + \kappa)^4}{(1 - 2^{\theta-1})|2\theta-1|} \left(\ln\left(\frac{8}{1-\theta}\right) + \frac{1}{\min\{\theta, 1-\theta\}} \right) & \text{otherwise.} \end{cases} \quad (2.2)$$

2.1 Main theorems

We are now ready to state our main results. The first result 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 $\{f_t : t \in \mathbb{N}_{T+1}\}$ by (1.5). Then we have that*

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \leq \inf_{f \in \mathcal{H}_K} \left\{ \|f_\rho - f\|_\rho + b_\theta \sqrt{\mu} T^{-(1-\theta)/2} \|f\|_K \right\}^2 + \frac{c_\rho c_\theta}{\mu} T^{-\min\{\theta, 1-\theta\}} \ln\left(\frac{8T}{1-\theta}\right) \quad (2.3)$$

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

We note that the earlier versions of error bounds usually rely on the capacity of the space \mathcal{H}_K measured by the entropy and covering numbers, see, for example, [11, 12, 30]. Our upper bound on the right-hand side of the equality (2.3) is *capacity (kernel) independent*.

The first term on the right-hand side of the inequality (2.3) concerns the approximation of the function f_ρ in the $\mathcal{L}_{\rho_X}^2$ space by functions from the space \mathcal{H}_K . We observe that it can be specifically represented by the square norm of the K -functional which is defined as

$$\mathcal{K}(s, f_\rho) = \inf_{f \in \mathcal{H}_K} \{ \|f_\rho - f\|_\rho + s \|f\|_K \}, \quad s > 0. \quad (2.4)$$

The K -functional plays a central role in interpolation theory, see, for example, [4, Chapter 3]. Therefore, the first term on the right-hand side of (2.3) can be characterized by requiring that f_ρ lies in the interpolation space $(\mathcal{L}_{\rho_X}^2, \mathcal{H}_K)_{\beta, \infty}$ defined as

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

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 $f_\rho \in (\mathcal{L}_{\rho_X}^2, \mathcal{H}_K)_{\beta, \infty}$. Intuitively, we can regard the interpolation space $(\mathcal{L}_{\rho_X}^2, \mathcal{H}_K)_{\beta, \infty}$ as an intermediate space between the metric space $\mathcal{L}_{\rho_X}^2$ and the much smaller approximation space \mathcal{H}_K . Similar ideas of using the K -functional to characterize the approximation property of the space \mathcal{H}_K can be found in [10, 11, 22].

In general, we have that $\lim_{s \rightarrow 0^+} \mathcal{K}(s, f_\rho) = \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho$. 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 (2.2). Define $\{f_t : t \in \mathbb{N}_{T+1}\}$ 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*

$$\lim_{T \rightarrow \infty} \mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] = \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2. \quad (2.6)$$

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 (2.6) equals zero. To see this, we say that \mathcal{H}_K is dense in $\mathcal{L}_{\rho_X}^2$ if

$$\inf_{f \in \mathcal{H}_K} \|f - g\|_\rho = 0, \quad \text{for all } g \in \mathcal{L}_{\rho_X}^2. \quad (2.7)$$

We note that, in [25], 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 (2.7), since $\|f\|_\rho \leq \|f\|_\infty$ and $\mathcal{C}(X)$ is dense in $\mathcal{L}_{\rho_X}^2$. Note that $f_\rho \in \mathcal{L}_{\rho_X}^2$. 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 $\mathcal{L}_{\rho_X}^2$ then we have that*

$$\lim_{T \rightarrow \infty} \mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] = 0. \quad (2.8)$$

By choosing θ appropriately, we can get explicit error rates if the regression function f_ρ lies in a certain hypothesis space. In the following, we concentrate on the hypothesis space $L_K^\beta(\mathcal{L}_{\rho_X}^2)$ (see below) since this facilitates a comparison of our results with those in the related literature [6, 24, 26, 31, 32, 33]. One can similarly derive error rates from (2.3) when f_ρ belongs to the interpolation space (2.5).

To define the space $L_K^\beta(\mathcal{L}_{\rho_X}^2)$, we introduce the integral operator $L_K : \mathcal{L}_{\rho_X}^2 \rightarrow \mathcal{L}_{\rho_X}^2$ defined as

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

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

$$L_K^\beta(\mathcal{L}_{\rho_X}^2) := \left\{ f = \sum_{j=1}^{\infty} \lambda_j^\beta a_j \phi_j : \|L_K^{-\beta} f\|_\rho := \sum_{j=1}^{\infty} a_j^2 < \infty \right\}, \quad (2.9)$$

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 $L_K^\beta(\mathcal{L}_{\rho_X}^2)$ will be. In particular, we know [24] that $L_K^\beta(\mathcal{L}_{\rho_X}^2) \subseteq \mathcal{H}_K$ for $\beta > 1/2$ and $L_K^{1/2}(\mathcal{L}_{\rho_X}^2) = \mathcal{H}_K$ with the norm satisfying

$$\|g\|_K = \|L_K^{-1/2} g\|_\rho, \quad \forall g \in \mathcal{H}_K. \quad (2.10)$$

One can find more details in [11, 24].

We now can state the following capacity independent rates. Hereafter, the expression $a_T = O(b_T)$ means that there exists a constant c such that $a_T \leq cb_T$ for all $T \in \mathbb{N}$.

Theorem 3. *Let $\theta \in (0, 1)$ and $\mu(\theta)$ be given by (2.2). Define $\{f_t : t \in \mathbb{N}_{T+1}\}$ by (1.5). If $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $0 < \beta \leq 1/2$ then, by selecting $\eta_t = \frac{1}{\mu(\frac{2\beta}{2\beta+1})} t^{-\frac{2\beta}{2\beta+1}}$ for $t \in \mathbb{N}$, there holds*

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] = O(T^{-\frac{2\beta}{2\beta+1}} \ln T). \quad (2.11)$$

In Section 6.1, we will derive the error rate (2.11) from some modified error bounds similar to (2.3). Since the best rate of the second term on the

right-hand side of (2.3) is $O(T^{-1/2} \ln T)$, the consequent error rate (2.11) is never faster than the rate $O(T^{-1/2} \ln T)$ achieved at $\beta = 1/2$. This means, under the hypothesis that $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta > 0$, that we can not expect better rates outside the range $\beta \in (0, 1/2]$. Hence, without loss of generality we only consider the case $\beta \in (0, 1/2]$ in Theorem 3.

We now turn our attention to the case that the step sizes of (1.5) are in the form of $\{\eta_t = \eta(T) : t \in \mathbb{N}_T\}$, that is, they equal to a constant depending on the sample number. In this scenario, the error bounds for the online 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 defined by (1.5). If $16(\kappa + 1)^4 \ln(8T)\eta \leq 1$ then $\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2]$ is bounded by*

$$\inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho + 2(1 + \kappa)(\eta T)^{-1/2} \|f\|_K \right\}^2 + c_\rho \eta \ln(8T). \quad (2.12)$$

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 $|f_{T+1}(x_0) - f_\rho(x_0)| \leq \kappa \|f_{T+1} - f_\rho\|_K$. 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 [24] 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.

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 $\mathcal{L}_{\rho_X}^2$ as well as in \mathcal{H}_K .

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

(1) *If the step size satisfies*

$$\lim_{T \rightarrow \infty} T\eta(T) = \infty, \quad \lim_{T \rightarrow \infty} \eta(T) \ln T = 0 \quad (2.13)$$

then there holds

$$\lim_{T \rightarrow \infty} \mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] = \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2. \quad (2.14)$$

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

$$\lim_{T \rightarrow \infty} T\eta(T) = \infty, \quad \lim_{T \rightarrow \infty} \eta^2(T)T = 0 \quad (2.15)$$

then we have that

$$\lim_{T \rightarrow \infty} \mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_K^2] = 0. \quad (2.16)$$

Note that, if the step size decays in the form of $\eta(T) = O(T^{-\theta})$, $T \rightarrow \infty$ with $\theta > 0$ then the hypothesis (2.13) allows the choice $\theta \in (0, 1)$ while (2.15) requires $\theta \in (1/2, 1)$.

We will prove Theorems 4 and 5 in Section 5. In Section 6.1, we shall establish the following error rates in $\mathcal{L}_{\rho_X}^2$ 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 defined by (1.5). If $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ 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*

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] = O\left(T^{-\frac{2\beta}{2\beta+1}} \ln T\right). \quad (2.17)$$

Moreover, if $\beta > 1/2$ then there holds

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_K^2] = O\left(T^{-\frac{2\beta-1}{2\beta+1}}\right). \quad (2.18)$$

As the last contribution of this paper, we further improve the preceding error rate [32] 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 (2.2). Define $\{f_t : t \in \mathbb{N}_{T+1}\}$ by (1.4). Assume that $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $0 < \beta \leq 1$. 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*

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] = O\left(T^{-\frac{2\beta}{2\beta+1} + \varepsilon}\right). \quad (2.19)$$

The preceding error rates of the online regularized algorithm (1.4) with $\lambda > 0$ were established in [32] under the assumption of $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with

some $\beta \in (0, 1]$. Namely, for any arbitrarily small $\varepsilon > 0$, by choosing $\lambda := \lambda(T)$ appropriately, there holds

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_\rho] = O\left(T^{-\frac{\beta}{2(\beta+1)} + \varepsilon}\right) \quad (2.20)$$

and, for $1/2 < \beta \leq 1$, we further have that

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{T+1} - f_\rho\|_K] = O\left(T^{-\frac{2\beta-1}{4\beta+2} + \varepsilon}\right). \quad (2.21)$$

By Cauchy-Schwarz inequality, we see from (2.19) that, for any arbitrarily small $\varepsilon > 0$, there holds

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

Therefore, the rate (2.19) is much better than (2.20).

We remark that the error rates obtained here for the online algorithm (1.5) are capacity (kernel) independent except the prior requirement that $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $\beta > 0$. It is an open problem to improve error bounds when some additional information is known such as the regularity of K or some polynomial decay of the eigenvalues of L_K [6, 11, 30]. In addition, it is noteworthy that all the rates are proved with respect to expectation norm and we do not know how to convert them into similar probabilistic bounds as [24, 31].

2.2 Comparisons and discussions

There is a large body of literature on online learning algorithms. Let us discuss some work relating to this paper.

The mistake (regret) bounds for the cumulative loss $1/T \sum_{t=1}^T (y_t - f_t(x_t))^2$ for general online algorithms have been well studied in the literature. See, for example, [2, 7, 8, 9, 16, 17, 18, 28, 35] and references therein. Specifically, mistake bounds were derived for the online density estimation in [2]. Section 6 in [16] showed, for a learning algorithm different from (1.5), the upper bounds for the relative expected instantaneous loss, measuring the predicting ability of the last output in the linear regression problem. The online algorithm studied in [9] 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 [18], 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 [28, Section 5]. There, generalization bounds for the average prediction were also derived from the cumulative loss bounds in reproducing kernel Hilbert spaces. More precisely, for each $t \in \mathbb{N}$, let $H_t : X \rightarrow \mathbb{R}$ 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 $\bar{H}_{T+1} = \frac{1}{T+1} \sum_{t=1}^{T+1} H_t$ and let $D \in \mathcal{H}_K$ with $D(x) \in Y := [-M, M]$ for all $x \in X$. It was shown [28, 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

$$\mathcal{E}(\bar{H}_{T+1}) \leq \mathcal{E}(D) + \frac{2M}{\sqrt{T+1}} \left(\sqrt{\kappa^2 + M^2} (\|D\|_K + 1) + 2M \sqrt{2 \ln \frac{2}{\delta}} \right).$$

The main difference of our generalization bounds (2.3) and (2.12) from the above one is that we have no assumption on the functions in the given upper bounds.

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 its last output f_{T+1} . In [21], the authors studied the performance of f_{T+1} in the \mathcal{H}_K norm where 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 [19, 32]. Using quite different methods from ours, [35] obtained similar generalization bounds as (2.12) for the online gradient descent algorithm associated with uniformly Lipschitz loss functions, linear kernels and constant step sizes (learning rates).

In this paper, we study the least-square online algorithm (1.5) without regularization terms in \mathcal{H}_K and conclude that it performs competitively in contrast to the commonly used least-square learning schemes. To explain this, in what follows we compare our capacity independent rates for the online algorithm (1.5) with the state-of-art ones for other least-square learning algorithms under the same hypothesis that $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta > 0$. In particular, it can be shown that our rates are almost optimal in a certain sense.

First, let us begin with the comparison with the rates for the online

regularized algorithms. In this discussion, we note that the rates (2.11), (2.17) and (2.18) of the online algorithm (1.5) are comparable to the corresponding rates (2.19), (2.21) of the online regularized algorithm (1.4) with $\lambda > 0$.

Second, under the same assumption that $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (0, 1/2]$, we present another 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 $t \in \mathbb{N}_T$. In addition, 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 [35, Theorem 5.2] for the average output \hat{f}_{T+1} in the kernel setting can be cast as

$$\mathbb{E}_{Z^T}[\mathcal{E}(\hat{f}_{T+1})] \leq \inf_{f \in \mathcal{H}_K} \left\{ \left(1 + \frac{2\kappa^2\sigma_{T+1}^2}{r_{T+1}}\right) \mathcal{E}(f) + \frac{1}{2r_{T+1}} \|f\|_K^2 \right\}, \quad (2.22)$$

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 $f \in \mathcal{L}_{\rho_X}^2$. If we denote the *regularization error* as

$$\mathcal{D}(\lambda) := \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\} \quad (2.23)$$

then the bound (2.22) tells us that $\mathbb{E}_{Z^T}[\|\hat{f}_{T+1} - f_\rho\|_\rho^2]$ is bounded by

$$\left(1 + \frac{2\kappa^2\sigma_{T+1}^2}{r_{T+1}}\right) \mathcal{D}\left(\frac{1}{2(r_{T+1} + 2\kappa^2\sigma_{T+1}^2)}\right) + \frac{2\kappa^2\sigma_{T+1}^2}{r_{T+1}} \mathcal{E}(f_\rho). \quad (2.24)$$

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 $\theta \neq 1/2$ and $\frac{\sigma_{T+1}^2}{r_{T+1}} = O(T^{-1/2} \ln T)$ for $\theta = 1/2$. Furthermore, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $\beta \in (0, 1/2]$ then we know from [24, Lemma 3] that

$$\mathcal{D}(\lambda) \leq \lambda^{2\beta} \|L_K^{-\beta} f_\rho\|_\rho^2, \quad (2.25)$$

which implies that

$$\mathcal{D}\left(\frac{1}{2(r_{T+1} + 2\kappa^2\sigma_{T+1}^2)}\right) = O(T^{-2\beta(1-\theta)}). \quad (2.26)$$

Therefore, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (0, 1/2]$, 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 (2.26) into (2.24) yields that

$$\mathbb{E}_{Z^T} [\|\hat{f}_{T+1} - f_\rho\|_\rho^2] = \begin{cases} O(T^{-1/2} \ln T) & \text{for } \beta = 1/2, \\ O(T^{-\frac{2\beta}{2\beta+1}}) & \text{for } \beta \in (0, 1/2). \end{cases} \quad (2.27)$$

Similarly, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (0, 1/2]$ 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 (2.24) and (2.25) that

$$\mathbb{E}_{Z^T} [\|\hat{f}_{T+1} - f_\rho\|_\rho^2] = O(T^{-\frac{2\beta}{2\beta+1}}). \quad (2.28)$$

We conclude that the rates (2.11) and (2.17) for the online algorithm (1.5) are almost the same as the corresponding rates (2.27) and (2.28) for the averaged stochastic gradient descent algorithm.

Third, we compare our rates with the offline regularization algorithm which is defined by

$$f_{\mathbf{z}, \lambda} := \arg \inf_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t \in \mathbb{N}_T} (f(x_t) - y_t)^2 + \lambda \|f\|_K^2 \right\}. \quad (2.29)$$

This offline algorithm is often referred to as a Tikhonov regularization scheme for learning, see, for example, [15].

The capacity independent generalization bounds in [33, 34] can be expressed as

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

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

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho^2] \leq \mathcal{D}(\lambda) + (\mathcal{E}(f_\rho) + \mathcal{D}(\lambda)) \left\{ \frac{4\kappa^2}{T\lambda} + \left(\frac{2\kappa^2}{T\lambda} \right)^2 \right\}. \quad (2.30)$$

But, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (0, 1/2]$ we know from (2.25) that $\mathcal{D}(\lambda) \leq \lambda^{2\beta} \|L_K^{-\beta} f_\rho\|_\rho^2$. Putting this into (2.30) and trading off T and λ , for $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $\beta \in (0, 1/2]$, the choice $\lambda = T^{-\frac{1}{2\beta+1}}$ gives the rate

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho^2] = O\left(T^{-\frac{2\beta}{2\beta+1}}\right). \quad (2.31)$$

When $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (1/2, 1]$, it was improved in [24], by choosing $\lambda = \lambda(T)$ appropriately, that¹

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{\mathbf{z}, \lambda} - f_\rho\|_\rho^2] = O(T^{-\frac{2\beta}{2\beta+1}}) \quad (2.32)$$

and

$$\mathbb{E}_{\mathbf{z} \in Z^T} [\|f_{\mathbf{z}, \lambda} - f_\rho\|_K^2] = O(T^{-\frac{2\beta-1}{2\beta+1}}). \quad (2.33)$$

Under the same assumptions, the rates (2.11), (2.17) and (2.18) of the online algorithm (1.5) are almost the same as the corresponding offline rates (2.31), (2.32) and (2.33).

We remark that the generalization bounds associated with $\mathcal{D}(\lambda)$ such as (2.30) for the regularization algorithms are neat and interesting, but it tends to suffer a saturation phenomenon with restricted β , since the regularization error $\mathcal{D}(\lambda)$ decays at most linearly as $\lambda \rightarrow 0+$.

To see why this is true, by (2.1), we know that $\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \} \geq \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \frac{\|f\|_\rho^2}{\kappa^2} \right\}$. Since $\|f - f_\rho\|_\rho \geq \| \|f\|_\rho - \|f_\rho\|_\rho \|$, by letting $t = \|f\|_\rho$, we have that

$$\mathcal{D}(\lambda) \geq \inf_{t \in \mathbb{R}} \left\{ (t - \|f_\rho\|_\rho)^2 + \frac{\lambda t^2}{\kappa^2} \right\} = \frac{\lambda}{\kappa^2 + \lambda} \|f_\rho\|_\rho^2. \quad (2.34)$$

If f_ρ is not identically zero (i.e., $\|f_\rho\|_\rho > 0$), the inequality (2.34) implies that the optimal decay of $\mathcal{D}(\lambda)$ is $O(\lambda)$, $\lambda \rightarrow 0+$ which, by (2.25), is achieved when $f_\rho \in L_K^{1/2}(\mathcal{L}_{\rho_X}^2) = \mathcal{H}_K$. Equivalently speaking, the rate of $\mathcal{D}(\lambda)$ is never faster than $O(\lambda)$, $\lambda \rightarrow 0+$ even if f_ρ lies in any smaller space $L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta > 1/2$ than \mathcal{H}_K . Therefore, in this case the consequent trade-off rate (2.31) derived from (2.30) is at most $O(T^{-1/2})$, $T \rightarrow \infty$, which is achieved at $\beta = 1/2$ and can not be improved beyond the range $\beta \in (0, 1/2]$. This is the so-called saturation phenomenon in the context of inverse problems [14].

For similar reasons, the rates (2.27) and (2.28) derived from (2.24) for the averaged stochastic gradient descent algorithm have the same problem. In contrast, our analysis shows that the capacity independent rate (2.17) for the un-regularized online algorithm (1.5) does not suffer this drawback and is arbitrarily close to T^{-1} as $\beta \rightarrow \infty$.

Finally, we explain in two folds that the error rates for the online algorithm (1.5) are actually almost optimal in the capacity independent

¹The rates are originally stated in the form of probability inequalities. We employ their expectation versions for consistency with other bounds throughout this paper.

sense. We initially illustrate this by citing the lower bounds in [6] when $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (1/2, 1]$. Second, we demonstrate a specific example to contrast our rates with the best possible rate in the non-parametric statistical literature (see, for example, [5, 26]).

We begin with the citation of the lower bound in [6]. Consider $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (1/2, 1]$ and the eigenvalues (arranged in decreasing order) of L_K have the decay $\lambda_i = O(i^{-b})$ with $b > 1$. Then it was proved in [6] that the rate $T^{-\frac{2\beta b}{2\beta b+1}}$ in $\mathcal{L}_{\rho_X}^2$ is optimal, see [6] 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, for example, [11]. 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 $b \rightarrow 1$ of 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 [31] for a discussion). In this sense, our rates (2.17) and (2.19) are almost optimal for the capacity independent case under the condition that $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta \in (1/2, 1]$.

Now, we illustrate by a specific example that our rates for the on-line algorithm (1.5) are comparable to the optimal rate in non-parametric statistics. For simplicity, we only consider smooth splines [29] 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 $s > 1/2$ which is defined by Fourier coefficients

$$H^s[0, 1] := \left\{ f = a_0 + \sum_{k \in \mathbb{N}} \sqrt{2} k^{-s} (a_k \sin(2\pi kx) + b_k \cos(2\pi kx)) : \right. \\ \left. \|f\|_s^2 := a_0^2 + \sum_{k \in \mathbb{N}} (a_k^2 + b_k^2) < \infty \right\}. \quad (2.35)$$

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

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

In addition, $\lambda_1 = 1$, $\lambda_{2j} = \lambda_{2j+1} = j^{-2s}$ for $j \in \mathbb{N}$ and the orthonormal eigenvalues are $\phi_1(x) = 1$, $\phi_{2j}(x) = \sqrt{2} \sin(2\pi jx)$ and $\phi_{2j+1}(x) = \sqrt{2} \cos(2\pi jx)$ for $j \in \mathbb{N}$. Therefore, in this case, the range space $L_{K_s}^\beta(\mathcal{L}_{dx}^2)$ with $\beta > 0$ defined by (2.9) is identical to the Sobolev space $H^{2\beta s}[0, 1]$.

It is well known (for example, [26]) in non-parametric statistics that the rate $O(T^{-\frac{4\beta s}{4\beta s+1}})$ is optimal if $f_\rho \in L_{K_s}^\beta(\mathcal{L}_{dx}^2) = H^{2\beta s}[0, 1]$ with $\beta > 1/2$. Therefore, if we assume $d\rho_X = dx$, $f_\rho \in L_{K_s}^\beta(\mathcal{L}_{dx}^2)$ with $\beta > 1/2$ and consider the online algorithm (1.5) with $\mathcal{H}_{K_s} = H^s[0, 1]$ with $s > 1/2$ then our capacity independent rate $O(T^{-\frac{2\beta}{2\beta+1}} \ln T)$ such as (2.17) is suboptimal. But it is arbitrarily close to the best one $O(T^{-\frac{4\beta s}{4\beta s+1}})$ as $s \rightarrow 1/2+$.

3 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 $\mathcal{L}_{\rho_X}^2$. 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 also, as a byproduct, allow us to improve the preceding error rate in [32] for the algorithm (1.4) with $\lambda > 0$ as shown in Section 6.2 below.

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

$$f_\lambda = \arg \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) + \lambda \|f\|_K^2 \right\}. \quad (3.1)$$

With a slight abuse of notation, we indicate f_ρ by f_0 .

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

$$L_K(f_\lambda - f_\rho) + \lambda f_\lambda = 0 \quad (3.2)$$

and

$$\|f_\lambda - f_\rho\|_\rho \leq \|f_\rho\|_\rho. \quad (3.3)$$

Proof. We introduce the functional $\mathcal{Q} : \mathcal{H}_K \rightarrow \mathbb{R}$ defined as $\mathcal{Q}(f) := \mathcal{E}(f) + \lambda \|f\|_K^2$ for any $f \in \mathcal{H}_K$. We know that \mathcal{Q} is differentiable and strictly convex. Therefore, it has a unique minimizer which we have called f_λ . Moreover, f_λ is determined by the fact that the gradient of \mathcal{Q} at f_λ is zero. Indeed, it can be verified for any $f, g \in \mathcal{H}_K$ that

$$\begin{aligned} \lim_{h \rightarrow 0} \frac{\mathcal{Q}(f + hg) - \mathcal{Q}(f)}{h} &= \left\langle \int_Z (f(x) - y) K_x + \lambda f, g \right\rangle_K d\rho(z) \\ &= \left\langle \int_X (f(x) - f_\rho(x)) K_x d\rho_X(x) + \lambda f, g \right\rangle_K \\ &= \langle L_K(f - f_\rho) + \lambda f, g \rangle_K. \end{aligned}$$

This proves the equality (3.2).

For the inequality (3.3), we first recall the property [11] of the least-square loss:

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2, \quad \forall f \in \mathcal{L}_{\rho_X}^2. \quad (3.4)$$

Therefore, (3.1) is equivalent to

$$f_\lambda = \arg \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho^2 + \lambda \|f\|_K^2 \right\}. \quad (3.5)$$

By taking $f = 0$, the above definition of f_λ yields that $\|f_\lambda - f_\rho\|_\rho^2 \leq \|f_\rho\|_\rho^2$. This completes the proof. \square

With (3.2) at hand, we can interpret the online algorithm (1.4) with $\lambda \geq 0$ as the following useful form.

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*

$$f_{t+1} - f_\lambda = (I - \eta_t(L_K + \lambda I))(f_t - f_\lambda) + \eta_t \mathcal{B}(f_t, z_t) \quad (3.6)$$

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

$$\mathcal{B}(f_t, z_t) := L_K(f_t - f_\rho) + (y_t - f_t(x_t))K_{x_t}. \quad (3.7)$$

Proof. The equality (3.6) 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 (3.2) 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

$$\begin{aligned} f_{t+1} - f_\lambda &= f_t - f_\lambda - \eta_t(f_t(x_t) - y_t)K_{x_t} - \eta_t \lambda f_t \\ &= (I - \eta_t(L_K + \lambda I))(f_t - f_\lambda) + \eta_t L_K(f_t - f_\lambda) \\ &\quad - \eta_t \lambda f_\lambda + \eta_t(y_t - f_t(x_t))K_{x_t}. \end{aligned} \quad (3.8)$$

But the equality (3.2) tells us that $-\lambda f_\lambda = L_K(f_\lambda - f_\rho)$. Hence, putting this back into (3.8) and arranging it yield the desired equality (3.6). \square

With the help of the formula (3.6), we can describe our approach by three steps in which the first step is referred to as *error decomposition*.

3.1 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 (3.6) and noting that $f_1 = 0$, for any $t \in \mathbb{N}_T$ we have that

$$f_{t+1} - f_\lambda = -\omega_1^t(L_K + \lambda I)f_\lambda + \sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t(L_K + \lambda I) \mathcal{B}(f_j, z_j). \quad (3.9)$$

Applying (3.9) with $t = T$, we get the following error decomposition

$$f_{T+1} - f_\lambda = -\omega_1^T(L_K + \lambda I)f_\lambda + \sum_{t \in \mathbb{N}_T} \eta_t \omega_{t+1}^T(L_K + \lambda I) \mathcal{B}(f_t, z_t). \quad (3.10)$$

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, for example, [21, 31, 32]. One can find similar ideas used for different learning schemes, see, for example, [3, 11, 13, 23, 24, 33] and references therein.

We will use the error decomposition (3.10) to estimate the expectation of $\|f_{T+1} - f_\rho\|_\rho^2$. To this end, we introduce some useful notations. Let L be a linear operator from $\mathcal{L}_{\rho_X}^2$ to itself, we use $\|L\|$ to denote its operator norm, that is, $\|L\| := \sup_{\|f\|_\rho=1} \|L(f)\|_\rho$. In addition, we always denote the expectation $\mathbb{E}_{z_1, \dots, z_t}$ as \mathbb{E}_{Z^t} for $t \in \mathbb{N}_T$ and adopt the convention $\mathbb{E}_{Z^0}[\xi] = \xi$ for any random variable ξ .

Now we can present the upper bound for the expectation of the error $\|f_{T+1} - f_\rho\|_\rho^2$ which is the foundation of our novel approach introduced in this paper.

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

$$\begin{aligned} & \|f_\lambda - f_\rho\|_\rho^2 + \|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho^2 + 2\|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho \|f_\lambda - f_\rho\|_\rho \\ & + \kappa^2 \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K + \lambda I)L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)]. \end{aligned} \quad (3.11)$$

Proof. Since $f_{T+1} - f_\rho = f_{T+1} - f_\lambda + f_\lambda - f_\rho$, we know that

$$\begin{aligned} \mathbb{E}_{Z^T}[\|f_{T+1} - f_\rho\|_\rho^2] &= \mathbb{E}_{Z^T}[\|f_{T+1} - f_\lambda\|_\rho^2] + \|f_\lambda - f_\rho\|_\rho^2 \\ &+ 2\mathbb{E}_{Z^T}[\langle f_{T+1} - f_\lambda, f_\lambda - f_\rho \rangle_\rho]. \end{aligned} \quad (3.12)$$

Hence, to prove (3.11), we need to estimate the first term and the last term on the right-hand side of (3.12).

For the first term on the right-hand side of (3.12), we first use the equality (3.10) to get that

$$\begin{aligned} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\lambda\|_\rho^2] &= \|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho^2 \\ &+ \mathbb{E}_{Z^T} \left[\left\| \sum_{t \in \mathbb{N}_T} \eta_t \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t) \right\|_\rho^2 \right] \\ &- 2\mathbb{E}_{Z^T} \left[\left\langle \sum_{t \in \mathbb{N}_T} \eta_t \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t), \omega_1^T(L_K + \lambda I)f_\lambda \right\rangle_\rho \right]. \end{aligned} \quad (3.13)$$

Then, we estimate the last two terms on the right-hand side of (3.13) separately.

For the second term on the right-hand side of (3.13), we write it as

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

Observe that the data $\mathbf{z} = \{z_t : t \in \mathbb{N}_T\}$ is independently distributed according to ρ and f_t is only dependent on $\{z_1, \dots, z_{t-1}\}$, 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

$$\mathbb{E}_{z_t} [\mathcal{B}(f_t, z_t)] = 0, \quad \forall t \in \mathbb{N}_T. \quad (3.14)$$

Therefore, for $t > t'$ we have that

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

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

$$\begin{aligned} &\mathbb{E}_{Z^T} \left[\left\| \sum_{t \in \mathbb{N}_T} \eta_t \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t) \right\|_\rho^2 \right] \\ &= \sum_{t \in \mathbb{N}_T} \eta_t^2 \mathbb{E}_{Z^t} \left[\left\| \omega_{t+1}^T(L_K + \lambda I)\mathcal{B}(f_t, z_t) \right\|_\rho^2 \right] \\ &\leq \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K + \lambda I)L_K^{1/2}\|^2 \mathbb{E}_{Z^t} \left[\left\| L_K^{-1/2}\mathcal{B}(f_t, z_t) \right\|_\rho^2 \right] \\ &= \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K + \lambda I)L_K^{1/2}\|^2 \mathbb{E}_{Z^t} \left[\left\| \mathcal{B}(f_t, z_t) \right\|_K^2 \right] \end{aligned} \quad (3.15)$$

where we have used the fact (2.10) in the last equality.

To estimate $\mathbb{E}_{Z^t} [\|\mathcal{B}(f_t, z_t)\|_K^2]$, note that the equality (3.14) implies that

$$\mathbb{E}_{z_t} [(f_t(x_t) - y_t)K_{x_t}] = L_K(f_t - f_\rho).$$

Hence, we can rewrite $\mathbb{E}_{z_t} [\|\mathcal{B}(f_t, z_t)\|_K^2]$ as

$$\mathbb{E}_{z_t} [\|(f_t(x_t) - y_t)K_{x_t}\|_K^2] - \|\mathbb{E}_{z_t} [(f_t(x_t) - y_t)K_{x_t}]\|_K^2. \quad (3.16)$$

Also, for any $x, x' \in X$ there holds

$$K(x, x') = \langle K_x, K_{x'} \rangle_K \leq \|K_x\|_K \|K_{x'}\|_K \leq \kappa^2. \quad (3.17)$$

Therefore,

$$\begin{aligned} \mathbb{E}_{Z^t} [\|\mathcal{B}(f_t, z_t)\|_K^2] &= \mathbb{E}_{z_1, \dots, z_{t-1}} \mathbb{E}_{z_t} [\|\mathcal{B}(f_t, z_t)\|_K^2] \\ &= \mathbb{E}_{Z^{t-1}} \mathbb{E}_{z_t} [\|(f_t(x_t) - y_t)K_{x_t}(\cdot)\|_K^2] \\ &\quad - \mathbb{E}_{Z^{t-1}} \|\mathbb{E}_{z_t} [(f_t(x_t) - y_t)K_{x_t}(\cdot)]\|_K^2 \\ &\leq \mathbb{E}_{Z^{t-1}} \mathbb{E}_{z_t} [\|(f_t(x_t) - y_t)K_{x_t}(\cdot)\|_K^2] \\ &\leq \kappa^2 \mathbb{E}_{Z^{t-1}} [\mathbb{E}_{z_t} [(f_t(x_t) - y_t)^2]] = \kappa^2 \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)]. \end{aligned} \quad (3.18)$$

Putting this into (3.15), we have that

$$\begin{aligned} \mathbb{E}_{Z^T} [\|\sum_{t \in \mathbb{N}_T} \eta_t \omega_{t+1}^T (L_K + \lambda I) \mathcal{B}(f_t, z_t)\|_\rho^2] \\ \leq \kappa^2 \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T (L_K + \lambda I) L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)]. \end{aligned} \quad (3.19)$$

For the last term on the right-hand side of (3.13), we use (3.14) to get that

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

Substituting this and (3.19) into (3.13) yields the following estimation for the first term on the right-hand side of (3.12)

$$\begin{aligned} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\lambda\|_\rho^2] &\leq \|\omega_1^T (L_K + \lambda I) f_\lambda\|_\rho^2 \\ &\quad + \kappa^2 \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T (L_K + \lambda I) L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)]. \end{aligned} \quad (3.20)$$

Now, it remains to estimate the last term on the right-hand side of (3.12). To do this, we apply (3.14) to the equality (3.10) and know that $\mathbb{E}_{Z^T}[f_{T+1} - f_\lambda] = -\omega_1^T(L_K + \lambda I)f_\lambda$. Therefore,

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

Putting this and (3.20) back into (3.12) yields the upper bound (3.11). \square

Now the error decomposition and Proposition 1 help us reduce the goal of our error analysis to the estimation of the four terms in (3.11). The first three terms of (3.11) are frequently referred to as the *approximation error* [23, 24]. The second term of (3.11) is called the *sample error*. We present their basic estimates separately in the following two subsections which constitute the second and last step of our approach.

3.2 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.

Since $f_0 = f_\rho$, to estimate the term $\|f_\lambda - f_\rho\|_\rho$, we only need to consider the case $\lambda > 0$. Recall Lemma 3 in [24].

Lemma 3. *Let $\lambda > 0$ and f_λ be defined by (3.1). If $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $0 < \beta \leq 1$ then there holds*

$$\|f_\lambda - f_\rho\|_\rho \leq \lambda^\beta \|L_K^{-\beta} f_\rho\|_\rho. \quad (3.21)$$

Moreover, if $1/2 < \beta \leq 1$ there holds

$$\|f_\lambda - f_\rho\|_K \leq \lambda^{\beta-1/2} \|L_K^{-\beta} f_\rho\|_\rho. \quad (3.22)$$

To bound the quantity $\|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho$ in (3.11), we need the following technical lemma which forms the essential estimates of our approach.

Lemma 4. *Let $\lambda \geq 0, \beta > 0$ and $\eta_t(\kappa^2 + \lambda) \leq 1$ for any integer $t \in [j, k]$. Then there holds*

$$\|\omega_j^k(L_K + \lambda I)L_K^\beta\|^2 \leq \frac{2((\beta/e)^\beta + \kappa^{2\beta})^2}{\exp\{\lambda \sum_{t=j}^k \eta_t\}((\sum_{t=j}^k \eta_t)^{2\beta} + 1)}. \quad (3.23)$$

Proof. By the definition of L_K and Schwarz inequality, for any $f \in \mathcal{L}_{\rho_X}^2$ there holds

$$\begin{aligned}\|L_K f\|_{\rho}^2 &= \int_X \left| \int_X K(x, x') f(x') d\rho_X(x') \right|^2 d\rho_X(x) \\ &\leq \int_X \left(\int_X |K(x, x')|^2 d\rho_X(x') \int_X |f(x')|^2 d\rho_X(x') \right) d\rho_X(x) \\ &= \|f\|_{\rho}^2 \int_X \int_X |K(x, x')|^2 d\rho_X(x') d\rho_X(x) \leq \kappa^4 \|f\|_{\rho}^2\end{aligned}\quad (3.24)$$

where we used (3.17) and the fact $\rho_X(X) = 1$. 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

$$\|L_K^{\beta}\| := \sup_{\ell \in \mathbb{N}} \lambda_{\ell}^{\beta} \leq \kappa^{2\beta}.\quad (3.25)$$

Moreover, for any $x \leq \kappa^2 + \lambda$, there holds

$$\|I - x(L_K + \lambda I)\| := \sup_{\ell \in \mathbb{N}} (1 - x(\lambda_{\ell} + \lambda)) \leq 1 - x\lambda.\quad (3.26)$$

Applying (3.26) with $x = \eta_t$ iteratively for $t \in [j, k]$ and (3.25), we have that

$$\begin{aligned}\|\omega_j^k(L_K + \lambda I)L_K^{\beta}\| &\leq \prod_{t=j}^k \|I - \eta_t(L_K + \lambda I)\| \|L_K^{\beta}\| \\ &\leq \kappa^{2\beta} \prod_{t=j}^k (1 - \eta_t \lambda) \leq \exp \left\{ -\lambda \sum_{t=j}^k \eta_t \right\} \kappa^{2\beta}\end{aligned}\quad (3.27)$$

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,

$$\begin{aligned}\|\omega_j^k(L_K + \lambda I)L_K^{\beta}\| &:= \sup_{\ell \in \mathbb{N}} \prod_{t=j}^k (1 - \eta_t(\lambda_{\ell} + \lambda)) \lambda_{\ell}^{\beta} \\ &\leq \sup_{\ell \in \mathbb{N}} \exp \left\{ -(\lambda + \lambda_{\ell}) \sum_{t=j}^k \eta_t \right\} \lambda_{\ell}^{\beta} \\ &\leq \exp \left\{ -\lambda \sum_{t=j}^k \eta_t \right\} \sup_{x \geq 0} \exp \left\{ -x \sum_{t=j}^k \eta_t \right\} x^{\beta} \\ &= \exp \left\{ -\lambda \sum_{t=j}^k \eta_t \right\} \left(\frac{\beta}{e} \right)^{\beta} \left(\sum_{t=j}^k \eta_t \right)^{-\beta}.\end{aligned}\quad (3.28)$$

Consequently, it follows from (3.27) and (3.28) that $\|\omega_t^T(L_K + \lambda I)L_K^\beta\|^2$ is bounded by

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

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

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) For $\lambda = 0$, $\|\omega_1^T(L_K)f_\rho\|_\rho$ is bounded by

$$\inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho + 2(1 + \kappa) \left(\sum_{t \in \mathbb{N}_T} \eta_t \right)^{-1/2} \|f\|_K \right\}. \quad (3.29)$$

(b) For $\lambda > 0$, there holds

$$\|\omega_1^T(L_K + \lambda I)f_\lambda\|_\rho \leq 2 \exp\left\{-\lambda \sum_{t \in \mathbb{N}_T} \eta_t\right\} \|f_\rho\|_\rho. \quad (3.30)$$

Proof. We first prove (3.29) in property (a). Since $\eta_t \kappa^2 \leq 1$ for $t \in \mathbb{N}_T$, using (3.26) 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

$$\begin{aligned} \|\omega_1^T(L_K)f_\rho\|_\rho &\leq \|\omega_1^T(L_K)(f - f_\rho)\|_\rho + \|\omega_1^T(L_K)f\|_\rho \\ &\leq \|f - f_\rho\|_\rho + \|\omega_1^T(L_K)L_K^{1/2}\| \|L_K^{-1/2}f\|_\rho \\ &= \|f - f_\rho\|_\rho + \|\omega_1^T(L_K)L_K^{1/2}\| \|f\|_K \end{aligned} \quad (3.31)$$

where (2.10) is used in the last equality. Applying (3.23) with $\lambda = 0, \beta = 1/2, j = 1$, and $k = T$ yields the inequality $\|\omega_1^T(L_K)L_K^{1/2}\| \leq 2(1 + \kappa) \left(\sum_{t \in \mathbb{N}_T} \eta_t \right)^{-1/2}$. Substituting this into the right-hand side of (3.31), because $f \in \mathcal{H}_K$ is arbitrary, yields that

$$\|\omega_1^T(L_K)f_\rho\|_\rho \leq \inf_{f \in \mathcal{H}_K} \left\{ \|f - f_\rho\|_\rho + 2(1 + \kappa) \left(\sum_{t \in \mathbb{N}_T} \eta_t \right)^{-1/2} \|f\|_K \right\}. \quad (3.32)$$

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 (3.26) again with $x = \eta_t$ for $t \in \mathbb{N}_T$ implies that

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

But, by (3.3), $\|f_\lambda\|_\rho \leq 2\|f_\rho\|_\rho$. This finishes the inequality (3.30). \square

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

3.3 Estimates for the sample error

We now move on to the estimate of the last term of the bound (3.11): the sample error. Later we adopt the convention $\sum_{j=t+1}^t \eta_j = 0$ for any $t \in \mathbb{N}$.

Lemma 6. *Let $\lambda \geq 0$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be defined by (1.4). If the step sizes satisfy $\eta_t(\kappa^2 + \lambda) \leq 1$ for any $t \in \mathbb{N}_T$ then there holds*

$$\begin{aligned} &\sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K + \lambda I)L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)] \\ &\leq 2(1 + \kappa)^2 \left(\sup_{t \in \mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)] \right) \sum_{t \in \mathbb{N}_T} \frac{\eta_t^2 \exp\left\{-\lambda \sum_{j=t+1}^T \eta_j\right\}}{\sum_{j=t+1}^T \eta_j + 1}. \end{aligned} \quad (3.33)$$

Proof. Applying (3.23) with $\beta = 1/2$, $k = T$, and $j = t + 1$ implies that

$$\|\omega_{t+1}^T(L_K + \lambda I)L_K^{1/2}\|^2 \leq \frac{2(1 + \kappa)^2}{\exp\left\{\lambda \sum_{j=t+1}^T \eta_j\right\} (\sum_{j=t+1}^T \eta_j + 1)}$$

where the convention $\sum_{t=T+1}^T \eta_t = 0$ is used. This immediately yields the inequality (3.33). \square

We would expect that the term on the right-hand side of the inequality (3.33) tends to zero as $T \rightarrow \infty$ under some conditions on the step sizes. Below, we roughly explain how to realize this expectation.

First, we uniformly bound the leaning sequence, for example,

$$\sup_{t \in \mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)] \leq 20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho). \quad (3.34)$$

This uniform bound is intuitively reasonable since we expect that the learning sequence $\{f_t : t \in \mathbb{N}_{T+1}\}$ approximates the regression function f_ρ . Actually, in the next section we will show that the expected bound (3.34) holds true as long as we enforce the following condition on the step sizes:

$$\sum_{j \in \mathbb{N}_t} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^t \eta_k\right\} / \left(\sum_{k=j+1}^t \eta_k + 1\right) \leq \frac{1}{8(1+\kappa)^4}, \quad \forall t \in \mathbb{N}_T. \quad (3.35)$$

Once the learning sequence is uniformly bounded, the term on the right-hand side of the inequality (3.33) tends to zero as T tends to infinity as long as

$$\lim_{T \rightarrow \infty} \sum_{t \in \mathbb{N}_T} \eta_t^2 \exp\left\{-\lambda \sum_{j=t+1}^T \eta_j\right\} / \left(\sum_{j=t+1}^T \eta_j + 1\right) = 0. \quad (3.36)$$

In the next section, we will discuss how the hypothesis (3.35) implies the uniform bound (3.34). In particular, we will prove that the commonly used step sizes $\{\eta_t = O(t^{-\theta}) : t \in \mathbb{N}\}$ and $\{\eta_t = O(T^{-\theta}) : t \in \mathbb{N}_T\}$ with $\theta \in (0, 1)$ satisfy (3.35) as well as (3.36).

4 Learning sequence and step sizes

We first prove that the uniform bound (3.34) for the learning sequence $\{f_t : t \in \mathbb{N}_{T+1}\}$ holds true when the step sizes satisfy the condition (3.35).

Proposition 2. *Let $\lambda \geq 0$ and $\{f_t : t \in \mathbb{N}_{T+1}\}$ be defined as (1.4). Moreover, if the step sizes satisfy $\eta_t(\kappa^2 + \lambda) \leq 1$ for $t \in \mathbb{N}_T$ and the hypothesis (3.35) then we have that*

$$\sup_{t \in \mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}} [\|f_t - f_\lambda\|_\rho^2] \leq 9\|f_\rho\|_\rho^2 + \mathcal{E}(f_\rho) \quad (4.1)$$

and

$$\sup_{t \in \mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)] \leq 20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho). \quad (4.2)$$

Proof. First recall from (3.4) that

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|_\rho^2, \quad \forall f \in \mathcal{L}_{\rho_X}^2. \quad (4.3)$$

Also, by (3.3), we know that $\|f - f_\rho\|_\rho^2 \leq 2\|f - f_\lambda\|_\rho^2 + 2\|f_\lambda - f_\rho\|_\rho^2 \leq 2\|f - f_\lambda\|_\rho^2 + 2\|f_\rho\|_\rho^2$. Combining this with (4.3) yields that

$$\mathcal{E}(f) \leq \mathcal{E}(f_\rho) + 2\|f - f_\lambda\|_\rho^2 + 2\|f_\rho\|_\rho^2. \quad (4.4)$$

Hence, the second inequality (4.2) follows from the first estimate (4.1).

Now, it is sufficient to estimate the first inequality (4.1) by induction. Since $\|f_\lambda\|_\rho \leq 2\|f_\rho\|_\rho$ from (3.3) and $f_k = 0$ for $k = 1$, we have that $\|f_1 - f_\lambda\|_\rho^2 = \|f_\lambda\|_\rho^2 \leq 4\|f_\rho\|_\rho^2$. It means that

$$\mathbb{E}_{Z^{k-1}} [\|f_k - f_\lambda\|_\rho^2] \leq 9\|f_\rho\|_\rho^2 + \mathcal{E}(f_\rho)$$

holds true for $k = 1$. As the induction assumption, we suppose the above inequality holds true for $k \in \mathbb{N}_t$. To advance the induction, we need to estimate $\mathbb{E}_{Z^t} [\|f_{t+1} - f_\lambda\|_\rho^2]$.

To see that, the formula (3.9) tells us that

$$\begin{aligned} \|f_{t+1} - f_\lambda\|_\rho^2 &= \|\omega_1^t(L_K + \lambda I)f_\lambda\|_\rho^2 + \left\| \sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t(L_K + \lambda I)\mathcal{B}(f_j, z_j) \right\|_\rho^2 \\ &\quad - 2\langle \omega_1^t(L_K + \lambda I)f_\lambda, \sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t(L_K + \lambda I)\mathcal{B}(f_j, z_j) \rangle_\rho. \end{aligned} \quad (4.5)$$

We estimate the expectations of the three terms on the right-hand side of (4.5) as follows.

Since $\eta_t(\kappa^2 + \lambda) \leq 1$ for $t \in \mathbb{N}_T$, by (3.26) we know that $\|\omega_1^t(L_K + \lambda I)\| \leq \prod_{j \in \mathbb{N}_t} \|I - \eta_j(L_K + \lambda I)\| \leq 1$. Hence, combining this with (3.3) implies that

$$\|\omega_1^t(L_K + \lambda I)f_\lambda\|_\rho^2 \leq \|\omega_1^t(L_K + \lambda I)\|_\rho^2 \|f_\lambda\|_\rho^2 \leq \|f_\lambda\|_\rho^2 \leq 4\|f_\rho\|_\rho^2. \quad (4.6)$$

To estimate the expectation of the second term on the right-hand side of (4.5), we argue similarly as in the proof of the inequality (3.15) to get that

$$\begin{aligned} &\mathbb{E}_{Z^t} \left[\left\| \sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t(L_K + \lambda I)\mathcal{B}(f_j, z_j) \right\|_\rho^2 \right] \\ &\leq \sum_{j \in \mathbb{N}_t} \eta_j^2 \|\omega_{j+1}^t(L_K + \lambda I)L_K^{1/2}\|_\rho^2 \mathbb{E}_{Z^j} [\|\mathcal{B}(f_j, z_j)\|_\rho^2]. \end{aligned} \quad (4.7)$$

Now, applying (3.23) with $\beta = 1/2$ yields that

$$\|\omega_{j+1}^t(L_K + \lambda I)L_K^{1/2}\|_\rho^2 \leq 2(1 + \kappa)^2 \frac{\exp\{-\lambda \sum_{k=j+1}^t \eta_k\}}{\sum_{k=j+1}^t \eta_k + 1}. \quad (4.8)$$

Also, for any $k \in \mathbb{N}_t$, using (3.18), (4.4) with $f = f_k$, we see from the induction assumption that

$$\begin{aligned} \mathbb{E}_{Z^k} [\|\mathcal{B}(f_k, z_k)\|_K^2] &\leq \kappa^2 \mathbb{E}_{Z^{k-1}} [\mathcal{E}(f_k)] \\ &\leq \kappa^2 (\mathcal{E}(f_\rho) + 2\mathbb{E}_{k-1} [\|f_k - f_\lambda\|_\rho^2] + 2\mathcal{E}(f_\rho)) \\ &\leq \kappa^2 (20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)). \end{aligned} \quad (4.9)$$

Thus, putting (4.8) and (4.9) into (4.7), we know from the hypothesis (3.35) that

$$\begin{aligned} \mathbb{E}_{Z^t} [\|\sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t (L_K + \lambda I) (\mathcal{B}(f_j, z_j))\|_\rho^2] \\ \leq 2\kappa^2 (\kappa + 1)^2 (20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)) \sum_{j \in \mathbb{N}_t} \eta_j^2 \frac{\exp\{-\lambda \sum_{k=j+1}^t \eta_k\}}{\sum_{k=j+1}^t \eta_k + 1} \\ \leq (20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)) / 4. \end{aligned} \quad (4.10)$$

For the last term on the right-hand side of (4.5), using (3.14) and the property that f_j is independent of z_j , we have the following equality

$$\begin{aligned} \mathbb{E}_{Z^t} \langle \omega_1^t (L_K + \lambda I) f_\lambda, \sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t (L_K) \mathcal{B}(f_j, z_j) \rangle_\rho \\ = \langle \omega_1^t (L_K + \lambda I) f_\lambda, \sum_{j \in \mathbb{N}_t} \eta_j \omega_{j+1}^t (L_K + \lambda I) (\mathbb{E}_{Z^t} [\mathcal{B}(f_j, z_j)]) \rangle_\rho = 0. \end{aligned} \quad (4.11)$$

Putting this, (4.6), (4.7), (4.10), and (4.11) together, we get from (4.5) that $\mathbb{E}_{Z^t} [\|f_{t+1} - f_\lambda\|_\rho^2]$ is bounded by

$$4\|f_\rho\|_\rho^2 + (20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho)) / 4 = 9\|f_\rho\|_\rho^2 + \mathcal{E}(f_\rho)$$

which advances the induction and completes the proof. \square

According to the discussion at the end of the last section, the key of our approach is to verify the hypotheses (3.35) and (3.36). The following lemma ensures that they hold true for commonly used step sizes.

Lemma 7. *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 $\sum_{j \in \mathbb{N}_\ell} \eta_j^2 \exp\{-\lambda \sum_{k=j+1}^\ell \eta_k\} / (\sum_{k=j+1}^\ell \eta_k + 1)$ is bounded by*

$$\frac{c_\theta}{\mu} \left(\exp\{-\lambda d_\theta \ell^{1-\theta} / \mu\} \ell^{-\min\{\theta, 1-\theta\}} + \ell^{-\theta} \right) \ln\left(\frac{8\ell}{1-\theta}\right), \quad (4.12)$$

where

$$d_\theta = \frac{1 - 2^{\theta-1}}{1 - \theta} \quad \text{and} \quad c_\theta = \begin{cases} 1 & \text{if } \theta = 0, \\ 13 & \text{if } \theta = 1/2, \\ \frac{13}{(1-2^{\theta-1})|2\theta-1|} & \text{otherwise.} \end{cases}$$

When the step sizes have the form of $\{\eta_t = \frac{1}{\mu}t^{-\theta} : t \in \mathbb{N}_T\}$ with $\theta \in (0, 1)$ and $\mu \geq \max\{\lambda, 1\} + \kappa^2$, applying (4.12) with $\ell = T$ directly verifies the hypothesis (3.36). At the same time, the following straightforward corollary of Lemma 7 tells us that the hypothesis (3.35) is satisfied by choosing the constant μ appropriately.

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 $\mu \geq \kappa^2 + \lambda$. Then, the hypothesis (3.35) holds true if $\mu - \lambda$ is larger than the quantity*

$$\mu(\theta) := \begin{cases} 16(1 + \kappa)^4 \ln(8T) & \text{if } \theta = 0, \\ 1248(1 + \kappa)^4 & \text{if } \theta = 1/2, \\ \frac{208(1 + \kappa)^4}{(1-2^{\theta-1})|2\theta-1|} \left(\ln\left(\frac{8}{1-\theta}\right) + \frac{1}{\min\{\theta, 1-\theta\}} \right) & \text{otherwise.} \end{cases} \quad (4.13)$$

Proof. For any $\ell \in \mathbb{N}_T$ and $\lambda \geq 0$, the quality (4.12) is bounded by $\frac{2c_\theta}{\mu} \ell^{-\min\{\theta, 1-\theta\}} \ln\left(\frac{8\ell}{1-\theta}\right)$. Thus, the hypothesis (3.35) holds true as long as $\mu \geq \kappa^2 + \max\{\lambda, 1\}$ and

$$\frac{16(1 + \kappa)^4 c_\theta}{\mu} t^{-\min\{\theta, 1-\theta\}} \ln\left(\frac{8t}{1-\theta}\right) \leq 1, \quad \forall t \in \mathbb{N}_T.$$

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

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

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

and the definition of c_θ given in Lemma 7 yield the desired result. \square

5 Error bounds and convergence

In this section, we develop error bounds and convergence results for the online algorithm (1.5). In this case, $\lambda = 0$ and $f_0 = f_\rho$. Consequently, the error decomposition formula (3.10) can be interpreted as

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

and the prior error bound (3.11) given in Proposition 1 becomes

$$\begin{aligned} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] &\leq \|\omega_1^T(L_K)f_\rho\|_\rho^2 \\ &+ \kappa^2 \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K)L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)]. \end{aligned} \quad (5.2)$$

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.

Let us first establish the following useful lemma. One can find a modified form in [31].

Lemma 8. *If $f \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with some $\beta > 0$ then*

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

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

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

Proof. Applying Lemma 4 with $\lambda = 0, j = 1$, and $k = T$, the desired estimates (5.3) and (5.4) follow immediately from the following observations

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

and

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

This completes our lemma. \square

Recall the definitions of $\mu(\theta)$ in Corollary 2 and c_θ in Lemma 7, our error bounds read as follows.

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

$$\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \leq \|\omega_1^T(L_K)f_\rho\|_\rho^2 + \frac{c_\rho c_\theta}{\mu} T^{-\min\{\theta, 1-\theta\}} \ln\left(\frac{8T}{1-\theta}\right) \quad (5.5)$$

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

Proof. We use the bound (5.2) to obtain the inequality (5.5). First, note that

$$\begin{aligned} & \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K)L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)] \\ & \leq \left(\sup_{t \in \mathbb{N}_T} \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)] \right) \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K)L_K^{1/2}\|^2. \end{aligned} \quad (5.6)$$

Since $\eta_t = \frac{1}{\mu}t^{-\theta}$ with $\mu \geq \mu(\theta)$ for any $t \in \mathbb{N}_T$, by Corollary 2 we know that the hypothesis (3.35) on the step sizes holds true for any $t \in \mathbb{N}_T$. Hence, by Proposition 2, we know that

$$\sup_{t \in \mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}}[\mathcal{E}(f_t)] \leq 20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho). \quad (5.7)$$

Also, using (3.23) with $j = 1$ and $k = T$ for the case $\lambda = 0$ yields that

$$\sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K)L_K^{1/2}\|^2 \leq 2(1 + \kappa)^2 \sum_{t \in \mathbb{N}_T} \frac{\eta_t^2}{\sum_{j=t+1}^T \eta_j + 1}.$$

Therefore, applying (4.12) with $\lambda = 0$ and $\ell = T$ to the right-hand side of the above inequality implies that

$$\sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T(L_K)L_K^{1/2}\|^2 \leq \frac{4(1 + \kappa)^2 c_\theta}{\mu} T^{-\min\{\theta, 1-\theta\}} \ln\left(\frac{8T}{1-\theta}\right).$$

Putting the above estimate, (5.6) and (5.7) together, the desired bound (5.5) follows from (5.2). \square

To apply the error bound (5.5) to prove Theorem 1, we need to estimate the summation $\sum_{t \in \mathbb{N}_T} \eta_t$ which leads to the following lemma.

Lemma 9. *If $0 < \theta < 1$ then, for any $t \in \mathbb{N}_T$, we have that*

$$\sum_{j=t}^T j^{-\theta} \geq \frac{(T+1)^{1-\theta} - t^{1-\theta}}{1-\theta}. \quad (5.8)$$

Proof. The result is directly from the inequality $\sum_{j=t}^T j^{-\theta} \geq \int_t^{T+1} x^{-\theta} dx$. \square

We now establish Theorem 1 stated in Section 2.1.

Proof of Theorem 1. Recall that $\eta_t = \frac{1}{\mu} t^{-\theta}$ for $\theta \in (0, 1)$. Therefore, by (5.8) we have that $\sum_{t=1}^T \eta_t \geq \frac{1 - 2^{\theta-1}}{(1-\theta)\mu} T^{1-\theta}$. Putting this with the estimate (3.29) for $\|\omega_1^T(L_K)f_\rho\|_\rho$ together, the bound (2.3) immediately follows from the bound (5.5) for the case $\theta \in (0, 1)$. \square

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 $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta > 1/2$.

We now 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 10. *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*

$$\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_K^2] \leq \|\omega_1^T(L_K)f_\rho\|_K^2 + \kappa^2\eta^2 \sum_{t \in \mathbb{N}_T} \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)]. \quad (5.9)$$

Proof. The argument here is similar to the proof of Proposition 1.

From the error decomposition (5.1), we know that

$$\begin{aligned} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_K^2] &= \|\omega_1^T(L_K)f_\rho\|_K^2 \\ &\quad + \eta^2 \mathbb{E}_{Z^T} \left[\left\| \sum_{t \in \mathbb{N}_T} \omega_{t+1}^T(L_K)\mathcal{B}(f_t, z_t) \right\|_K^2 \right] \\ &\quad - 2\eta \sum_{t \in \mathbb{N}_T} \langle \omega_1^T(L_K)f_\rho, \mathbb{E}_{Z^T} [\omega_{t+1}^T(L_K)\mathcal{B}(f_t, z_t)] \rangle_K. \end{aligned} \quad (5.10)$$

To estimate the second term on right-hand side of the above inequality, we write its expectation part as

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

By the vanishing property (3.14) of $\mathcal{B}(f_t, z_t)$, for $t > t'$ we have that

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

Using the symmetry of t and t' , the above equality also holds true for $t' > t$. Consequently, combining this with the property (2.10) implies that

$$\begin{aligned} \mathbb{E}_{Z^T} \left[\left\| \sum_{t \in \mathbb{N}_T} \omega_{t+1}^T(L_K) \mathcal{B}(f_t, z_t) \right\|_K^2 \right] &= \sum_{t \in \mathbb{N}_T} \mathbb{E}_{Z^t} \left[\left\| \omega_{t+1}^T(L_K) L_K^{-1/2} \mathcal{B}(f_t, z_t) \right\|_\rho^2 \right] \\ &\leq \sum_{t \in \mathbb{N}_T} \left\| \omega_{t+1}^T(L_K) \right\|^2 \mathbb{E}_{Z^t} \left[\left\| L_K^{-1/2} \mathcal{B}(f_t, z_t) \right\|_\rho^2 \right] \\ &= \sum_{t \in \mathbb{N}_T} \left\| \omega_{t+1}^T(L_K) \right\|^2 \mathbb{E}_{Z^t} \left[\left\| \mathcal{B}(f_t, z_t) \right\|_K^2 \right]. \end{aligned} \quad (5.11)$$

To estimate $\mathbb{E}_{Z^t} \left[\left\| \mathcal{B}(f_t, z_t) \right\|_K^2 \right]$, we use (3.18) to get that

$$\mathbb{E}_{Z^t} \left[\left\| \mathcal{B}(f_t, z_t) \right\|_K^2 \right] \leq \kappa^2 \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)]. \quad (5.12)$$

Also, since $\eta \kappa^2 \leq 1$, the estimate (3.26) yields that, for any $t \in \mathbb{N}_T$ there holds $\left\| \omega_{t+1}^T(L_K) \right\| \leq 1$. Putting these estimates back into (5.11), we have that

$$\begin{aligned} \mathbb{E}_{Z^T} \left[\left\| \sum_{t \in \mathbb{N}_T} \omega_{t+1}^T(L_K) \mathcal{B}(f_t, z_t) \right\|_K^2 \right] &\leq \kappa^2 \sum_{t=1}^T \left\| \omega_{t+1}^T(L_K) \right\|^2 \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)] \\ &\leq \kappa^2 \sum_{t=1}^T \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)]. \end{aligned} \quad (5.13)$$

For the last term on the right-hand side of (5.10), we see from (3.14) that

$$\sum_{t \in \mathbb{N}_T} \langle \omega_1^T(L_K) f_\rho, \mathbb{E}_{Z^T} [\omega_{t+1}^T(L_K) \mathcal{B}(f_t, z_t)] \rangle_K = 0.$$

Cascading this equality, (5.10) and (5.13) yields the desired estimate. \square

We now present error bounds in $\mathcal{L}_{\rho_X}^2$ as well as in \mathcal{H}_K when the step sizes are in the form of $\{\eta_t = \eta : t \in \mathbb{N}_T\}$ 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 $16(\kappa + 1)^4 \ln(8T)\eta \leq 1$ then there holds

$$\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \leq \|\omega_1^T(L_K)f_\rho\|_\rho^2 + c_\rho \eta \ln(8T). \quad (5.14)$$

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

$$\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_K^2] \leq \|\omega_1^T(L_K)f_\rho\|_K^2 + c_\rho \eta^2 T. \quad (5.15)$$

Proof. We regard η as $\frac{1}{\mu}$. Then, the inequality $16(\kappa + 1)^4 \ln(8T)\eta \leq 1$ means that $\mu \geq 16(1 + \kappa)^4 \ln(8T)$. By Corollary 2, the hypothesis (3.35) on the step sizes is satisfied. Hence, the estimate (4.2) in Proposition 2 holds true.

The first estimate (5.14) follows from Proposition 3 for the case $\theta = 0$.

The bound (5.9) in Lemma 10 and (4.2) yield the estimate (5.15) in the \mathcal{H}_K norm. \square

From the above proposition, we can prove Theorem 4 stated in the Section 2.1.

Proof of Theorem 4. The error bound (2.12) is derived from (5.14) and the estimate (3.29) for $\|\omega_1^T(L_K)f_\rho\|_\rho$. \square

5.2 Proofs of convergence results

We are in a position to apply the general bounds (2.3) and (2.12) in Theorems 1 and 4 to prove Theorems 2 and 5 while the bounds (5.5) and (5.14) will be used to derive the explicit rates in the next section.

For the proof of Theorem 2, we need the following well-known property of the K -functional, see [4]. We include its proof for completeness.

Lemma 11. Let $s > 0$ and $\mathcal{K}(s, f_\rho)$ be defined by (2.4). Then there holds

$$\lim_{s \rightarrow 0^+} \mathcal{K}(s, f_\rho) = \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho. \quad (5.16)$$

Proof. By the definition (2.4) of \mathcal{K} , it is straightforward that

$$\underline{\lim}_{s \rightarrow 0^+} \mathcal{K}(s, f_\rho) \geq \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho.$$

Hence, we only need to prove

$$\overline{\lim}_{s \rightarrow 0+} \mathcal{K}(s, f_\rho) \leq \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho. \quad (5.17)$$

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 $\mathcal{K}(s, f_\rho) := \inf_{f \in \mathcal{H}_K} \{\|f - f_\rho\|_\rho + s\|f\|_K\}$ is bounded by

$$\varepsilon + \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho + s\|f_\varepsilon\|_K.$$

Letting $s \rightarrow 0+$ yields (5.17), and hence (5.16). \square

We now prove Theorem 2 by Lemma 11 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 (5.16) with $s = b_\theta \sqrt{\mu} T^{-(1-\theta)/2}$ to (2.3) yields that

$$\overline{\lim}_{T \rightarrow \infty} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \leq \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2.$$

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

$$\inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2 \leq \|f_{T+1} - f_\rho\|_\rho^2,$$

which leads to

$$\underline{\lim}_{T \rightarrow \infty} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \geq \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2.$$

This completes Theorem 2. \square

Turn to the proof of Theorem 5 which is involving convergence in \mathcal{H}_K , we further need the following observation.

Lemma 12. *Let $f_\rho \in \mathcal{H}_K$, the step sizes $\{\eta_t = \eta(T) : t \in \mathbb{N}_T\}$ satisfy $\eta(T)\kappa^2 \leq 1$ for any $T \in \mathbb{N}$ and $\lim_{T \rightarrow \infty} \eta(T)T = \infty$. Then we have that*

$$\lim_{T \rightarrow \infty} \|\omega_1^T(L_K)f_\rho\|_K = 0. \quad (5.18)$$

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

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

To prove (5.18), in what follows we construct functions in $L_K^1(\mathcal{L}_{\rho_X}^2)$ to approximate f_ρ arbitrarily in \mathcal{H}_K while for functions in $L_K^1(\mathcal{L}_{\rho_X}^2)$, we apply (5.4) in Lemma 8 with $\beta = 1$ to deal with it.

To this end, write $f_\rho = \sum_{\lambda_j > 0, j=1}^N \lambda_j^{1/2} a_j \phi_j + \sum_{\lambda_j > 0, j=N+1}^{\infty} \lambda_j^{1/2} a_j \phi_j$.

Then $f_N = \sum_{\lambda_j > 0, j=1}^N \lambda_j^1 (\lambda_j^{-1/2} a_j) \phi_j \in L_K^1(\mathcal{L}_{\rho_X}^2)$ for any $N \in \mathbb{N}$ because

$$\|L_K^{-1} f_N\|_\rho^2 = \sum_{\lambda_j > 0, j=1}^N (\lambda_j^{-1/2} a_j)^2 < \infty.$$

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

$$\sum_{j=N(\varepsilon)+1}^{\infty} a_j^2 \leq \varepsilon^2.$$

This implies that

$$\|f_\rho - f_{N(\varepsilon)}\|_K^2 = \left\| \sum_{j=N(\varepsilon)+1}^{\infty} \lambda_j^{1/2} a_j \phi_j \right\|_K^2 = \sum_{j=N(\varepsilon)+1}^{\infty} a_j^2 \leq \varepsilon^2. \quad (5.19)$$

Also, if we indicate the function $K(x, \cdot)$ by $K_x(\cdot)$ then the operator L_K can be reinterpreted as $\int_X K_x \otimes K_x d\rho_X(x) : \mathcal{H}_K \rightarrow \mathcal{H}_K$ defined as

$$\left(\int_X K_x \otimes K_x d\rho_X(x) \right) f := \int_X f(x) K_x d\rho_X(x), \quad \text{for } f \in \mathcal{H}_K. \quad (5.20)$$

By (2.10), we have that $\|(I - \eta(T)L_K)f\|_K = \|(I - \eta(T)L_K)L_K^{-1/2}f\|_\rho \leq \|I - \eta(T)L_K\| \|L_K^{-1/2}f\|_\rho = \|I - \eta(T)L_K\| \|f\|_K$. Also, since $\eta(T)\kappa^2 \leq 1$, applying (3.26) with $x = \eta$ implies that $\|(I - \eta(T)L_K)\| \leq 1$. Hence, for any $f \in \mathcal{H}_K$ there holds

$$\|(I - \eta(T)L_K)f\|_K \leq \|f\|_K.$$

Consequently, applying the above inequality with $f = f_\rho - f_{N(\varepsilon)}$ and (5.19) tells us that

$$\begin{aligned} \|(I - \eta(T)L_K)^T f_\rho\|_K &\leq \|(I - \eta(T)L_K)^T f_{N(\varepsilon)}\|_K \\ &\quad + \|(I - \eta(T)L_K)^T (f_\rho - f_{N(\varepsilon)})\|_K \\ &\leq \|(I - \eta(T)L_K)^T f_{N(\varepsilon)}\|_K + \|f_\rho - f_{N(\varepsilon)}\|_K \\ &\leq \|(I - \eta(T)L_K)^T f_{N(\varepsilon)}\|_K + \varepsilon. \end{aligned} \quad (5.21)$$

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

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

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

$$\|(I - \eta(T)L_K)^T f_{N(\varepsilon)}\|_K \leq \varepsilon \quad \forall T \geq T(\varepsilon). \quad (5.22)$$

Putting (5.21) and (5.22) together, we know that

$$\|\omega_1^T(L_K)f_\rho\|_K = \|(I - \eta(T)L_K)^T f_\rho\|_K \leq 2\varepsilon, \quad \forall T \geq T(\varepsilon).$$

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

We now move on to prove Theorem 5.

Proof of Theorem 5. Observe that we have assumed $\eta = \eta(T)$ depending on the sample number T .

Since either $\lim_{T \rightarrow \infty} \eta(T) \ln T = 0$ or $\lim_{T \rightarrow \infty} T\eta^2(T) = 0$, there exists $T_1(\varepsilon) \in \mathbb{N}$ such that $16(\kappa + 1)^4 \eta(T) \ln T \leq 1$ for all $T \geq T_1(\varepsilon)$. The hypotheses in Lemma 10 and Theorem 4 are satisfied for any $T \geq T_1(\varepsilon)$. Therefore, (2.12) holds true for any $T \geq T_1(\varepsilon)$.

We first prove the convergence in $\mathcal{L}_{\rho_X}^2$. Applying (5.16) with $s = 2(1 + \kappa)(\eta(T)T)^{-1/2}$ to (2.12), it follows from the hypothesis $\lim_{T \rightarrow \infty} T\eta(T) = \infty$ that

$$\overline{\lim}_{T \rightarrow \infty} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \leq \inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2.$$

On the other hand, since $f_{T+1} \in \mathcal{H}_K$, $\inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2 \leq \|f_{T+1} - f_\rho\|_\rho^2$ for any data $\mathbf{z} = \{z_t : t \in \mathbb{N}_T\}$ which leads to

$$\inf_{f \in \mathcal{H}_K} \|f - f_\rho\|_\rho^2 \leq \underline{\lim}_{T \rightarrow \infty} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2].$$

This verifies the first part of Theorem 5.

The proof for the second part is similar. Since $\lim_{T \rightarrow \infty} T\eta^2(T) = 0$, combining (5.15) in Proposition 4 with (5.18) in Lemma 12 yields that

$$\overline{\lim}_{T \rightarrow \infty} \mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_K^2] \leq 0 \quad (5.23)$$

which completes Theorem 5. \square

6 Explicit error rates

In this section we use the general bounds (5.5) and (5.14) given in Propositions 3 and 4 to derive explicit error rates for $\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2]$ by choosing the step sizes appropriately, when the regression function f_ρ lies in $L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta > 0$.

6.1 Rates for online learning without regularization

We first establish explicit error rates for the algorithm (1.5) stated in Theorems 3 and 6.

Proof of Theorem 3. We use Proposition 3 with $\theta \in (0, 1)$ to prove our theorem. Recall that $\mu(\theta)$ is defined by (4.13) for $0 < \theta < 1$. Therefore, if $\eta_t = \frac{t^{-\theta}}{\mu(\theta)}$ then, by (5.5), we know that

$$\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2] \leq \|\omega_1^T(L_K)f_\rho\|_\rho^2 + O\left(T^{-\min\{\theta, 1-\theta\}} \ln T\right). \quad (6.1)$$

Since $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$, applying (5.3) with $f = f_\rho$ implies that $\|\omega_1^T(L_K)f_\rho\|_\rho = O\left(\left(\sum_{t \in \mathbb{N}_T} \eta_t\right)^{-\beta}\right)$. Hence, by (5.8) we have that

$$\|\omega_1^T(L_K)f_\rho\|_\rho^2 = O\left(T^{-2(1-\theta)\beta}\right).$$

Note that

$$O\left(T^{-\min\{\theta, 1-\theta\}} \ln T\right) = O\left(T^{-\theta} \ln T + T^{-(1-\theta)} \ln T\right).$$

Putting these estimates into (6.1) and noting the hypothesis $\beta \in (0, 1/2]$, we know that

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

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

By Proposition 4, we can establish Theorem 6.

Proof of Theorem 6. In order to use Proposition 4, observe that $\ln(8T) \leq \frac{(8T)^\theta}{\theta} \leq \frac{8T^\theta}{\theta}$ for any $\theta \in (0, 1)$. Hence, we know that the choice of the step size $\eta := \frac{\beta}{64(1+\kappa)^4(2\beta+1)} T^{-\frac{2\beta}{2\beta+1}}$ for $\beta > 0$ satisfies the hypothesis $16(\kappa + 1)^4 \ln(8T)\eta \leq 1$ 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 (5.14) in Proposition 4 implies that

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

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 (2.17).

Similarly, if $f_\rho \in L_K^\beta(\mathcal{L}_{\rho_X}^2)$ with $\beta > 1/2$ then, applying (5.4) with $f = f_\rho$ to (5.15) yields that

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

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 (2.18). \square

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 $0 < \lambda \leq 1$ 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 (3.35). Consequently, by Proposition 2 we have that

$$\sup_{k \in \mathbb{N}_{T+1}} \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)] \leq 20\|f_\rho\|_\rho^2 + 3\mathcal{E}(f_\rho). \quad (6.2)$$

Also, observe that $\mu := \mu(\theta) + 1 \geq \kappa^2 + \lambda$ for $\lambda \in (0, 1]$. Therefore, applying Lemma 7 with $\ell = T$ tells us that $\sum_{t \in \mathbb{N}_T} \eta_t^2 \exp\left\{-\lambda \sum_{j=t+1}^T \eta_j\right\} / \left(\sum_{j=t+1}^T \eta_j + \right.$

1) is bounded by

$$\frac{c_\theta}{\mu} \left(\exp\{-\lambda d_\theta T^{1-\theta}/\mu\} T^{-\min\{\theta, 1-\theta\}} + T^{-\theta} \right) \ln\left(\frac{8T}{1-\theta}\right), \quad (6.3)$$

where $d_\theta = \frac{1-2^{\theta-1}}{1-\theta}$. Putting (6.2) and (6.3) back into the term on the right-hand side of (3.33) in Lemma 6, we have that

$$\begin{aligned} & \kappa^2 \sum_{t \in \mathbb{N}_T} \eta_t^2 \|\omega_{t+1}^T (L_K + \lambda I) L_K^{1/2}\|^2 \mathbb{E}_{Z^{t-1}} [\mathcal{E}(f_t)] \\ & \leq \frac{c_\rho c_\theta}{\mu} \left(\exp\{-\lambda d_\theta T^{1-\theta}/\mu\} T^{-\min\{\theta, 1-\theta\}} + T^{-\theta} \right) \ln\left(\frac{8T}{1-\theta}\right), \end{aligned}$$

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

Combining this with the bound (3.11) given by Proposition 1 for the case $\lambda \in (0, 1]$, $\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2]$ is bounded by

$$\begin{aligned} & \|f_\lambda - f_\rho\|_\rho^2 + \|\omega_1^T (L_K + \lambda I) f_\lambda\|_\rho^2 + 2\|f_\lambda - f_\rho\|_\rho \|\omega_1^T (L_K + \lambda I) f_\lambda\|_\rho \\ & + \frac{c_\rho c_\theta}{\mu} \left(\exp\{-\lambda d_\theta T^{1-\theta}/\mu\} T^{-\min\{\theta, 1-\theta\}} + T^{-\theta} \right) \ln\left(\frac{8T}{1-\theta}\right). \end{aligned} \quad (6.4)$$

Hence, it remains to estimate the first three terms of the above quantity involving $\|f_\lambda - f_\rho\|_\rho^2$ and $\|\omega_1^T (L_K + \lambda I) f_\lambda\|_\rho$.

To do that, we see from Lemma 3 that

$$\|f_\lambda - f_\rho\|_\rho \leq \lambda^\beta \|L_K^{-\beta} f_\rho\|_\rho, \quad \forall \beta \in (0, 1]. \quad (6.5)$$

Also, the estimate (3.30) implies that

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

This in connection with the inequality (5.8) with $t = 1$ implies that

$$\|\omega_1^T (L_K + \lambda I) f_\lambda\|_\rho \leq 2 \exp\{-\lambda d_\theta T^{1-\theta}/\mu\} \|f_\rho\|_\rho. \quad (6.6)$$

Substituting the estimates (6.5) and (6.6) into (6.4), it follows that $\mathbb{E}_{Z^T} [\|f_{T+1} - f_\rho\|_\rho^2]$ is bounded by

$$\begin{aligned} & \lambda^{2\beta} \|L_K^{-\beta} f_\rho\|_\rho^2 + \frac{c_\rho c_\theta}{\mu} T^{-\theta} \ln\left(\frac{8T}{1-\theta}\right) + 4\|L_K^{-\beta} f_\rho\|_\rho \|f_\rho\|_\rho \lambda^\beta \exp\{-\lambda d_\theta T^{1-\theta}/\mu\} \\ & + \left(\frac{c_\rho c_\theta}{\mu} + 4\|f_\rho\|_\rho^2 \right) \exp\{-\lambda d_\theta T^{1-\theta}/\mu\} \ln\left(\frac{8T}{1-\theta}\right). \end{aligned} \quad (6.7)$$

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

$$\exp\{-cT^\epsilon\} = O(T^{-s}). \quad (6.8)$$

Therefore, for any $0 < \epsilon < \frac{\beta}{2\beta+1}$, choosing $\lambda = T^{-\frac{1}{2\beta+1} + \frac{\epsilon}{2\beta}}$ and $\theta = \frac{2\beta}{2\beta+1}$ in (6.7) yields the desired error rate (2.19). \square

Acknowledgement. We would like to thank Mark Herbster and Yuan Yao for helpful discussions.

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, *Journal of the American Statistical Association*, 2005, to appear.
- [4] J. Bergh and J. Löfström, *Interpolation spaces, an introduction*, Springer-Verlag, New York, 1976.
- [5] T. Cai, Rates of convergence and adaption over Besov spaces under pointwise risk, *Statistica Sinica* **13** (2003), 881-902.
- [6] 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>
- [7] 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.
- [8] N. Cesa-Bianchi, A. Conconi and C. Gentile, A second-order perceptron algorithm, *SIAM J. Comput.* **34** (2005), 640-688.

- [9] 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.
- [10] 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.
- [11] F. Cucker and S. Smale, On the mathematical foundations of learning, *Bull. Amer. Math. Soc.* **39** (2001), 1–49.
- [12] F. Cucker and S. Smale, Best choices for regularization parameters in learning theory, *Found. Comput. Math.* **2** (2002), 413–428.
- [13] 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.
- [14] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, The Netherlands, 1996.
- [15] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.
- [16] J. Forster and M. K. Warmuth, Relative expected instantaneous loss bounds, *J. Computer and System Sciences* **64** (2002), 76–102.
- [17] M. Herbster and M. K. Warmuth, Tracking the best expert, *Machine Learning* **32** (1998), 151–178.
- [18] J. Kivinen, A. J. Smola, and R. C. Williamson, Online learning with kernels, *IEEE Trans. Signal Processing* **52** (2004), 2165–2176.
- [19] 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.
- [20] H. Robbins and S. Monro, A stochastic approximation method, *The Annals of Mathematical Statistics*, **22** (1951), 400–407.
- [21] S. Smale and Y. Yao, Online learning algorithms, *Found. Comp. Math.*, online version, September 2005.

- [22] S. Smale and D. X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.* **1** (2003), 17–41.
- [23] S. Smale and D. X. Zhou, Shannon sampling and function reconstruction from point values, *Bull. Amer. Math. Soc.* **41** (2004), 279-305.
- [24] S. Smale and D. X. Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.*, to appear.
- [25] I. Steinwart, Support vector machines are universally consistent, *J. Complexity* **18** (2002), 768–791.
- [26] C. J. Stone, Optimal global rates of convergence for nonparametric regression, *Annals of Statistics*, **10** (1982), 1040-1053.
- [27] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
- [28] V. Vovk, On-line regression competitive with reproducing kernel Hilbert spaces, Technical report, Second version, January 2006. Available at <http://arxiv.org/abs/cs.LG/0511058>.
- [29] G. Walba, *Spline models for observational data*, SIAM, 1990.
- [30] Q. Wu, Y. Ying, and D. X. Zhou, Learning rates of least-square regularized regression, *Found. Comput. Math.*, online version, September 2005.
- [31] Y. Yao, L. Rosasco, and A. Caponnetto, Early stopping in gradient descent boosting, September 2005. Available at <http://math.berkeley.edu/~yao/publications/earlystop.pdf>
- [32] Y. Ying and D. X. Zhou, Online regularized classification algorithms, submitted, June 2005.
- [33] T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comp.* **15** (2003), 1397–1437.
- [34] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, *Ann. Stat.* **32** (2004), 56–85.
- [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.

Appendix: Proof of Lemma 7

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

$$\begin{aligned} & \sum_{j \in \mathbb{N}_{\ell}} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) \\ &= \sum_{j \in \mathbb{N}_{\ell-1}} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) + \frac{1}{\mu^2 \ell^{2\theta}}. \end{aligned} \quad (6.9)$$

Let us first prove the case $\theta \neq 0$. The last term on the right-hand side of (6.9) 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 (6.9), we divide it into two terms:

$$\begin{aligned} \sum_{j \in \mathbb{N}_{\ell-1}} \frac{\eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\}}{\sum_{k=j+1}^{\ell} \eta_k + 1} &= \sum_{\substack{j \in \mathbb{N} \\ j \leq \frac{\ell-1}{2}}} \frac{\eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\}}{\sum_{k=j+1}^{\ell} \eta_k + 1} \\ &+ \sum_{j > \frac{\ell-1}{2}, j \in \mathbb{N}}^{\ell-1} \frac{\eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\}}{\sum_{k=j+1}^{\ell} \eta_k + 1}. \end{aligned} \quad (6.10)$$

By (5.8), we know that $\sum_{t=j+1}^{\ell} t^{-\theta} \geq \frac{1}{1-\theta} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta})$. Then, for any $j \leq \ell-1$, there holds

$$\begin{aligned} & \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) \\ & \leq \frac{\exp\left\{-\frac{\lambda}{(1-\theta)\mu} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta})\right\}}{\frac{1}{\mu(1-\theta)} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta}) + 1}. \end{aligned} \quad (6.11)$$

Below, using (6.11) we estimate the terms on the right-hand side of (6.10) 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 (6.11) implies that

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

Also,

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

Consequently, it follows that

$$\begin{aligned} & \sum_{\substack{j \in \mathbb{N} \\ j \leq \frac{\ell-1}{2}}} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) \\ & \leq \frac{\tilde{c}_\theta}{\mu} \ell^{-\min\{\theta, 1-\theta\}} \exp\left\{-\frac{\lambda(1-2^{\theta-1})}{(1-\theta)\mu} (\ell+1)^{1-\theta}\right\} \ln\left(\frac{e\ell}{2}\right) \end{aligned} \quad (6.12)$$

$$\text{where } \tilde{c}_\theta = \begin{cases} 2 & \text{for } \theta = 1/2, \\ \frac{2}{(1-2^{\theta-1})|2\theta-1|} & \text{for } \theta \neq 0, 1/2. \end{cases}$$

We now turn to the second term on the right-hand side of (6.10). Since $\exp\left\{-\frac{\lambda}{(1-\theta)\mu} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta})\right\} \leq 1$, from (6.11), we have that

$$\begin{aligned} & \sum_{j > \frac{\ell-1}{2}, j \in \mathbb{N}}^{\ell-1} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) \\ & \leq \frac{1}{\mu^2} \sum_{j > \frac{\ell-1}{2}, j \in \mathbb{N}}^{\ell-1} \frac{j^{-2\theta}}{\frac{1}{\mu(1-\theta)} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta}) + 1} \\ & \leq \frac{4}{\mu^2} \ell^{-\theta} \sum_{j > \frac{\ell-1}{2}, j \in \mathbb{N}}^{\ell-1} \frac{j^{-\theta}}{\frac{1}{\mu(1-\theta)} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta}) + 1}. \end{aligned}$$

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

$$\begin{aligned} & \sum_{j > \frac{\ell-1}{2}}^{\ell-1} \frac{j^{-\theta}}{\frac{1}{\mu(1-\theta)} ((\ell+1)^{1-\theta} - (j+1)^{1-\theta}) + 1} \\ & \leq 3 \sum_{j > \frac{\ell-1}{2}}^{\ell-1} \int_j^{j+1} \frac{(x+1)^{-\theta}}{\frac{1}{\mu(1-\theta)} ((\ell+1)^{1-\theta} - (x+1)^{1-\theta}) + 1} dx \\ & \leq 3 \int_{\frac{\ell-1}{2}}^{\ell} \frac{(x+1)^{-\theta}}{\frac{1}{\mu(1-\theta)} ((\ell+1)^{1-\theta} - (x+1)^{1-\theta}) + 1} dx \\ & = 3\mu \ln \left(1 + \left(\frac{1-2^{\theta-1}}{(1-\theta)\mu}\right) (\ell+1)^{1-\theta}\right) \leq 3\mu \ln \left(\frac{2(\ell+1)}{1-\theta}\right) \end{aligned}$$

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

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

Putting this and (6.12) together into (6.9), we have, for $\theta \in (0, 1)$, that

$$\begin{aligned} & \sum_{j \in \mathbb{N}_\ell} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) \\ & \leq \frac{c_\theta}{\mu} \left(\exp\left\{-\frac{\lambda(1-2^{\theta-1})}{(1-\theta)\mu}(\ell+1)^{1-\theta}\right\} \ell^{-\min\{\theta, 1-\theta\}} + \ell^{-\theta} \right) \ln\left(\frac{8\ell}{1-\theta}\right). \end{aligned}$$

where $c_\theta = \begin{cases} 13 & \text{for } \theta = 1/2, \\ \frac{13}{(1-2^{\theta-1})|2\theta-1|} & \text{for } \theta \in (0, 1/2) \cup (1/2, 1). \end{cases}$

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

$$\begin{aligned} \sum_{j \in \mathbb{N}_\ell} \eta_j^2 \exp\left\{-\lambda \sum_{k=j+1}^{\ell} \eta_k\right\} / \left(\sum_{k=j+1}^{\ell} \eta_k + 1\right) &= \frac{1}{\mu^2} \sum_{j=1}^{\ell} e^{-\frac{(\ell-j)\lambda}{\mu}} / \left(\frac{\ell-j}{\mu} + 1\right) \\ &\leq \frac{1}{\mu^2} \sum_{j \in \mathbb{N}_\ell} \frac{1}{\frac{\ell-j}{\mu} + 1} \leq \frac{1}{\mu} \ln(e\ell). \end{aligned}$$

This finishes the proof of Lemma 7. □