Online Learning Algorithms Can Converge Comparably Fast as Batch Learning

Junhong Lin and Ding-Xuan Zhou

***Abstract*— Online learning algorithms in a reproducing kernel Hilbert space associated with convex loss functions are studied. We show that in terms of the expected excess generalization error, they can converge comparably fast as corresponding kernel-based batch learning algorithms. Under mild conditions on loss functions and approximation errors, fast learning rates and finite sample upper bounds are established using polynomially decreasing step-size sequences. For some commonly used loss functions for classification, such as the logistic and the p-norm hinge loss functions with $p \in [1, 2]$, the learning rates are the same as those for Tikhonov regularization and can be of order $O(T^{-(1/2)} \log T)$, which are nearly optimal up to a logarithmic factor. Our novelty lies in a sharp estimate for the expected values of norms of the learning sequence (or an inductive argument to uniformly bound the expected risks of the learning sequence in expectation) and a refined error decomposition for online learning algorithms.**

***Index Terms*— Approximation error, learning theory, online learning, reproducing kernel Hilbert space (RKHS).**

I. Introduction

NONPARAMETRIC regression or classification aims at learning predictors from samples. To measure the performance of a predictor, one may use a loss function and its induced generalization error. Given a prediction function $f : X \to \mathbb{R}$, defined on a separable metric space X (input space), a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$ gives a local error $V(y, f(x))$ at $(x, y) \in Z := X \times Y$ with an output space $Y \subseteq \mathbb{R}$. The *generalization error* $\mathcal{E} = \mathcal{E}^V$ associated with the loss V and a Borel probability measure ρ on Z, defined as

$$\mathcal{E}(f) = \int_Z V(y, f(x)) d\rho,$$

measures the performance of f.

Kernel methods provide efficient nonparametric learning algorithms for dealing with nonlinear features, where reproducing kernel Hilbert spaces (RKHSs) are often used as hypothesis spaces in the design of learning algorithms. With suitable choices of kernels, RKHSs can be used to approximate functions in $L^2_{\rho_X}$, the space of square integrable functions with respect to the marginal probability measure ρ_X. A reproducing kernel $K : X \times X \to \mathbb{R}$ is a symmetric function such that $(K(u_i, u_j))_{i,j=1}^{\ell}$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^{\ell}$ in X. The RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ is the completion of the linear span of the set $\{K_x := K(x, \cdot) : x \in X\}$ with respect to the inner product given by $\langle K_x, K_u\rangle_K = K(x, u)$.

Batch learning algorithms perform learning tasks by using a whole batch of sample $\mathbf{z} = \{z_i = (x_i, y_i) \in Z\}_{i=1}^T$. Throughout this paper, we assume that the sample $\{z_i = (x_i, y_i)\}_i$ is drawn independently according to the measure ρ on Z. A large family of batch learning algorithms are generated by Tikhonov regularization

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t=1}^{T} V(y_t, f(x_t)) + \lambda \|f\|_K^2 \right\}, \ \lambda > 0. \quad (1)$$

Tikhonov regularization scheme (1) associated with convex loss functions has been extensively studied in the literature, and sharp learning rates have been well developed due to many results, as described in the books (see [1], [2], and references therein). But in practice, it may be difficult to implement when the sample size T is extremely large, as its standard complexity is about $O(T^3)$ for many loss functions. For example, for the hinge loss $V(y, f) = (1 - yf)_+ = \max\{1 - yf, 0\}$ or the square hinge loss $V(y, f) = (1 - yf)_+^2$ in classification corresponding to support vector machines, solving the scheme (1) is equivalent to solving a constrained quadratic program, with complexity of order $O(T^3)$.

With complexity $O(T)$ or $O(T^2)$, online learning represents an important family of efficient and scalable machine learning algorithms for large-scale applications. Over the past years, a variety of online learning algorithms have been proposed (see [3]–[7] and references therein). Most of them take the form of regularized online learning algorithms, i.e., given $f_1 = 0$,

$$f_{t+1} = f_t - \eta_t (V'_-(y_t, f_t(x_t)) K_{x_t} + \lambda_t f_t), \quad t = 1, \ldots, T-1 \quad (2)$$

where $\{\lambda_t\}$ is a regularization sequence and $\{\eta_t > 0\}$ is a step-size sequence. In particular, $\{\lambda_t\}$ is chosen as a constant sequence $\{\lambda > 0\}$ in [4] and [5] or as a time-varying regularization sequence in [8] and [9]. Throughout this paper, we assume that V is convex with respect to the second variable. That is, for any fixed $y \in Y$, the univariate function $V(y, \cdot)$ on $\mathbb{R}$ is convex. Hence, its left derivative $V'_-(y, f)$ exists at every $f \in \mathbb{R}$ and is nondecreasing.

We study the following online learning algorithm without regularization.

Manuscript received July 10, 2016; revised November 28, 2016; accepted February 17, 2017. This work was supported in part by the the Research Grants Council of Hong Kong under Project CityU 104113 and in part by the National Science Foundation of China under Grant 11461161006 and Grant 11471292. *(Corresponding author: Ding-Xuan Zhou.)*

J. Lin is with the Department of Mathematics, City University of Hong Kong, Hong Kong, also with the Laboratory for Computational and Statistical Learning, Massachusetts Institute of Technology, Cambridge, MA, USA, and also with the Istituto Italiano di Tecnologia, Genoa, Italy (e-mail: jhlin5@hotmail.com).

D.-X. Zhou is with the Department of Mathematics, City University of Hong Kong, Hong Kong (e-mail: mazhou@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2677970

Definition 1: The *online learning algorithm* without regularization associated with the loss V and the kernel K is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t V'_-(y_t, f_t(x_t))K_{x_t}, \quad t = 1, \ldots, T-1 \quad (3)$$

where $\{\eta_t > 0\}$ is a step-size sequence.

Let f_ρ^V be a minimizer of the generalization error $\mathcal{E}(f)$ among all measurable functions $f : X \to Y$. The main purpose of this paper is to estimate the expected excess generalization error $\mathbb{E}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$, where f_T is generated by the unregularized online learning algorithm (3) with a convex loss V. Under a mild condition on approximation errors and a growth condition on the loss V, we derive upper bounds for the expected excess generalization error using polynomially decaying step-size sequences. Our bounds are independent of the capacity of the RKHS $\mathcal{H}_K$, and are comparable to those for Tikhonov regularization (1), see more details in Section III. In particular, for some loss functions, such as the logistic loss, the p-absolute value loss, and the p-hinge loss with $p \in [1, 2]$, our learning rates are of order $O(T^{-(1/2)} \log T)$, which is nearly optimal in the sense that up to a logarithmic factor, it matches the minimax rates of order $O(T^{-(1/2)})$ in [10] for stochastic approximation in the nonstrongly convex case. In our approach, an inductive argument is involved, to develop sharp estimates for the expected values of $\|f_t\|_K^2$, which is better than uniform bounds in the existing literature, or to bound the expected values of $\mathcal{E}(f_t)$ uniformly. Our second novelty is a refined error decomposition, which might be used for other online or gradient descent algorithms [11], [12] and is of independent interest.

The rest of this paper is organized as follows. We introduce in Section II some basic assumptions that underlie our analysis, and give our main results as well as examples, illustrating our upper bounds for the expected excess generalization error for different kinds of loss functions in learning theory. Section III contributes to discussions and comparisons with previous results, mainly on online learning algorithms with or without regularization, and the common Tikhonov regularization batch learning algorithms. Section IV deals with the proof of our main results, which relies on an error decomposition as well as the lemmas proved in the Appendix. Finally, in Section V, we will discuss the numerical simulation of the studied algorithms, and give some numerical simulations, which complements our theoretical results.

II. Main Results

In this section, we first state our main assumptions, following with some comments. We then present our main results with simple discussions.

A. Assumptions on the Kernel and Loss Function

Throughout this paper, we assume that the kernel is bounded on $X \times X$ with the constant

$$\kappa = \sup_{x \in X} \max(\sqrt{K(x,x)}, 1) < \infty \quad (4)$$

and that $|V|_0 := \sup_{y \in Y} V(y, 0) < \infty$. These bounded conditions on K and V are common in learning theory. They are satisfied when X is compact and Y is a bounded subset of $\mathbb{R}$. Moveover, the condition $|V|_0 < \infty$ implies that $\mathcal{E}(f_\rho^V)$ is finite

$$\mathcal{E}(f_\rho^V) \le \mathcal{E}(0) = \int_Z V(y, 0) d\rho \le |V|_0.$$

The assumption on the loss function V is a growth condition for its left derivative $V'_-(y, \cdot)$.

Assumption 1.a: Assume that for some $q \ge 0$ and constant $c_q > 0$, there holds

$$|V'_-(y, f)| \le c_q(1 + |f|^q), \quad \forall f \in \mathbb{R}, y \in Y. \quad (5)$$

The growth condition (5) is implied by the requirement for the loss function to be Nemitiski [2], [13]. It is weaker than, either assuming the loss or its gradient, to be Lipschitz in its second variable as often done in learning theory, or assuming the loss to be α-activating with $\alpha \in (0, 1]$ in [14].

An alterative to Assumption 1.a made for V in the literature is the following assumption [15], [16].

Assumption 1.b: Assume that for some $a_V, b_V \ge 0$, there holds

$$|V'_-(y, f)|^2 \le a_V V(y, f) + b_V, \quad \forall f \in \mathbb{R}, y \in Y. \quad (6)$$

Assumption 1.b is satisfied for most loss functions commonly used in learning theory, when Y is a bounded subset of $\mathbb{R}$. In particular, when $V(y, \cdot)$ is smooth, it is satisfied with $b_V = 0$ and some appropriate a_V [16, Lemma 2.1].

B. Assumption on the Approximation Error

The performance of online learning algorithm (3) depends on how well the target function f_ρ^V can be approximated by functions from the hypothesis space $\mathcal{H}_K$. For our purpose of estimating the excess generalization error, the approximation is measured by $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ with $f \in \mathcal{H}_K$. Moreover, the output function f_T produced by the online learning algorithm lies in a ball of $\mathcal{H}_K$ with the radius increasing with T (as shown in Lemma 7). So we measure the approximation ability of the hypothesis space $\mathcal{H}_K$ with respect to the generalization error $\mathcal{E}(f)$ and f_ρ^V by penalizing the functions with their norm squares [17] as follows.

Definition 2: The approximation error associated with the triplet (ρ, V, K) is defined by

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \{\mathcal{E}(f) - \mathcal{E}(f_\rho^V) + \lambda \|f\|_K^2\}, \quad \lambda > 0. \quad (7)$$

When $f_\rho^V \in \mathcal{H}_K$, we can take $f = f_\rho^V$ in (7) and find $\mathcal{D}(\lambda) \le \|f_\rho^V\|_K^2 \lambda = O(\lambda)$. When $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ can be arbitrarily small as f runs over $\mathcal{H}_K$, we know that $\mathcal{D}(\lambda) \to 0$ as $\lambda \to 0$. To derive explicit convergence rates for the studied online algorithm, we make the following assumption on the decay of the approximation error to be $O(\lambda^\beta)$.

Assumption 3: Assume that for some $\beta \in (0, 1]$ and $c_\beta > 0$, the approximation error satisfies

$$\mathcal{D}(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0. \quad (8)$$

C. Alternative Conditions on the Approximation Error

Assumption (8) on the approximation error is standard in analyzing both Tikhonov regularization schemes [1], [2] and online learning algorithms [8], [9], [18]. It is independent of the sample, and measures the approximation ability of the space $\mathcal{H}_K$ to f_ρ^V with respect to (ρ, V). It may be replaced by alterative simple conditions for specified loss functions.

For a Lipschitz continuous loss function meaning that

$$\sup_{y\in Y, f, f'\in\mathbb{R}} \frac{|V(y,f)-V(y,f')|}{|f-f'|} = l < \infty$$

it is easy to see that $\mathcal{E}(f)-\mathcal{E}(f_\rho^V) \le l\|f-f_\rho^V\|_{L^1_{\rho_X}}$, and thus a sufficient condition for (8) is

$$\inf_{f\in\mathcal{H}_K}\left\{\left\|f-f_\rho^V\right\|_{L^1_{\rho_X}} + \lambda\|f\|_K^2\right\} = O(\lambda^\beta).$$

In particular, for the hinge loss in classification, we have $l = 1$. Such a condition measures quantitatively the approximation of the function f_ρ^V in the space $L^1_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$, and can be characterized [2], [17] by requiring f_ρ^V to lie in some interpolation space between $\mathcal{H}_K$ and $L^1_{\rho_X}$.

For the least squares loss, $f_\rho^V = f_\rho$ and there holds $\mathcal{E}(f)-\mathcal{E}(f_\rho) = \|f-f_\rho\|^2_{L^2_{\rho_X}}$. Here, f_ρ is the regression function defined at $x \in X$ to be the expectation of the conditional distribution $\rho(y|x)$ given x. In this case, condition (8) is exactly

$$\inf_{f\in\mathcal{H}_K}\left\{\left\|f-f_\rho\right\|^2_{L^2_{\rho_X}} + \lambda\|f\|_K^2\right\} = O(\lambda^\beta).$$

This condition is about the approximation of the function f_ρ in the space $L^2_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$. It can be characterized [17] by requiring that f_ρ lies in $L_K^{\beta/2}(L^2_{\rho_X})$, the range of the operator $L_K^{\beta/2}$. Recall that the integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ is defined by

$$L_K(f) = \int_X f(x)K_x d\rho_X, \quad f \in L^2_{\rho_X}.$$

Since K is a reproducing kernel with finite κ, the operator L_K is symmetric, compact, and positive, and its power $L_K^{\beta/2}$ is well defined.

D. Stating Main Results

Our first main result of this paper, to be proved in Section IV, is stated as follows.

Theorem 1: Under Assumption 1.a, let $\eta_t = \eta_1 t^{-\theta}$ with $\max((1/2), q/(q+1)) < \theta < 1$ and η_1 satisfying

$$0 < \eta_1 \le \min\left(\sqrt{\frac{(q^*-1)(1-\theta)}{12c_q^2(1+\kappa)^{2q+2}q^*}}, \frac{1-\theta}{2(1+2|V|_0)}\right) \quad (9)$$

where we denote $q^* = 2\theta-(1-\theta)\cdot\max(0, q-1) > 0$. Then

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\} \le \widetilde{C}\{\mathcal{D}(T^{\theta-1})+T^{\theta-1}\} \quad (10)$$

where $\widetilde{C}$ is a positive constant depending on η_1, q, κ, and θ (independent of T and given explicitly in the proof).

Combining Theorem 1 with Assumption 3, we get the following explicit learning rates.

Corollary 2: Under the conditions of Theorem 1 and Assumption 3, we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-(1-\theta)\beta}).$$

Replacing Assumption 1.a by Assumption 1.b, we can relax the restriction on θ in Theorem 1 as $\theta \in (0, 1)$, which thus improves the learning rates. Concretely, we have the following convergence results.

Theorem 3: Under Assumption 1.b, let $\eta_t = \eta_1 t^{-\theta}$ with $0 < \theta < 1$ and η_1 satisfying

$$0 < \eta_1 \le \frac{\min(\theta, 1-\theta)}{2a_V\kappa^2}. \quad (11)$$

Then

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\} \le \widetilde{C}'\{\mathcal{D}(T^{\theta-1})+T^{-\min(\theta,1-\theta)}\}\log T \quad (12)$$

where $\widetilde{C}'$ is a positive constant depending on $\eta_1, a_V, b_V\kappa$, and θ (independent of T and given explicitly in the proof).

Corollary 4: Under the conditions of Theorem 3 and Assumption 3, let $\theta = \beta/(\beta+1)$. Then, we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-\frac{\beta}{\beta+1}}\log T).$$

To illustrate the above-mentioned results, we give the following examples of commonly used loss functions in learning theory with corresponding learning rates for online learning algorithms (3).

Example 1: Assume $|y| \le M$, and conditions (4) and (8) hold with $0 < \beta \le 1$. For the least squares loss $V(y,a) = (y-a)^2$, the p-norm loss $V(y,a) = |y-a|^p$ with $p \in [1,2)$, the hinge loss $V(y,a) = (1-ya)_+$, the logistic loss $V(y,a) = \log(1+e^{-ya})$, and the p-norm hinge loss $V(y,a) = ((1-ya)_+)^p$ with $p \in (1,2]$, choosing $\eta_t = \eta_1 t^{-\beta/(\beta+1)}$ with η_1 satisfying (11), we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-\frac{\beta}{\beta+1}}\log T)$$

which is of order $O(T^{-(1/2)}\log T)$ if $\beta = 1$.

Example 1 follows from Corollary 4, while the conclusion of the next example is seen from Corollary 2.

Example 2: Under the assumption of Example 1, for the p-norm loss $V(y,a) = |y-a|^p$ and the p-norm hinge loss $V(y,a) = ((1-ya)_+)^p$ with $p > 2$, selecting $\eta_t = \eta_1 t^{-((p-1)/p+\epsilon)}$ with $\epsilon \in (0,(1/p))$ and η_1 such that (9) holds with $q = p-1$, we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-(\frac{1}{p}-\epsilon)\beta})$$

which is of order $O(T^{\epsilon-(1/p)})$ if $\beta = 1$.

Remark 1:
1) The learning rates given in Example 1 are optimal in the sense that they are the same as those for the Tikhonov regularization [2, Ch. 7].
2) According to Example 1, the optimal learning rates are achieved when $\eta_t \simeq t^{-\beta/(1+\beta)}$. Since β is not known in general, in practice, a hold-out cross-validation method can be used to tune the ideal exponential parameter θ.
3) Our analysis can be extended to the case of constant step sizes. In fact, following our proofs given in the following, the readers can see that, when $\eta_t = T^{-\beta/(\beta+1)}$ for

$t = 1, \ldots, T-1$, the results stated in Example 1 still hold.

E. Classification Problem

The binary classification problem in learning theory is a special case of our learning problems. In this case, $Y = \{1, -1\}$. A classifier for classification is a function f from X to Y and its misclassification error $\mathcal{R}(f)$ is defined as the probability of the event $\{(x, y) \in Z : y \neq f(x)\}$ of making wrong predictions. A minimizer of the misclassification error is the Bayes rule $f_c : X \to Y$ given by

$$f_c(x) = \begin{cases} 1, & \text{if } \rho(y = 1|x) \geq 1/2 \\ -1, & \text{otherwise.} \end{cases}$$

The performance of a classification algorithm can be measured by the excess misclassification error $\mathcal{R}(f) - \mathcal{R}(f_c)$. For the online learning algorithms (3), our classifier is given by $\operatorname{sign}(f_T)$

$$\operatorname{sign}(f_T)(x) = \begin{cases} 1, & \text{if } f_T(x) \geq 0 \\ -1, & \text{otherwise.} \end{cases}$$

So our error analysis aims at the excess misclassification error

$$\mathcal{R}(\operatorname{sign}(f_T)) - \mathcal{R}(f_c).$$

This can be often done [15], [19], [20] by bounding the excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ and using the so-called comparison theorems. For example, for the hinge loss $V(y, f(x)) = (1 - yf(x))_+$, it was shown in [21] that $f_\rho^V = f_c$ and the comparison theorem in [15] asserts that

$$\mathcal{R}(\operatorname{sign}(f)) - \mathcal{R}(f_c) \leq \mathcal{E}(f) - \mathcal{E}(f_c)$$

for any measurable function f. For the least squares loss, the logistic loss, and the p-norm hinge loss with $p > 1$, the comparison theorem [19], [20] states that there exists a constant c_V such that for any measurable function f

$$\mathcal{R}(\operatorname{sign}(f)) - \mathcal{R}(f_c) \leq c_V \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho^V)}.$$

Furthermore, if the distribution ρ satisfies a Tsybakov noise condition, then there is a refined comparison relation for a so-called admissible loss function, see more details in [19] and [20].

III. Related Work and Discussion

There is a large amount of work on online learning algorithms and, more generally, stochastic approximations (see [3]–[9], [12], [14]–[16], [18], [22], [23], and the references therein). In this section, we discuss some of the previous results related to this paper.

The regret bounds for online algorithms have been well studied in the literature [22]–[24]. Most of these results assume that the hypothesis space is of finite dimension, or the gradient is bounded, or the objective functions are strongly convex. Using an "online-to-batch" approach, generalization error bounds can be derived from the regret bounds.

For the nonparametric regression or classification setting, online algorithms have been studied in [3]–[6], [8], [9], [14], and [18]. Recently, Ying and Zhou [14] showed that for a loss function V satisfying

$$|V'_-(y, f) - V'_-(y, g)| \leq L|f - g|^\alpha, \quad \forall y \in Y,\ f, g \in \mathbb{R} \tag{13}$$

for some $0 < \alpha \leq 1$ and $0 < L < \infty$, under the assumption of existence of $\arg\inf_{f \in \mathcal{H}_K} \mathcal{E}(f) = f_{\mathcal{H}_K} \in \mathcal{H}_K$, by selecting $\eta_t = \eta_1 t^{-2/(\alpha+2)}$, there holds

$$\mathbb{E}_{z_1, z_2, \ldots, z_{T-1}}[\mathcal{E}(f_T) - \mathcal{E}(f_{\mathcal{H}_K})] = O(T^{-\frac{\alpha}{\alpha+2}}).$$

It is easy to see that such a loss function always satisfies the growth condition (5) with $q = \alpha$, when $\sup_{y \in Y} |V'_-(y, 0)| < \infty$. Therefore, as shown in Corollary 2, our learning rates for such a loss function are of order $O(T^{-(\beta/2)+\epsilon})$, which reduces to $O(T^{-(1/2)+\epsilon})$, if we further assume the existence of $f_{\mathcal{H}_K} = \arg\inf_{f \in \mathcal{H}_K} \mathcal{E}(f) \in \mathcal{H}_K$, as in [14]. Note that in general, $f_{\mathcal{H}_K}$ may not exist, thus our results require weaker assumptions, involving approximation errors in the error bounds. Also, our obtained upper bounds are better and are especially of great improvements when α is close to 0. In the cases of $\beta = 1$, these bounds are nearly optimal and up to a logarithmic factor, coincide with the minimax rates of order $O(T^{-(1/2)})$ in [10] for stochastic approximations in the nonstrongly convex case. Besides, in comparison with [14], where only loss functions satisfying (13) with $\alpha \in (0, 1]$ are considered, a broader class of convex loss functions are considered in this paper. At last, let us mention that for the least squares loss, the obtained learning rate $O(T^{-\beta/(\beta+1)} \log T)$ from Example 1 is the same as that derived in [18].

Our learning rates are also better than those for online classification in [5] and [8]. For example, for the hinge loss, the upper bound obtained in [5] is of the form $O(T^{\epsilon - \beta/(2(\beta+1))})$, while the bound in Example 1 is of the form $O(T^{-\beta/(1+\beta)} \log T)$, which is better. For a p-norm hinge loss with $p > 1$, the bound obtained in [5] is of order $O(T^{\epsilon - \beta/(2[(2-\beta)p+3\beta])})$, while the bounds in Examples 1 and 2 are of order $O(T^{\epsilon - (\beta/\max(p,2))})$.

We now compare our learning rates with those for batch learning algorithms. For general convex loss functions, the method for which sharp bounds are available is Tikhonov regularization (1). If no noise condition is imposed, the best capacity-independent error bounds for (1) with Lipschitz loss functions [2, Ch. 7], are of order $O(T^{-\beta/(\beta+1)})$. The obtained bounds in Example 1 for Lipschitz loss functions are the same as the best one available for the Tikhonov regularization, up to a logarithmic factor.

We conclude this section with some possible future work. First, it would be interesting to prove sharper rates by considering the capacity assumptions on the hypothesis spaces. Second, in this paper, we only consider the i.i.d. (independent identically distributed) setting. However, our analysis can be extended to some non-i.i.d. settings, such as the setting with Markov sampling as in [25] and [26]. Finally, our analysis may also be applied to other stochastic learning models, such as online learning with random features [27], which will be studied in our future work.

IV. Proof of Main Results

In this section, we prove our main results, Theorems 1 and 3.

A. Preliminary Lemmas

To prove Theorems 1 and 3, we need several lemmas to be proved in the Appendix.

Lemma 1 is key and will be used several times for the proof of Theorem 1. It is inspired by the recent work in [14], [28], and [29].

Lemma 1: Under Assumption 1.a, for any $f \in \mathcal{H}_K$, and $t = 1, \ldots, T-1$

$$\|f_{t+1} - f\|_K^2 \le \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t [V(y_t, f(x_t)) - V(y_t, f_t(x_t))] \quad (14)$$

where

$$G_t = \kappa c_q \big(1 + \kappa^q \|f_t\|_K^q\big). \quad (15)$$

Using Lemma 1 and an inductive argument, we can estimate the expected value $\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1}\|_K^2]$ and provide a novel bound as follows. For notational simplicity, we denote by $\mathcal{A}(f_*)$ the excess generalization error of $f_* \in \mathcal{H}_K$ with respect to (ρ, V) as

$$\mathcal{A}(f_*) = \mathcal{E}(f_*) - \mathcal{E}\big(f_\rho^V\big). \quad (16)$$

Lemma 2: Under Assumption 1.a, let $\eta_t = \eta_1 t^{-\theta}$ with $\max((1/2), q/(q+1)) < \theta < 1$ and η_1 satisfying (9). Then, for an arbitrarily fixed $f_* \in \mathcal{H}_K$ and $t = 1, \ldots, T-1$

$$\mathbb{E}_{z_1,\ldots,z_t}\big[\|f_{t+1}\|_K^2\big] \le 6\|f_*\|_K^2 + 4\mathcal{A}(f_*) t^{1-\theta} + 4 \quad (17)$$

and

$$\eta_{t+1}^2 \mathbb{E}_{z_1,\ldots,z_t}\big[G_{t+1}^2\big] \le \big(3\|f_*\|_K^2 + 2\mathcal{A}(f_*) t^{1-\theta} + 3\big)(t+1)^{-q^*} \quad (18)$$

where q^* is defined in Theorem 1.

Lemma 2 asserts that for a suitable choice of decaying step sizes, $\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1}\|_K^2]$ can be well bounded if there exists some $f_* \in \mathcal{H}_K$ such that $\mathcal{A}(f_*)$ is small. It improves uniform bounds found in the existing literature.

Replacing Assumption 1.a with Assumption 1.b in Lemma 1, we can prove the following result.

Lemma 3: Under Assumption 1.b, we have for any arbitrary $f \in \mathcal{H}_K$, and $t = 1, \ldots, T-1$

$$\|f_{t+1} - f\|_K^2 \le \|f_t - f\|_K^2 + \eta_t^2 \kappa^2 b_V + a_V \eta_t^2 \kappa^2 V(y_t, f_t(x_t)) + 2\eta_t [V(y_t, f(x_t)) - V(y_t, f_t(x_t))]. \quad (19)$$

Using Lemma 3, and an induction argument, we can bound the expected risks of the learning sequence as follows.

Lemma 4: Under Assumption 1.b, let $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in (0,1)$ and η_1 such that (11). Then, for any $t = 1, \ldots, T-1$, there holds

$$\mathbb{E}_{z_1,\ldots,z_{t-1}} \mathcal{E}(f_t) \le \tilde{B} \quad (20)$$

where $\tilde{B}$ is a positive constant depending only on $\eta_1, \theta, b_V, \kappa^2$, and $|V|_0$ (given explicitly in the proof).

We also need the following elementary inequalities, which, for completeness, will be proved in the Appendix using a similar approach as that in [28].

Lemma 5: For any $q^* \ge 0$, there holds

$$\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \le 2T^{-\min(1,q^*)} \log(eT).$$

Furthermore, if $q^* > 1$, then

$$\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \le 2\left(2^{q^*} + \frac{q^*}{q^*-1}\right) T^{-1}.$$

B. Deriving Convergence From Averages

An essential tool in our error analysis is to derive the convergence of a sequence $\{u_t\}_t$ from its averages of the form $(1/T)\sum_{j=1}^T u_j$ and $(1/k)\sum_{j=T-k+1}^T u_j$. Lemma 6 is elementary for sequences and the idea is from [7]. We provide a proof in the Appendix.

Lemma 6: Let $\{u_t\}_t$ be a real-valued sequence. We have

$$u_T = \frac{1}{T}\sum_{j=1}^T u_j + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{j=T-k+1}^{T} (u_j - u_{T-k}). \quad (21)$$

From Lemma 6, we see that if the average $(1/T)\sum_{j=1}^T u_j$ tends to some u^* and the moving average $\sum_{k=1}^{T-1} 1/(k(k+1)) \sum_{j=T-k+1}^T (u_j - u_{T-k})$ tends to zero, then u_T tends to u^* as well.

Recall that our goal is to derive upper bounds for the expected excess generalization error $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$. We can easily bound the weighted average $(1/T)\sum_{t=1} 2\eta_t \mathbb{E}_{z_1,\ldots,z_{T-1}}[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)]$ from (14) [or (19)]. This, together with Lemma 6, demonstrates how to bound the weighted excess generalization error $2\eta_T \mathbb{E}_{z_1,\ldots,z_{T-1}}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$ in terms of the weighted average and the moving weighted average. Interestingly, the bounds on the weighted average and the moving weighted average are essentially the same, as shown in Sections IV-D and IV-E.

C. Error Decomposition

Our proofs rely on a novel error decomposition derived from Lemma 6. In what follows, we shall use the notation $\mathbb{E}$ for $\mathbb{E}_{z_1,\ldots,z_{T-1}}$. Choosing $u_t = 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\}$ in Lemma 6, we get

$$\begin{aligned} 2\eta_T \mathbb{E}\big\{\mathcal{E}(f_T) - \mathcal{E}\big(f_\rho^V\big)\big\} &= \frac{1}{T}\sum_{j=1}^T 2\eta_j \mathbb{E}\big\{\mathcal{E}(f_j) - \mathcal{E}\big(f_\rho^V\big)\big\} \\ &\quad + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{j=T-k+1}^{T} \big(2\eta_j \mathbb{E}\big\{\mathcal{E}(f_j) - \mathcal{E}\big(f_\rho^V\big)\big\} \\ &\quad - 2\eta_{T-k} \mathbb{E}\big\{\mathcal{E}(f_{T-k}) - \mathcal{E}\big(f_\rho^V\big)\big\}\big) \end{aligned}$$

which can be rewritten as

$$
\begin{aligned}
&2\eta_T \mathbb{E}\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\} \\
&\quad = \frac{1}{T}\sum_{t=1}^{T} 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\} \\
&\qquad + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\} \\
&\qquad + \sum_{k=1}^{T-1} \frac{1}{k+1}\left[\frac{2}{k}\sum_{t=T-k+1}^{T} \eta_t - \eta_{T-k}\right] \\
&\qquad \times \mathbb{E}\{\mathcal{E}(f_{T-k}) - \mathcal{E}(f_\rho^V)\}. \qquad (22)
\end{aligned}
$$

Since, $\mathcal{E}(f_{T-k}) - \mathcal{E}(f_\rho^V) \geq 0$ and that $\{\eta_t\}_{t\in\mathbb{N}}$ is a nonincreasing sequence, we know that the last term of (22) is at most zero. Therefore, we get

$$
\begin{aligned}
&2\eta_T \mathbb{E}\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\} \\
&\quad \leq \frac{1}{T}\sum_{t=1}^{T} 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\} \\
&\qquad + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\}. \qquad (23)
\end{aligned}
$$

D. Proof of Theorem 1

In this section, we prove Theorem 1. We first prove the following general result, from which we can derive Theorem 1.

Theorem 5: Under Assumption 1.a, let $\eta_t = \eta_1 t^{-\theta}$ with $\max((1/2), q/(q+1)) < \theta < 1$ and η_1 satisfying (9). Then, for any fixed $f_* \in \mathcal{H}_K$

$$
\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_{T-1}}\{\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)\} \\
&\quad \leq \bar{C}_1 \mathcal{A}(f_*) + \bar{C}_2 \|f_*\|_K^2 T^{-1+\theta} + \bar{C}_3 T^{-1+\theta} \qquad (24)
\end{aligned}
$$

where $\bar{C}_1, \bar{C}_2$, and $\bar{C}_3$ are positive constants depending on η_1, q, κ, and θ (independent of T or f_* and given explicitly in the proof).

Proof: Let us first bound the average error, the first term of (23). Choosing $f = f_*$ in (14), taking expectation on both sides, and noting that f_t depends only on $z_1, z_2, \ldots, z_{t-1}$, we have

$$
\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\
&\quad \leq \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f_*\|_K^2\right] + \eta_t^2 \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[G_t^2\right] \\
&\qquad + 2\eta_t \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\mathcal{E}(f_*) - \mathcal{E}(f_t)\right] \\
&\quad = \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f_*\|_K^2\right] + \eta_t^2 \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[G_t^2\right] \\
&\qquad + 2\eta_t \mathcal{A}(f_*) - 2\eta_t \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \qquad (25)
\end{aligned}
$$

which implies

$$
\begin{aligned}
&2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \\
&\quad \leq \mathbb{E}\left[\|f_t - f_*\|_K^2\right] - \mathbb{E}\left[\|f_{t+1} - f_*\|_K^2\right] \\
&\qquad + 2\eta_t \mathcal{A}(f_*) + \eta_t^2 \mathbb{E}\left[G_t^2\right].
\end{aligned}
$$

Summing over $t = 1, \ldots, T$, with $f_1 = 0$ and $\eta_t = \eta_1 t^{-\theta}$

$$
\begin{aligned}
&\sum_{t=1}^{T} 2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \\
&\quad \leq \|f_*\|_K^2 + 2\eta_1 \mathcal{A}(f_*) \sum_{t=1}^{T} t^{-\theta} + \sum_{t=1}^{T} \eta_t^2 \mathbb{E}\left[G_t^2\right].
\end{aligned}
$$

This together with (18) yields

$$
\begin{aligned}
&\sum_{t=1}^{T} 2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \\
&\quad \leq \|f_*\|_K^2 + 2\eta_1 \mathcal{A}(f_*) \sum_{t=1}^{T} t^{-\theta} \\
&\qquad + \left(3\|f_*\|_K^2 + 2\mathcal{A}(f_*) T^{1-\theta} + 3\right) \sum_{t=1}^{T} t^{-q^*}.
\end{aligned}
$$

Applying the elementary inequalities

$$
\sum_{j=1}^{t} j^{-\theta'} \leq 1 + \int_1^t u^{-\theta'} du \leq \begin{cases} \dfrac{t^{1-\theta'}}{1-\theta'}, & \text{when } \theta' < 1 \\ \log(et), & \text{when } \theta' = 1 \\ \dfrac{\theta'}{\theta'-1}, & \text{when } \theta' > 1 \end{cases} \qquad (26)
$$

with $\theta' = \theta$ and $q^* > 1$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} 2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \\
&\quad \leq \left(\frac{2\eta_1}{1-\theta} + \frac{2q^*}{q^*-1}\right) \mathcal{A}(f_*) T^{1-\theta} + \left(4\|f_*\|_K^2 + 3\right) \frac{q^*}{q^*-1}.
\end{aligned}
$$

Dividing both sides by T, we get a bound for the first term of (23) as

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} 2\eta_t \mathbb{E}\left[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\right] \\
&\quad \leq \left(\frac{2\eta_1}{1-\theta} + \frac{2q^*}{q^*-1}\right) \mathcal{A}(f_*) T^{-\theta} \\
&\qquad + \left(4\|f_*\|_K^2 + 3\right) \frac{q^*}{q^*-1} T^{-1}. \qquad (27)
\end{aligned}
$$

Then, we turn to the moving average error, the second term of (23). Let $k \in \{1, \ldots, T-1\}$. Note that f_{T-k} depends only on $z_1, \ldots, z_{T-k-1}$. Taking expectation on both sides of (14), and rearranging terms, we have that for $t \geq T-k$

$$
\begin{aligned}
&2\eta_t \mathbb{E}[\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})] \\
&\quad \leq \mathbb{E}\left[\|f_t - f_{T-k}\|_K^2\right] - \mathbb{E}\left[\|f_{t+1} - f_{T-k}\|_K^2\right] + \eta_t^2 \mathbb{E}\left[G_t^2\right].
\end{aligned}
$$

Using this inequality repeatedly for $t = T-k, \ldots, T$, we have

$$
\begin{aligned}
&\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k+1}^{T} 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_{T-k})\} \\
&\quad \leq \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} \eta_t^2 \mathbb{E}\left[G_t^2\right].
\end{aligned}
$$

Combining this with (18) implies

$$\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}$$
$$\leq \big(3\|f_*\|_K^2+2\mathcal{A}(f_*)T^{1-\theta}+3\big)\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}t^{-q^*}.$$

Applying Lemma 5, we have

$$\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}$$
$$\leq 2\left(2^{q^*}+\frac{q^*}{q^*-1}\right)\big(3\|f_*\|_K^2+2\mathcal{A}(f_*)T^{1-\theta}+3\big)T^{-1}. \tag{28}$$

Finally, putting (27) and (28) into the error decomposition (23), and then dividing both sides by $2\eta_T = 2\eta_1 T^{-\theta}$, by a direct calculation, we get our desired bound (24) with

$$\bar{C}_1 = \frac{1}{1-\theta}+\frac{3q^*}{\eta_1(q^*-1)}+\frac{2^{q^*+1}}{\eta_1}$$
$$\bar{C}_2 = \frac{5q^*}{\eta_1(q^*-1)}+\frac{3\cdot 2^{q^*}}{\eta_1}$$

and

$$\bar{C}_3 = \frac{9q^*}{2\eta_1(q^*-1)}+\frac{3\cdot 2^{q^*}}{\eta_1}.$$

The proof is complete. □

We are in a position to prove Theorem 1.

Proof of Theorem 1: By Theorem 5, we have

$$\mathbb{E}\big\{\mathcal{E}(f_T)-\mathcal{E}\big(f_\rho^V\big)\big\}$$
$$\leq (\bar{C}_1+\bar{C}_2)\big\{\mathcal{E}(f_*)-\mathcal{E}\big(f_\rho^V\big)+\|f_*\|_K^2T^{\theta-1}\big\}+\bar{C}_3T^{\theta-1}.$$

Since the constants $\bar{C}_1, \bar{C}_2$, and $\bar{C}_3$ are independent of $f_*\in\mathcal{H}_K$, we take the infimum over $f_*\in\mathcal{H}_K$ on both sides, and conclude that

$$\mathbb{E}\big\{\mathcal{E}(f_T)-\mathcal{E}\big(f_\rho^V\big)\big\}\leq(\bar{C}_1+\bar{C}_2)\mathcal{D}(T^{\theta-1})+\bar{C}_3T^{\theta-1}.$$

The proof of Theorem 1 is complete by taking $\widetilde{C}=\bar{C}_1+\bar{C}_2+\bar{C}_3$.

E. Proof of Theorem 3

In this section, we give the proof of Theorem 3. It follows from the following more general theorem, as shown in the proof of Theorem 1.

Theorem 6: Under Assumption 1.b, let $\eta_t = \eta_1 t^{-\theta}$ with $0<\theta<1$ and η_1 satisfying (11). Then, for any fixed $f_*\in\mathcal{H}_K$

$$\mathbb{E}_{z_1,\ldots,z_{T-1}}\big\{\mathcal{E}(f_T)-\mathcal{E}\big(f_\rho^V\big)\big\}$$
$$\leq\big(2\mathcal{A}(f_*)+(2\eta_1)^{-1}\|f_*\|_K^2T^{-1+\theta}+\bar{B}_1T^{-\min(\theta,1-\theta)}\big)\log T \tag{29}$$

where $\bar{B}_1$ is a positive constant depending only on η_1, a_V, b_V, κ, and θ (independent of T or f_* and given explicitly in the proof).

Proof: The proof parallels to that of Theorem 5. Note that we have the error decomposition (23). We only need to estimate the last two terms of (23).

To bound the first term of the right-hand side of (23), we first apply Lemma 3 with a fixed $f\in\mathcal{H}_K$ and subsequently take the expectation on both sides of (19) to get

$$\mathbb{E}\big[\|f_{l+1}-f\|_K^2\big]$$
$$\leq\mathbb{E}\big[\|f_l-f\|_K^2\big]$$
$$+\eta_l^2\kappa^2(a_V\mathbb{E}[\mathcal{E}(f_l)]+b_V)+2\eta_l\mathbb{E}(\mathcal{E}(f)-\mathcal{E}(f_l)). \tag{30}$$

By Lemma 4, we have (20). Introducing (20) into (30) with $f=f_*$, and rearranging terms

$$2\eta_l\mathbb{E}\big(\mathcal{E}(f_l)-\mathcal{E}\big(f_\rho^V\big)\big)\leq\mathbb{E}\big[\|f_l-f_*\|_K^2-\|f_{l+1}-f_*\|_K^2\big]$$
$$+2\eta_l\mathcal{A}(f_*)+\eta_l^2\kappa^2(a_V\tilde{B}+b_V).$$

Summing up over $l=1,\ldots,T$, rearranging terms, and then dividing both sides by T, we get

$$\frac{1}{T}\sum_{l=1}^{T}2\eta_l\mathbb{E}(\mathcal{E}(f_l)-\mathcal{E}(f_*))$$
$$\leq\frac{\|f_*\|_K^2}{T}+\frac{2\eta_1}{T}\mathcal{A}(f_*)\sum_{t=1}^{T}t^{-\theta}+\eta_1^2\kappa^2(a_V\tilde{B}+b_V)\frac{1}{T}\sum_{l=1}^{T}l^{-2\theta}.$$

By using the elementary inequality with $q\geq 0, T\geq 3$

$$\sum_{t=1}^{T}t^{-q}\leq T^{\max(1-q,0)}\sum_{t=1}^{T}t^{-1}\leq 2T^{\max(1-q,0)}\log T$$

one can get

$$\frac{1}{T}\sum_{l=1}^{T}2\eta_l\mathbb{E}(\mathcal{E}(f_l)-\mathcal{E}(f_*))$$
$$\leq\frac{\|f_*\|_K^2}{T}+4\eta_1\mathcal{A}(f_*)T^{-\theta}\log T$$
$$+\eta_1^2 2\kappa^2(a_V\tilde{B}+b_V)T^{-\min(2\theta,1)}\log T. \tag{31}$$

To bound the last term of (23), we let $1\leq k\leq t-1$ and $i\in\{t-k,\ldots,t\}$. Note that f_i depends only on $z_1,\ldots,z_{i-1}$ when $i>1$. We apply Lemma 3 with $f=f_{t-k}$, and then take the expectation on both sides of (19) to derive

$$2\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]$$
$$\leq\mathbb{E}\big[\|f_i-f_{t-k}\|_K^2-\|f_{i+1}-f_{t-k}\|_K^2\big]$$
$$+\eta_i^2\kappa^2(a_V\mathbb{E}[\mathcal{E}(f_i)]+b_V).$$

Summing up over $i=t-k,\ldots,t$

$$\sum_{i=t-k}^{t}2\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]\leq\kappa^2\sum_{i=t-k}^{t}\eta_i^2(a_V\mathbb{E}[\mathcal{E}(f_i)]+b_V).$$

Note that the left-hand side is exactly $\sum_{i=t-k+1}^{t} \eta_i \mathbb{E}[\mathcal{E}(f_i) - \mathcal{E}(f_{t-k})]$. We thus know that

$$\sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k+1}^{t} \eta_i \mathbb{E}[\mathcal{E}(f_i) - \mathcal{E}(f_{t-k})]$$
$$\leq \frac{\kappa^2}{2} \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t} \eta_i^2 (a_V \mathbb{E}[\mathcal{E}(f_i)] + b_V)$$
$$\leq \frac{\kappa^2}{2} \Big(a_V \sup_{1 \leq i \leq t} \mathbb{E}[\mathcal{E}(f_i)] + b_V\Big) \sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k}^{t} \eta_i^2.$$

With $\eta_t = \eta_1 t^{-\theta}$, by using Lemma 5, this can be relaxed as

$$\sum_{k=1}^{t-1} \frac{1}{k(k+1)} \sum_{i=t-k+1}^{t} \eta_i \mathbb{E}[\mathcal{E}(f_i) - \mathcal{E}(f_{t-k})]$$
$$\leq \eta_1^2 \kappa^2 t^{-\min(2\theta, 1)} \log(et) (a_V \sup_{1 \leq i \leq t} \mathbb{E}[\mathcal{E}(f_i)] + b_V). \quad (32)$$

Introducing (31) and (32) into (23), plugging with (20), and dividing both sides by $2\eta_T = 2\eta_1 T^{-\theta}$, one can prove the desired result with $\bar{B}_1 = 2\eta_1 \kappa^2 (a_V \tilde{B} + b_V)$. □

V. Numerical Simulations

The simplest case to implement online learning algorithm (3) is when $X = \mathbb{R}^d$ for some $d \in \mathbb{N}$ and K is the linear kernel given by $K(x, w) = w^T x$. In this case, it is straightforward to see that $f_{t+1}(x) = w_{t+1}^\top x$ with $w_1 = 0 \in \mathbb{R}^d$ and

$$w_{t+1} = w_t - \eta_t V'_-(y_t, w_t^\top x_t) x_t, \quad t = 1, \ldots, T.$$

For a general kernel, by induction, it is easy to see that $f_{t+1}(x) = \sum_{j=1}^{T} c_{t+1}^j K(x, x_j)$ with

$$c_{t+1} = c_t - \eta_t V'_- \left(y_t, \sum_{j=1}^{T} c_t^j K(x_t, x_j) \right) \mathbf{e}_t, \quad t = 1, \ldots, T$$

for $c_1 = 0 \in \mathbb{R}^T$. Here, $c_t = (c_t^1, \ldots, c_t^T)^\top$ for $1 \leq t \leq T$, and $\{\mathbf{e}_1, \ldots, \mathbf{e}_T\}$ is a standard basis of $\mathbb{R}^T$. Indeed, it is straightforward to check by induction that

$$f_{t+1} = \sum_{j=1}^{T} c_t^j K_{x_j} - \eta_t V'_-(y_t, f_t(x_t)) K_{x_t}$$
$$= \sum_{j=1}^{T} K_{x_j} \big(c_t^j - \eta_t V'_-(y_j, f_t(x_j)) \mathbf{e}_t^j\big).$$

To see how the step-size decaying rate indexed by θ affects the performance of the studied algorithm, we carry out simple numerical simulations on the *Adult*[1] data set with the hinge loss and the Gaussian kernel with kernel width $\sigma = 4$. We consider a subset of *Adult* with $T = 1000$, and run the algorithm for different θ values with $\eta_1 = 1/4$. The test and training errors (with respect to the hinge loss) for different θ values are shown in Fig. 1. We see that the minimal test error (with respect to the hinge loss) is achieved at some $\theta^* < 1/2$,

[1]The data set can be downloaded from archive.ics.uci.edu/ml and www.csie.ntu.edu.tw/cjlin/libsvmtools/

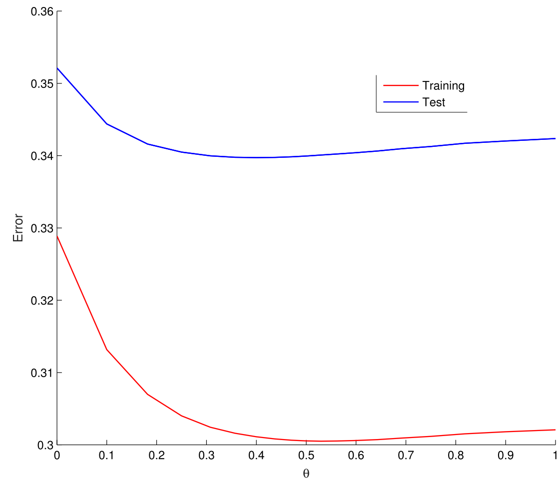

Fig. 1. Test and training errors for online learning with different θ values on *Adult* ($T = 1000$).

TABLE I
Comparison of Online Learning Using Cross Validation With LIBSVM

Algorithm	test classification error	training time
online learning	$16.2 \pm 0.2\%$	5.4 ± 0.3
LIBSVM	$18.7 \pm 0.0\%$	5.8 ± 0.5

which complements our obtained results. We also compare the performance of online learning algorithm (3) in terms of test error and training time with that of LIBSVM, a state-of-the-art batch learning algorithm for classification [30]. The test classification error and training time, for the online learning algorithm using cross validation (for choosing the best θ) and LIBSVM, are summarized in Table I, from which we see that the online learning algorithm is comparable to LIBSVM on both test error and running time.

Appendix

In this appendix, we prove the lemmas stated before.

Proof of Lemma 1: Since f_{t+1} is given by (3), by expanding the inner product, we have

$$\|f_{t+1} - f\|_K^2 = \|f_t - f\|_K^2 + \eta_t^2 \|V'_-(y_t, f_t(x_t)) K_{x_t}\|_K^2$$
$$+ 2\eta_t V'_-(y_t, f_t(x_t)) \langle K_{x_t}, f - f_t \rangle_K.$$

Observe that $\|K_{x_t}\|_K = (K(x_t, x_t))^{1/2} \leq \kappa$ and that

$$\|f\|_\infty \leq \kappa \|f\|_K, \quad \forall f \in \mathcal{H}_K.$$

These together with the incremental condition (5) yield

$$\|V'_-(y_t, f_t(x_t)) K_{x_t}\|_K$$
$$\leq \kappa |V'_-(y_t, f_t(x_t))|$$
$$\leq \kappa c_q (1 + |f_t(x_t)|^q) \leq \kappa c_q \big(1 + \kappa^q \|f_t\|_K^q\big).$$

Therefore, $\|f_{t+1} - f\|_K^2$ is bounded by

$$\|f_t - f\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t V'_-(y_t, f_t(x_t)) \langle K_{x_t}, f - f_t \rangle_K.$$

Using the reproducing property, we get

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2$$
$$+ 2\eta_t V'_-(y_t, f_t(x_t)) (f(x_t) - f_t(x_t)). \quad (33)$$

Since $V(y_t, \cdot)$ is a convex function, we have

$$V'_-(y_t, a)(b-a) \le V(y_t, b) - V(y_t, a), \quad \forall a, b \in \mathbb{R}.$$

Using this relation to (33), we get our desired result.

In order to prove Lemma 2, we first bound the learning sequence uniformly as follows.

Lemma 7: Under Assumption 1.a, let $0 \le \theta < 1$ satisfy $\theta \ge \frac{q}{q+1}$ and $\eta_t = \eta_1 t^{-\theta}$ with η_1 satisfying

$$0 < \eta_1 \le \min\left\{\frac{\sqrt{1-\theta}}{\sqrt{8}c_q(\kappa+1)^{q+1}}, \frac{1-\theta}{4|V|_0}\right\}. \tag{34}$$

Then, for $t = 1, \ldots, T-1$

$$\|f_{t+1}\|_K \le t^{\frac{1-\theta}{2}}. \tag{35}$$

Proof: We prove our statement by induction.

Taking $f = 0$ in Lemma 1, we know that

$$\begin{aligned}\|f_{t+1}\|_K^2 &\le \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t[V(y_t, 0) - V(y_t, f_t(x_t))] \\ &\le \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t |V|_0. \end{aligned}\tag{36}$$

Since $f_1 = 0$, G_1 is given by (15) and by (34), $\eta_1^2 c_q^2 \kappa^2 + 2\eta_1 |V|_0 \le 1$, we thus get (35) for the case $t = 1$.

Now, assume $\|f_t\|_K \le (t-1)^{(1-\theta)/2}$ with $t \ge 2$. Then

$$\begin{aligned}G_t^2 &\le \kappa^2 c_q^2 (1+\kappa^q)^2 \max\left(1, \|f_t\|_K^{2q}\right) \\ &\le 4c_q^2(\kappa+1)^{2q+2}(t-1)^{(1-\theta)q}\end{aligned}\tag{37}$$

where for the last inequality, we used $\kappa \le \kappa + 1$ and $1 + \kappa^q \le 2(\kappa+1)^q$. Hence, using (36)

$$\begin{aligned}&\|f_{t+1}\|_K^2 \\ &\le (t-1)^{1-\theta} + \eta_1^2 t^{-2\theta} 4c_q^2(\kappa+1)^{2q+2} t^{(1-\theta)q} + 2\eta_1 t^{-\theta}|V|_0 \\ &= t^{1-\theta}\left\{\left(1 - \frac{1}{t}\right)^{1-\theta} + \frac{\eta_1^2 4c_q^2(\kappa+1)^{2q+2}}{t^{(q+1)\theta+1-q}} + \frac{2\eta_1|V|_0}{t}\right\}.\end{aligned}$$

Since $(1-(1/t))^{1-\theta} \le 1 - (1-\theta)/t$ and the condition $\theta \ge q/(q+1)$ implies $(q+1)\theta + 1 - q \ge 1$, we see that $\|f_{t+1}\|_K^2$ is bounded by

$$t^{1-\theta}\left\{1 - \frac{1-\theta}{t} + \frac{\eta_1^2 4c_q^2(\kappa+1)^{2q+2}}{t} + \frac{2\eta_1|V|_0}{t}\right\}.$$

Finally, we use the restriction (34) for η_1 and find $\|f_{t+1}\|_K^2 \le t^{1-\theta}$. This completes the induction procedure and proves our conclusion. □

Now, we are ready to prove Lemma 2.

Proof of Lemma 2: Recall an iterative relation (25) of error terms in the proof of Theorem 5. It follows from $\mathcal{E}(f_t) \ge \mathcal{E}(f_\rho^V)$ that

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] &\le \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f_*\|_K^2\right] \\ &\quad + \eta_t^2 \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[G_t^2\right] + 2\eta_t \mathcal{A}(f_*).\end{aligned}\tag{38}$$

Since G_t is given by (15), applying Schwarz's inequality

$$\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[G_t^2\right] \le 2\kappa^2 c_q^2\left(1 + \kappa^{2q}\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^{2q}\right]\right).$$

If $q \le 1$, using Hölder's inequality

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^{2q}\right] &\le \left(\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^2\right]\right)^q \\ &\le 1 + \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^2\right].\end{aligned}$$

If $q > 1$, noting that (9) implies (34), we have (35) and thus

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^{2q}\right] &\le \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^2\right] t^{(q-1)(1-\theta)} \\ &= \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^2\right] t^{2\theta - q^*}.\end{aligned}$$

Combining the above-mentioned two cases yields

$$\begin{aligned}&\eta_t^2 \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[G_t^2\right] \\ &\quad \le 2\kappa^2 c_q^2 \eta_t^2\left(1 + \kappa^{2q}\left(1 + \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t\|_K^2\right]\right)t^{2\theta-q^*}\right) \\ &\quad \le 2\kappa^2 c_q^2 \eta_t^2\left(1 + \kappa^{2q} t^{2\theta - q^*}\right. \\ &\qquad \left.\cdot\left(1 + 2\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f^*\|_K^2\right] + 2\|f_*\|_K^2\right)\right) \\ &\quad \le C_1\left(1 + \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f^*\|_K^2\right] + \|f_*\|_K^2\right)t^{-q^*}\end{aligned}\tag{39}$$

where

$$C_1 = 4\eta_1^2 c_q^2 (1+\kappa)^{2q+2}. \tag{40}$$

Putting (39) into (38) yields

$$\begin{aligned}&\mathbb{E}_{z_1,\ldots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\ &\quad \le \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f_*\|_K^2\right] + 2\eta_1 t^{-\theta}\mathcal{A}(f_*) \\ &\qquad + C_1\left(1 + \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|f_t - f^*\|_K^2\right] + \|f_*\|_K^2\right)t^{-q^*}.\end{aligned}$$

Applying this inequality iteratively, with $f_1 = 0$, we derive

$$\begin{aligned}&\mathbb{E}_{z_1,\ldots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\ &\quad \le \|f_*\|_K^2 + 2\eta_1 \mathcal{A}(f_*)\sum_{j=1}^t j^{-\theta} \\ &\qquad + C_1\left(1 + \|f_*\|_K^2\right. \\ &\qquad \left. + \max_{j=1,\ldots,t} \mathbb{E}_{z_1,\ldots,z_{j-1}}\left[\|f_j - f^*\|_K^2\right]\right)\sum_{j=1}^t j^{-q^*}.\end{aligned}$$

Note that $\theta \in (1/2, 1)$ and that from the restriction on θ, $q^* > 1$. Applying the elementary inequality (26) to bound $\sum_{j=1}^t j^{-q^*}$ and $\sum_{j=1}^t j^{-\theta}$, we get

$$\begin{aligned}&\mathbb{E}_{z_1,\ldots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\ &\le \|f_*\|_K^2 + \frac{2\eta_1}{1-\theta}\mathcal{A}(f_*)t^{1-\theta} \\ &\quad + \frac{C_1 q^*}{q^*-1}\left(1 + \|f_*\|_K^2 + \max_{j=1,\ldots,t} \mathbb{E}_{z_1,\ldots,z_{j-1}}\left[\|f_j - f^*\|_K^2\right]\right).\end{aligned}$$

Now, we derive upper bounds for $\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1} - f_*\|_K^2]$ by induction for $t = 1, \ldots, T-1$. Assume that $\mathbb{E}_{z_1,\ldots,z_{j-1}}[\|f_j - f_*\|_K^2] \le 2(\|f_*\|_K^2 + \mathcal{A}(f_*)(j-1)^{1-\theta} + 1)$ holds for

$j = 1, \dots, t$. Then

$$
\begin{aligned}
&\mathbb{E}_{z_1,\dots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\
&\quad \le \|f_*\|_K^2 + \frac{C_1 q^*}{q^*-1}(3 + 3\|f_*\|_K^2 + 2\mathcal{A}(f_*)t^{1-\theta}]) \\
&\qquad + \frac{2\eta_1}{1-\theta}\mathcal{A}(f_*)t^{1-\theta} \\
&\quad \le \left(1 + \frac{3C_1 q^*}{q^*-1}\right)(1 + \|f_*\|_K^2) \\
&\qquad + \left(\frac{2C_1 q^*}{q^*-1} + \frac{2\eta_1}{1-\theta}\right)\mathcal{A}(f_*)t^{1-\theta}.
\end{aligned}
$$

Recall that C_1 is given by (40). We see from (9) that $3C_1 q^*/(q^*-1) \le 1-\theta \le 1$ and $2\eta_1/(1-\theta) \le 1$. It follows that

$$
\mathbb{E}_{z_1,\dots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \le 2\left(\|f_*\|_K^2 + \mathcal{A}(f_*)t^{1-\theta} + 1\right). \quad (41)
$$

From the above-mentioned induction procedure, we conclude that for $t = 1, \dots, T-1$, the bound (41) holds, which leads to the desired bound (17) using $\|f_t\|_K^2 \le 2\|f_t - f_*\|_K^2 + 2\|f_*\|_K^2$. Applying (41) into (39), and noting that $C_1 \le 1$ by the restriction (9), we get the other desired bound (18). The proof is complete.

Proof of Lemma 3: Following the proof of Lemma 1, we have:

$$
\begin{aligned}
\|f_{t+1} - f\|_K^2 &\le \|f_t - f\|_K^2 + \eta_t^2\kappa^2|V_-(y_t, f_t(x_t))|^2 \\
&\quad + 2\eta_t\left[V(y_t, f(x_t)) - V(y_t, f_t(x_t))\right].
\end{aligned}
$$

Applying Assumption 1.b to the above, we get the desired result.

Proof of Lemma 4: The proof is divided into several steps.

Basic Decomposition: We choose $\mu_t = \eta_t\mathbb{E}[\mathcal{E}(f_t)]$ in Lemma 6 to get

$$
\begin{aligned}
&\eta_t\mathbb{E}[\mathcal{E}(f_t)] \\
&\quad = \frac{1}{t}\sum_{i=1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)] \\
&\qquad + \sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k+1}^{t}(\eta_i\mathbb{E}[\mathcal{E}(f_i)] - \eta_{t-k}\mathbb{E}[\mathcal{E}(f_{t-k})]).
\end{aligned}
$$

Since $\{\eta_t\}_t$ is decreasing and $\mathbb{E}[\mathcal{E}(f_{t-k})]$ is nonnegative, the above can be relaxed as

$$
\begin{aligned}
\eta_t\mathbb{E}[\mathcal{E}(f_t)] &\le \frac{1}{t}\sum_{i=1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)] \\
&\quad + \sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k+1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i) - \mathcal{E}(f_{t-k})].
\end{aligned} \quad (42)
$$

In the rest of the proof, we will bound the last two terms in the above-mentioned estimate.

Bounding the Average: To bound the first term on the right-hand side of (42), we apply (30) with $f = 0$ to get

$$
\begin{aligned}
\mathbb{E}\left[\|f_{l+1}\|_K^2\right] &\le \mathbb{E}\left[\|f_l\|_K^2\right] + \eta_l^2\kappa^2(a_V\mathbb{E}[\mathcal{E}(f_l)] + b_V) \\
&\quad + 2\eta_l\mathbb{E}(\mathcal{E}(0) - \mathcal{E}(f_l)).
\end{aligned}
$$

Rearranging terms, and using the fact that $\mathcal{E}(0) \le |V|_0$

$$
\begin{aligned}
&\eta_l(2 - a_V\eta_l\kappa^2)\mathbb{E}[\mathcal{E}(f_l)] \\
&\quad \le \mathbb{E}[\|f_l\|_K^2 - \|f_{l+1}\|_K^2] + b_V\eta_l^2\kappa^2 + 2\eta_l|V|_0.
\end{aligned}
$$

It thus follows from $a_V\eta_l\kappa^2 \le 1$, implied by (11), that

$$
\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \mathbb{E}\left[\|f_l\|_K^2 - \|f_{l+1}\|_K^2\right] + b_V\eta_l^2\kappa^2 + 2\eta_l|V|_0. \quad (43)
$$

Summing up over $l = 1, \dots, t$, introducing $f_1 = 0$, $\|f_{t+1}\|_K^2 \ge 0$, and then multiplying both sides by $1/t$, we get

$$
\frac{1}{t}\sum_{l=1}^{t}\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \frac{1}{t}\sum_{l=1}^{t}\left(b_V\eta_l^2\kappa^2 + 2\eta_l|V|_0\right).
$$

Since $\eta_t = \eta_1 t^{-\theta}$, we have

$$
\frac{1}{t}\sum_{l=1}^{t}\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \left(b_V\eta_1^2\kappa^2 + 2\eta_1|V|_0\right)\frac{1}{t}\sum_{l=1}^{t}l^{-\theta}.
$$

Using (26), we get

$$
\frac{1}{t}\sum_{l=1}^{t}\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \frac{b_V\eta_1^2\kappa^2 + 2\eta_1|V|_0}{1-\theta}t^{-\theta}. \quad (44)
$$

Bounding the Moving Average: To bound the last term of (42), we let $1 \le k \le t-1$ and $i \in \{t-k, \dots, t\}$. Recall the inequality (32) in the proof of Theorem 6. Applying the basic inequality $e^{-x} \le (ex)^{-1}, x > 0$, which implies $t^{-\min(\theta, 1-\theta)}\log(et) \le (1/\min(\theta, 1-\theta))$, we see that the last term of (42) can be upper bounded by

$$
\frac{\eta_1^2\kappa^2}{\min(\theta, 1-\theta)}t^{-\theta}\left(a_V\sup_{1\le i\le t}\mathbb{E}[\mathcal{E}(f_i)] + b_V\right).
$$

Induction: Introducing (32) and (44) into the decomposition (42), and then dividing both sides by $\eta_t = \eta_1 t^{-\theta}$, we get

$$
\mathbb{E}[\mathcal{E}(f_t)] \le A\sup_{1\le i\le t}\mathbb{E}[\mathcal{E}(f_i)] + B \quad (45)
$$

where we set $A = (\eta_1 a_V\kappa^2/\min(\theta, 1-\theta))$ and

$$
B = \frac{b_V\eta_1\kappa^2 + 2|V|_0}{1-\theta} + \frac{\eta_1 b_V\kappa^2}{\min(\theta, 1-\theta)}.
$$

The restriction (11) on η_1 tells us that $A \le 1/2$. Then, using (45) with an inductive argument, we find that for all $t \le T$

$$
\mathbb{E}[\mathcal{E}(f_t)] \le 2B \quad (46)
$$

which leads to the desired result with $\tilde{B} = 2B$. In fact, the case $t = 2$ can be verified directly from (43), by plugging with $f_1 = 0$. Now, assume that (46) holds for any $k \le t-1$, where $t \ge 3$. Under this hypothesis condition, if $\mathbb{E}[\mathcal{E}(f_t)] \le \sup_{1\le i\le t-1}\mathbb{E}[\mathcal{E}(f_i)]$, then using the hypothesis condition, we know that $\mathbb{E}[\mathcal{E}(f_t)] \le 2B$. If $\mathbb{E}[\mathcal{E}(f_t)] \ge \sup_{1\le i\le t-1}\mathbb{E}[\mathcal{E}(f_i)]$, we use (45) to get

$$
\mathbb{E}[\mathcal{E}(f_t)] \le A\mathbb{E}[\mathcal{E}(f_t)] + B \le \mathbb{E}[\mathcal{E}(f_t)]/2 + B
$$

which implies $\mathbb{E}[\mathcal{E}(f_t)] \le 2B$. The proof is thus complete.

Proof of Lemma 5: Exchanging the order in the sum, we have

$$\begin{aligned}
&\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \\
&= \sum_{t=1}^{T-1} \sum_{k=T-t}^{T-1} \frac{1}{k(k+1)} t^{-q^*} + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} T^{-q^*} \\
&= \sum_{t=1}^{T-1} \left(\frac{1}{T-t} - \frac{1}{T} \right) t^{-q^*} + \left(1 - \frac{1}{T}\right) T^{-q^*} \\
&\leq \sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*}.
\end{aligned}$$

What remains is to estimate the term $\sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*}$. Note that

$$\sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*} = \sum_{t=1}^{T-1} \frac{t^{1-q^*}}{(T-t)t} \leq T^{\max(1-q^*,0)} \sum_{t=1}^{T-1} \frac{1}{(T-t)t}$$

and that by (26)

$$\begin{aligned}
\sum_{t=1}^{T-1} \frac{1}{(T-t)t} &= \frac{1}{T} \sum_{t=1}^{T-1} \left(\frac{1}{T-t} + \frac{1}{t} \right) \\
&= \frac{2}{T} \sum_{t=1}^{T-1} \frac{1}{t} \leq \frac{2}{T} \log(eT).
\end{aligned}$$

From the above-mentioned analysis, we see the first statement of the lemma.

To prove the second part of the lemma, we split the term $\sum_{t=1}^{T-1} 1/(T-t) t^{-q^*}$ into two parts

$$\begin{aligned}
&\sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*} \\
&= \sum_{T/2 \leq t \leq T-1} \frac{1}{T-t} t^{-q^*} + \sum_{1 \leq t < T/2} \frac{1}{T-t} t^{-q^*} \\
&\leq 2^{q^*} T^{-q^*} \sum_{T/2 \leq t \leq T-1} \frac{1}{T-t} + 2T^{-1} \sum_{1 \leq t < T/2} t^{-q^*} \\
&= 2^{q^*} T^{-q^*} \sum_{1 \leq t \leq T/2} t^{-1} + 2T^{-1} \sum_{1 \leq t < T/2} t^{-q^*}.
\end{aligned}$$

Applying (26) to the above and then using $T^{-q^*+1} \log T \leq 1/(2(q^* - 1))$, we see the second statement of Lemma 5.

Proof of Lemma 6: For $k = 1, \ldots, T-1$

$$\begin{aligned}
&\frac{1}{k} \sum_{j=T-k+1}^{T} u_j - \frac{1}{k+1} \sum_{j=T-k}^{T} u_j \\
&= \frac{1}{k(k+1)} \left\{ (k+1) \sum_{j=T-k+1}^{T} u_j - k \sum_{j=T-k}^{T} u_j \right\} \\
&= \frac{1}{k(k+1)} \sum_{j=T-k+1}^{T} (u_j - u_{T-k}).
\end{aligned}$$

Summing over $k = 1, \ldots, T-1$, and rearranging terms, we get (21).

Acknowledgment

The authors would like to thank the reviewers and editors for their constructive comments, and Dr. Raffaello Camoriano and Prof. Lorenzo Rosasco for sharing codes for our simulations.

References

[1] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[2] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.

[3] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of on-line learning algorithms," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2050–2057, Sep. 2004.

[4] S. Smale and Y. Yao, "Online learning algorithms," *Found. Comput. Math.*, vol. 6, no. 2, pp. 145–170, 2006.

[5] Y. Ying and D. X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[6] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 773–781.

[7] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 71–79.

[8] G. B. Ye and D. X. Zhou, "Fully online classification by regularization," *Appl. Comput. Harmon. Anal.*, vol. 23, no. 2, pp. 198–214, 2007.

[9] P. Tarres and Y. Yao, "Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5716–5735, Sep. 2014.

[10] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar, "Information-theoretic lower bounds on the oracle complexity of convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1–9.

[11] T. Hu, J. Fan, Q. Wu, and D. X. Zhou, "Regularization schemes for minimum error entropy principle," *Anal. Appl.*, vol. 13, no. 4, pp. 437–455, 2015.

[12] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 817–824.

[13] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri, "Some properties of regularized kernel methods," *J. Mach. Learn. Res.*, vol. 5, pp. 1363–1390, Oct. 2004.

[14] Y. Ying and D. X. Zhou, "Unregularized online learning algorithms with general loss functions," *Appl. Comput. Harmon. Anal.*, vol. 42, pp. 224–244, Aug. 2017.

[15] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, no. 1, pp. 56–85, 2004.

[16] N. Srebro, K. Sridharan, and A. Tewari. (Sep. 2010). "Optimistic rates for learning with a smooth loss." [Online]. Available: https://arxiv.org/abs/1009.3896

[17] S. Smale and D. X. Zhou, "Estimating the approximation error in learning theory," *Anal. Appl.*, vol. 1, no. 1, pp. 17–41, 2003.

[18] Y. Ying and M. Pontil, "Online gradient descent learning algorithms," *Found. Comput. Math.*, vol. 8, no. 5, pp. 561–596, 2008.

[19] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[20] D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou, "Support vector machine soft margin classifiers: Error analysis," *J. Mach. Learn. Res.*, vol. 5, pp. 1143–1175, 2004.

[21] G. Wahba, *Spline Models for Observational Data*, vol. 59. Philadelphia, PA, USA: SIAM, 1990.

[22] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Inf. Comput.*, vol. 132, no. 1, pp. 1–63, 1997.

[23] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 928–936.

[24] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: A meta-algorithm and applications.," *Theory Comput.*, vol. 8, no. 1, pp. 121–164, 2012.

[25] S. Smale and D.-X. Zhou, "Online learning with Markov sampling," *Anal. Appl.*, vol. 7, no. 1, pp. 87–113, 2009.

[26] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu, "The generalization ability of online SVM classification based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 628–639, Mar. 2015.

[27] Z. Hu, M. Lin, and C. Zhang, "Dependent online kernel learning with constant number of random Fourier features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2464–2476, Oct. 2015.

[28] J. Lin, L. Rosasco, and D.-X. Zhou, "Iterative regularization for learning with convex loss functions," *J. Mach. Learn. Res.*, vol. 17, no. 77, pp. 1–38, 2016.

[29] J. Lin and D.-X. Zhou, "Learning theory of randomized Kaczmarz algorithm," *J. Mach. Learn. Res.*, vol. 16, pp. 3341–3365, Jan. 2015.

[30] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

Junhong Lin received the Ph.D. degree in applied mathematics from Zhejiang University, Hangzhou, China, in 2013.

From 2013 to 2015, he was a Post-Doctoral Fellow with the Department of Mathematics, City University of Hong Kong, Hong Kong. He is currently with the Laboratory for Computational and Statistical Learning, Massachusetts Institute of Technology, Cambridge, MA, USA, and also with the Istituto Italiano di Tecnologia, Genoa, Italy. His current research interests include compressed sensing and learning theory.

Ding-Xuan Zhou received the B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 1988 and 1991, respectively.

He joined the Faculty of the City University of Hong Kong, Hong Kong, in 1996, where he is currently a Chair Professor of the Department of Mathematics. He has authored over 100 research papers. His current research interests include learning theory, data science, wavelet analysis, and approximation theory.

Dr. Zhou is serving on the Editorial Board of over 10 international journals, and is an Editor-in-Chief of the *Journal Analysis and Application*. He received a Joint Research Fund from the National Science Fund of China for Distinguished Young Scholars in 2005 and the Humboldt Research Fellowship in 1993, and was rated in 2014, 2015, and 2016 by Thomson Reuters as a Highly-cited Researcher. He has co-organized over 20 international conferences and conducted over 20 research grants.

Online Learning Algorithms Can Converge Comparably Fast as Batch Learning

Junhong Lin and Ding-Xuan Zhou

***Abstract*— Online learning algorithms in a reproducing kernel Hilbert space associated with convex loss functions are studied. We show that in terms of the expected excess generalization error, they can converge comparably fast as corresponding kernel-based batch learning algorithms. Under mild conditions on loss functions and approximation errors, fast learning rates and finite sample upper bounds are established using polynomially decreasing step-size sequences. For some commonly used loss functions for classification, such as the logistic and the p-norm hinge loss functions with $p \in [1, 2]$, the learning rates are the same as those for Tikhonov regularization and can be of order $O(T^{-(1/2)} \log T)$, which are nearly optimal up to a logarithmic factor. Our novelty lies in a sharp estimate for the expected values of norms of the learning sequence (or an inductive argument to uniformly bound the expected risks of the learning sequence in expectation) and a refined error decomposition for online learning algorithms.**

***Index Terms*— Approximation error, learning theory, online learning, reproducing kernel Hilbert space (RKHS).**

I. Introduction

NONPARAMETRIC regression or classification aims at learning predictors from samples. To measure the performance of a predictor, one may use a loss function and its induced generalization error. Given a prediction function $f : X \to \mathbb{R}$, defined on a separable metric space X (input space), a loss function $V : \mathbb{R}^2 \to \mathbb{R}_+$ gives a local error $V(y, f(x))$ at $(x, y) \in Z := X \times Y$ with an output space $Y \subseteq \mathbb{R}$. The *generalization error* $\mathcal{E} = \mathcal{E}^V$ associated with the loss V and a Borel probability measure ρ on Z, defined as

$$\mathcal{E}(f) = \int_Z V(y, f(x)) d\rho,$$

measures the performance of f.

Kernel methods provide efficient nonparametric learning algorithms for dealing with nonlinear features, where reproducing kernel Hilbert spaces (RKHSs) are often used as hypothesis spaces in the design of learning algorithms. With suitable choices of kernels, RKHSs can be used to approximate functions in $L^2_{\rho_X}$, the space of square integrable functions with respect to the marginal probability measure ρ_X. A reproducing kernel $K : X \times X \to \mathbb{R}$ is a symmetric function such that $(K(u_i, u_j))_{i,j=1}^{\ell}$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^{\ell}$ in X. The RKHS $(\mathcal{H}_K, \|\cdot\|_K)$ is the completion of the linear span of the set $\{K_x := K(x, \cdot) : x \in X\}$ with respect to the inner product given by $\langle K_x, K_u \rangle_K = K(x, u)$.

Batch learning algorithms perform learning tasks by using a whole batch of sample $\mathbf{z} = \{z_i = (x_i, y_i) \in Z\}_{i=1}^{T}$. Throughout this paper, we assume that the sample $\{z_i = (x_i, y_i)\}_i$ is drawn independently according to the measure ρ on Z. A large family of batch learning algorithms are generated by Tikhonov regularization

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t=1}^{T} V(y_t, f(x_t)) + \lambda \|f\|_K^2 \right\}, \ \lambda > 0. \quad (1)$$

Tikhonov regularization scheme (1) associated with convex loss functions has been extensively studied in the literature, and sharp learning rates have been well developed due to many results, as described in the books (see [1], [2], and references therein). But in practice, it may be difficult to implement when the sample size T is extremely large, as its standard complexity is about $O(T^3)$ for many loss functions. For example, for the hinge loss $V(y, f) = (1 - yf)_+ = \max\{1 - yf, 0\}$ or the square hinge loss $V(y, f) = (1 - yf)_+^2$ in classification corresponding to support vector machines, solving the scheme (1) is equivalent to solving a constrained quadratic program, with complexity of order $O(T^3)$.

With complexity $O(T)$ or $O(T^2)$, online learning represents an important family of efficient and scalable machine learning algorithms for large-scale applications. Over the past years, a variety of online learning algorithms have been proposed (see [3]–[7] and references therein). Most of them take the form of regularized online learning algorithms, i.e., given $f_1 = 0$,

$$f_{t+1} = f_t - \eta_t (V'_-(y_t, f_t(x_t)) K_{x_t} + \lambda_t f_t), \quad t = 1, \ldots, T-1 \quad (2)$$

where $\{\lambda_t\}$ is a regularization sequence and $\{\eta_t > 0\}$ is a step-size sequence. In particular, $\{\lambda_t\}$ is chosen as a constant sequence $\{\lambda > 0\}$ in [4] and [5] or as a time-varying regularization sequence in [8] and [9]. Throughout this paper, we assume that V is convex with respect to the second variable. That is, for any fixed $y \in Y$, the univariate function $V(y, \cdot)$ on $\mathbb{R}$ is convex. Hence, its left derivative $V'_-(y, f)$ exists at every $f \in \mathbb{R}$ and is nondecreasing.

We study the following online learning algorithm without regularization.

Manuscript received July 10, 2016; revised November 28, 2016; accepted February 17, 2017. This work was supported in part by the the Research Grants Council of Hong Kong under Project CityU 104113 and in part by the National Science Foundation of China under Grant 11461161006 and Grant 11471292. *(Corresponding author: Ding-Xuan Zhou.)*

J. Lin is with the Department of Mathematics, City University of Hong Kong, Hong Kong, also with the Laboratory for Computational and Statistical Learning, Massachusetts Institute of Technology, Cambridge, MA, USA, and also with the Istituto Italiano di Technologia, Genoa, Italy (e-mail: jhlin5@hotmail.com).

D.-X. Zhou is with the Department of Mathematics, City University of Hong Kong, Hong Kong (e-mail: mazhou@cityu.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2677970

Definition 1: The *online learning algorithm* without regularization associated with the loss V and the kernel K is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t V'_-(y_t, f_t(x_t))K_{x_t}, \quad t = 1, \ldots, T-1 \quad (3)$$

where $\{\eta_t > 0\}$ is a step-size sequence.

Let f_ρ^V be a minimizer of the generalization error $\mathcal{E}(f)$ among all measurable functions $f : X \to Y$. The main purpose of this paper is to estimate the expected excess generalization error $\mathbb{E}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$, where f_T is generated by the unregularized online learning algorithm (3) with a convex loss V. Under a mild condition on approximation errors and a growth condition on the loss V, we derive upper bounds for the expected excess generalization error using polynomially decaying step-size sequences. Our bounds are independent of the capacity of the RKHS $\mathcal{H}_K$, and are comparable to those for Tikhonov regularization (1), see more details in Section III. In particular, for some loss functions, such as the logistic loss, the p-absolute value loss, and the p-hinge loss with $p \in [1, 2]$, our learning rates are of order $O(T^{-(1/2)} \log T)$, which is nearly optimal in the sense that up to a logarithmic factor, it matches the minimax rates of order $O(T^{-(1/2)})$ in [10] for stochastic approximation in the nonstrongly convex case. In our approach, an inductive argument is involved, to develop sharp estimates for the expected values of $\|f_t\|_K^2$, which is better than uniform bounds in the existing literature, or to bound the expected values of $\mathcal{E}(f_t)$ uniformly. Our second novelty is a refined error decomposition, which might be used for other online or gradient descent algorithms [11], [12] and is of independent interest.

The rest of this paper is organized as follows. We introduce in Section II some basic assumptions that underlie our analysis, and give our main results as well as examples, illustrating our upper bounds for the expected excess generalization error for different kinds of loss functions in learning theory. Section III contributes to discussions and comparisons with previous results, mainly on online learning algorithms with or without regularization, and the common Tikhonov regularization batch learning algorithms. Section IV deals with the proof of our main results, which relies on an error decomposition as well as the lemmas proved in the Appendix. Finally, in Section V, we will discuss the numerical simulation of the studied algorithms, and give some numerical simulations, which complements our theoretical results.

II. Main Results

In this section, we first state our main assumptions, following with some comments. We then present our main results with simple discussions.

A. Assumptions on the Kernel and Loss Function

Throughout this paper, we assume that the kernel is bounded on $X \times X$ with the constant

$$\kappa = \sup_{x \in X} \max(\sqrt{K(x,x)}, 1) < \infty \quad (4)$$

and that $|V|_0 := \sup_{y \in Y} V(y, 0) < \infty$. These bounded conditions on K and V are common in learning theory. They are satisfied when X is compact and Y is a bounded subset of $\mathbb{R}$. Moveover, the condition $|V|_0 < \infty$ implies that $\mathcal{E}(f_\rho^V)$ is finite

$$\mathcal{E}(f_\rho^V) \le \mathcal{E}(0) = \int_Z V(y, 0) d\rho \le |V|_0.$$

The assumption on the loss function V is a growth condition for its left derivative $V'_-(y, \cdot)$.

Assumption 1.a: Assume that for some $q \ge 0$ and constant $c_q > 0$, there holds

$$|V'_-(y, f)| \le c_q(1 + |f|^q), \quad \forall f \in \mathbb{R}, y \in Y. \quad (5)$$

The growth condition (5) is implied by the requirement for the loss function to be Nemitiski [2], [13]. It is weaker than, either assuming the loss or its gradient, to be Lipschitz in its second variable as often done in learning theory, or assuming the loss to be α-activating with $\alpha \in (0, 1]$ in [14].

An alterative to Assumption 1.a made for V in the literature is the following assumption [15], [16].

Assumption 1.b: Assume that for some $a_V, b_V \ge 0$, there holds

$$|V'_-(y, f)|^2 \le a_V V(y, f) + b_V, \quad \forall f \in \mathbb{R}, y \in Y. \quad (6)$$

Assumption 1.b is satisfied for most loss functions commonly used in learning theory, when Y is a bounded subset of $\mathbb{R}$. In particular, when $V(y, \cdot)$ is smooth, it is satisfied with $b_V = 0$ and some appropriate a_V [16, Lemma 2.1].

B. Assumption on the Approximation Error

The performance of online learning algorithm (3) depends on how well the target function f_ρ^V can be approximated by functions from the hypothesis space $\mathcal{H}_K$. For our purpose of estimating the excess generalization error, the approximation is measured by $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ with $f \in \mathcal{H}_K$. Moreover, the output function f_T produced by the online learning algorithm lies in a ball of $\mathcal{H}_K$ with the radius increasing with T (as shown in Lemma 7). So we measure the approximation ability of the hypothesis space $\mathcal{H}_K$ with respect to the generalization error $\mathcal{E}(f)$ and f_ρ^V by penalizing the functions with their norm squares [17] as follows.

Definition 2: The approximation error associated with the triplet (ρ, V, K) is defined by

$$\mathcal{D}(\lambda) = \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) - \mathcal{E}(f_\rho^V) + \lambda \|f\|_K^2 \right\}, \quad \lambda > 0. \quad (7)$$

When $f_\rho^V \in \mathcal{H}_K$, we can take $f = f_\rho^V$ in (7) and find $\mathcal{D}(\lambda) \le \|f_\rho^V\|_K^2 \lambda = O(\lambda)$. When $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ can be arbitrarily small as f runs over $\mathcal{H}_K$, we know that $\mathcal{D}(\lambda) \to 0$ as $\lambda \to 0$. To derive explicit convergence rates for the studied online algorithm, we make the following assumption on the decay of the approximation error to be $O(\lambda^\beta)$.

Assumption 3: Assume that for some $\beta \in (0, 1]$ and $c_\beta > 0$, the approximation error satisfies

$$\mathcal{D}(\lambda) \le c_\beta \lambda^\beta, \quad \forall \lambda > 0. \quad (8)$$

C. Alternative Conditions on the Approximation Error

Assumption (8) on the approximation error is standard in analyzing both Tikhonov regularization schemes [1], [2] and online learning algorithms [8], [9], [18]. It is independent of the sample, and measures the approximation ability of the space $\mathcal{H}_K$ to f_ρ^V with respect to (ρ, V). It may be replaced by alterative simple conditions for specified loss functions.

For a Lipschitz continuous loss function meaning that

$$\sup_{y\in Y, f, f'\in\mathbb{R}} \frac{|V(y,f)-V(y,f')|}{|f-f'|} = l < \infty$$

it is easy to see that $\mathcal{E}(f)-\mathcal{E}(f_\rho^V) \le l\|f-f_\rho^V\|_{L^1_{\rho_X}}$, and thus a sufficient condition for (8) is

$$\inf_{f\in\mathcal{H}_K}\left\{\left\|f-f_\rho^V\right\|_{L^1_{\rho_X}} + \lambda\|f\|_K^2\right\} = O(\lambda^\beta).$$

In particular, for the hinge loss in classification, we have $l = 1$. Such a condition measures quantitatively the approximation of the function f_ρ^V in the space $L^1_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$, and can be characterized [2], [17] by requiring f_ρ^V to lie in some interpolation space between $\mathcal{H}_K$ and $L^1_{\rho_X}$.

For the least squares loss, $f_\rho^V = f_\rho$ and there holds $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f-f_\rho\|^2_{L^2_{\rho_X}}$. Here, f_ρ is the regression function defined at $x \in X$ to be the expectation of the conditional distribution $\rho(y|x)$ given x. In this case, condition (8) is exactly

$$\inf_{f\in\mathcal{H}_K}\left\{\left\|f-f_\rho\right\|^2_{L^2_{\rho_X}} + \lambda\|f\|_K^2\right\} = O(\lambda^\beta).$$

This condition is about the approximation of the function f_ρ in the space $L^2_{\rho_X}$ by functions from the RKHS $\mathcal{H}_K$. It can be characterized [17] by requiring that f_ρ lies in $L_K^{\beta/2}(L^2_{\rho_X})$, the range of the operator $L_K^{\beta/2}$. Recall that the integral operator $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ is defined by

$$L_K(f) = \int_X f(x)K_x d\rho_X, \quad f \in L^2_{\rho_X}.$$

Since K is a reproducing kernel with finite κ, the operator L_K is symmetric, compact, and positive, and its power $L_K^{\beta/2}$ is well defined.

D. Stating Main Results

Our first main result of this paper, to be proved in Section IV, is stated as follows.

Theorem 1: Under Assumption 1.a, let $\eta_t = \eta_1 t^{-\theta}$ with $\max((1/2), q/(q+1)) < \theta < 1$ and η_1 satisfying

$$0 < \eta_1 \le \min\left(\sqrt{\frac{(q^*-1)(1-\theta)}{12c_q^2(1+\kappa)^{2q+2}q^*}}, \frac{1-\theta}{2(1+2|V|_0)}\right) \quad (9)$$

where we denote $q^* = 2\theta - (1-\theta)\cdot\max(0, q-1) > 0$. Then

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T) - \mathcal{E}\left(f_\rho^V\right)\right\} \le \widetilde{C}\{\mathcal{D}(T^{\theta-1}) + T^{\theta-1}\} \quad (10)$$

where $\widetilde{C}$ is a positive constant depending on η_1, q, κ, and θ (independent of T and given explicitly in the proof).

Combining Theorem 1 with Assumption 3, we get the following explicit learning rates.

Corollary 2: Under the conditions of Theorem 1 and Assumption 3, we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T) - \mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-(1-\theta)\beta}).$$

Replacing Assumption 1.a by Assumption 1.b, we can relax the restriction on θ in Theorem 1 as $\theta \in (0,1)$, which thus improves the learning rates. Concretely, we have the following convergence results.

Theorem 3: Under Assumption 1.b, let $\eta_t = \eta_1 t^{-\theta}$ with $0 < \theta < 1$ and η_1 satisfying

$$0 < \eta_1 \le \frac{\min(\theta, 1-\theta)}{2a_V\kappa^2}. \quad (11)$$

Then

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T) - \mathcal{E}\left(f_\rho^V\right)\right\} \le \widetilde{C}'\{\mathcal{D}(T^{\theta-1}) + T^{-\min(\theta,1-\theta)}\}\log T \quad (12)$$

where $\widetilde{C}'$ is a positive constant depending on $\eta_1, a_V, b_V\kappa$, and θ (independent of T and given explicitly in the proof).

Corollary 4: Under the conditions of Theorem 3 and Assumption 3, let $\theta = \beta/(\beta+1)$. Then, we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T) - \mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-\frac{\beta}{\beta+1}}\log T).$$

To illustrate the above-mentioned results, we give the following examples of commonly used loss functions in learning theory with corresponding learning rates for online learning algorithms (3).

Example 1: Assume $|y| \le M$, and conditions (4) and (8) hold with $0 < \beta \le 1$. For the least squares loss $V(y,a) = (y-a)^2$, the p-norm loss $V(y,a) = |y-a|^p$ with $p \in [1,2)$, the hinge loss $V(y,a) = (1-ya)_+$, the logistic loss $V(y,a) = \log(1+e^{-ya})$, and the p-norm hinge loss $V(y,a) = ((1-ya)_+)^p$ with $p \in (1,2]$, choosing $\eta_t = \eta_1 t^{-\beta/(\beta+1)}$ with η_1 satisfying (11), we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T) - \mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-\frac{\beta}{\beta+1}}\log T)$$

which is of order $O(T^{-(1/2)}\log T)$ if $\beta = 1$.

Example 1 follows from Corollary 4, while the conclusion of the next example is seen from Corollary 2.

Example 2: Under the assumption of Example 1, for the p-norm loss $V(y,a) = |y-a|^p$ and the p-norm hinge loss $V(y,a) = ((1-ya)_+)^p$ with $p > 2$, selecting $\eta_t = \eta_1 t^{-((p-1)/p+\epsilon)}$ with $\epsilon \in (0,(1/p))$ and η_1 such that (9) holds with $q = p-1$, we have

$$\mathbb{E}_{z_1,z_2,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T) - \mathcal{E}\left(f_\rho^V\right)\right\} = O(T^{-(\frac{1}{p}-\epsilon)\beta})$$

which is of order $O(T^{\epsilon-(1/p)})$ if $\beta = 1$.

Remark 1:

1) The learning rates given in Example 1 are optimal in the sense that they are the same as those for the Tikhonov regularization [2, Ch. 7].
2) According to Example 1, the optimal learning rates are achieved when $\eta_t \simeq t^{-\beta/(1+\beta)}$. Since β is not known in general, in practice, a hold-out cross-validation method can be used to tune the ideal exponential parameter θ.
3) Our analysis can be extended to the case of constant step sizes. In fact, following our proofs given in the following, the readers can see that, when $\eta_t = T^{-\beta/(\beta+1)}$ for

$t = 1, \ldots, T-1$, the results stated in Example 1 still hold.

E. Classification Problem

The binary classification problem in learning theory is a special case of our learning problems. In this case, $Y = \{1, -1\}$. A classifier for classification is a function f from X to Y and its misclassification error $\mathcal{R}(f)$ is defined as the probability of the event $\{(x, y) \in Z : y \neq f(x)\}$ of making wrong predictions. A minimizer of the misclassification error is the Bayes rule $f_c : X \to Y$ given by

$$f_c(x) = \begin{cases} 1, & \text{if } \rho(y = 1|x) \geq 1/2 \\ -1, & \text{otherwise.} \end{cases}$$

The performance of a classification algorithm can be measured by the excess misclassification error $\mathcal{R}(f) - \mathcal{R}(f_c)$. For the online learning algorithms (3), our classifier is given by $\operatorname{sign}(f_T)$

$$\operatorname{sign}(f_T)(x) = \begin{cases} 1, & \text{if } f_T(x) \geq 0 \\ -1, & \text{otherwise.} \end{cases}$$

So our error analysis aims at the excess misclassification error

$$\mathcal{R}(\operatorname{sign}(f_T)) - \mathcal{R}(f_c).$$

This can be often done [15], [19], [20] by bounding the excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_\rho^V)$ and using the so-called comparison theorems. For example, for the hinge loss $V(y, f(x)) = (1 - yf(x))_+$, it was shown in [21] that $f_\rho^V = f_c$ and the comparison theorem in [15] asserts that

$$\mathcal{R}(\operatorname{sign}(f)) - \mathcal{R}(f_c) \leq \mathcal{E}(f) - \mathcal{E}(f_c)$$

for any measurable function f. For the least squares loss, the logistic loss, and the p-norm hinge loss with $p > 1$, the comparison theorem [19], [20] states that there exists a constant c_V such that for any measurable function f

$$\mathcal{R}(\operatorname{sign}(f)) - \mathcal{R}(f_c) \leq c_V \sqrt{\mathcal{E}(f) - \mathcal{E}(f_\rho^V)}.$$

Furthermore, if the distribution ρ satisfies a Tsybakov noise condition, then there is a refined comparison relation for a so-called admissible loss function, see more details in [19] and [20].

III. Related Work and Discussion

There is a large amount of work on online learning algorithms and, more generally, stochastic approximations (see [3]–[9], [12], [14]–[16], [18], [22], [23], and the references therein). In this section, we discuss some of the previous results related to this paper.

The regret bounds for online algorithms have been well studied in the literature [22]–[24]. Most of these results assume that the hypothesis space is of finite dimension, or the gradient is bounded, or the objective functions are strongly convex. Using an "online-to-batch" approach, generalization error bounds can be derived from the regret bounds.

For the nonparametric regression or classification setting, online algorithms have been studied in [3]–[6], [8], [9], [14], and [18]. Recently, Ying and Zhou [14] showed that for a loss function V satisfying

$$|V'_-(y, f) - V'_-(y, g)| \leq L|f - g|^\alpha, \quad \forall y \in Y, f, g \in \mathbb{R} \tag{13}$$

for some $0 < \alpha \leq 1$ and $0 < L < \infty$, under the assumption of existence of $\arg\inf_{f \in \mathcal{H}_K} \mathcal{E}(f) = f_{\mathcal{H}_K} \in \mathcal{H}_K$, by selecting $\eta_t = \eta_1 t^{-2/(\alpha+2)}$, there holds

$$\mathbb{E}_{z_1, z_2, \ldots, z_{T-1}}[\mathcal{E}(f_T) - \mathcal{E}(f_{\mathcal{H}_K})] = O(T^{-\frac{\alpha}{\alpha+2}}).$$

It is easy to see that such a loss function always satisfies the growth condition (5) with $q = \alpha$, when $\sup_{y \in Y} |V'_-(y, 0)| < \infty$. Therefore, as shown in Corollary 2, our learning rates for such a loss function are of order $O(T^{-(\beta/2)+\epsilon})$, which reduces to $O(T^{-(1/2)+\epsilon})$, if we further assume the existence of $f_{\mathcal{H}_K} = \arg\inf_{f \in \mathcal{H}_K} \mathcal{E}(f) \in \mathcal{H}_K$, as in [14]. Note that in general, $f_{\mathcal{H}_K}$ may not exist, thus our results require weaker assumptions, involving approximation errors in the error bounds. Also, our obtained upper bounds are better and are especially of great improvements when α is close to 0. In the cases of $\beta = 1$, these bounds are nearly optimal and up to a logarithmic factor, coincide with the minimax rates of order $O(T^{-(1/2)})$ in [10] for stochastic approximations in the nonstrongly convex case. Besides, in comparison with [14], where only loss functions satisfying (13) with $\alpha \in (0, 1]$ are considered, a broader class of convex loss functions are considered in this paper. At last, let us mention that for the least squares loss, the obtained learning rate $O(T^{-\beta/(\beta+1)} \log T)$ from Example 1 is the same as that derived in [18].

Our learning rates are also better than those for online classification in [5] and [8]. For example, for the hinge loss, the upper bound obtained in [5] is of the form $O(T^{\epsilon-\beta/(2(\beta+1))})$, while the bound in Example 1 is of the form $O(T^{-\beta/(1+\beta)} \log T)$, which is better. For a p-norm hinge loss with $p > 1$, the bound obtained in [5] is of order $O(T^{\epsilon-\beta/(2[(2-\beta)p+3\beta])})$, while the bounds in Examples 1 and 2 are of order $O(T^{\epsilon-(\beta/\max(p,2))})$.

We now compare our learning rates with those for batch learning algorithms. For general convex loss functions, the method for which sharp bounds are available is Tikhonov regularization (1). If no noise condition is imposed, the best capacity-independent error bounds for (1) with Lipschitz loss functions [2, Ch. 7], are of order $O(T^{-\beta/(\beta+1)})$. The obtained bounds in Example 1 for Lipschitz loss functions are the same as the best one available for the Tikhonov regularization, up to a logarithmic factor.

We conclude this section with some possible future work. First, it would be interesting to prove sharper rates by considering the capacity assumptions on the hypothesis spaces. Second, in this paper, we only consider the i.i.d. (independent identically distributed) setting. However, our analysis can be extended to some non-i.i.d. settings, such as the setting with Markov sampling as in [25] and [26]. Finally, our analysis may also be applied to other stochastic learning models, such as online learning with random features [27], which will be studied in our future work.

IV. Proof of Main Results

In this section, we prove our main results, Theorems 1 and 3.

A. Preliminary Lemmas

To prove Theorems 1 and 3, we need several lemmas to be proved in the Appendix.

Lemma 1 is key and will be used several times for the proof of Theorem 1. It is inspired by the recent work in [14], [28], and [29].

Lemma 1: Under Assumption 1.a, for any $f \in \mathcal{H}_K$, and $t = 1, \ldots, T-1$

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t[V(y_t, f(x_t)) - V(y_t, f_t(x_t))] \quad (14)$$

where

$$G_t = \kappa c_q \big(1 + \kappa^q \|f_t\|_K^q\big). \quad (15)$$

Using Lemma 1 and an inductive argument, we can estimate the expected value $\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1}\|_K^2]$ and provide a novel bound as follows. For notational simplicity, we denote by $\mathcal{A}(f_*)$ the excess generalization error of $f_* \in \mathcal{H}_K$ with respect to (ρ, V) as

$$\mathcal{A}(f_*) = \mathcal{E}(f_*) - \mathcal{E}\big(f_\rho^V\big). \quad (16)$$

Lemma 2: Under Assumption 1.a, let $\eta_t = \eta_1 t^{-\theta}$ with $\max((1/2), q/(q+1)) < \theta < 1$ and η_1 satisfying (9). Then, for an arbitrarily fixed $f_* \in \mathcal{H}_K$ and $t = 1, \ldots, T-1$

$$\mathbb{E}_{z_1,\ldots,z_t}\big[\|f_{t+1}\|_K^2\big] \leq 6\|f_*\|_K^2 + 4\mathcal{A}(f_*)t^{1-\theta} + 4 \quad (17)$$

and

$$\eta_{t+1}^2 \mathbb{E}_{z_1,\ldots,z_t}\big[G_{t+1}^2\big] \leq \big(3\|f_*\|_K^2 + 2\mathcal{A}(f_*)t^{1-\theta} + 3\big)(t+1)^{-q^*} \quad (18)$$

where q^* is defined in Theorem 1.

Lemma 2 asserts that for a suitable choice of decaying step sizes, $\mathbb{E}_{z_1,\ldots,z_t}[\|f_{t+1}\|_K^2]$ can be well bounded if there exists some $f_* \in \mathcal{H}_K$ such that $\mathcal{A}(f_*)$ is small. It improves uniform bounds found in the existing literature.

Replacing Assumption 1.a with Assumption 1.b in Lemma 1, we can prove the following result.

Lemma 3: Under Assumption 1.b, we have for any arbitrary $f \in \mathcal{H}_K$, and $t = 1, \ldots, T-1$

$$\|f_{t+1} - f\|_K^2 \leq \|f_t - f\|_K^2 + \eta_t^2 \kappa^2 b_V + a_V \eta_t^2 \kappa^2 V(y_t, f_t(x_t)) + 2\eta_t[V(y_t, f(x_t)) - V(y_t, f_t(x_t))]. \quad (19)$$

Using Lemma 3, and an induction argument, we can bound the expected risks of the learning sequence as follows.

Lemma 4: Under Assumption 1.b, let $\eta_t = \eta_1 t^{-\theta}$ with $\theta \in (0,1)$ and η_1 such that (11). Then, for any $t = 1, \ldots, T-1$, there holds

$$\mathbb{E}_{z_1,\ldots,z_{t-1}}\mathcal{E}(f_t) \leq \tilde{B} \quad (20)$$

where $\tilde{B}$ is a positive constant depending only on $\eta_1, \theta, b_V, \kappa^2$, and $|V|_0$ (given explicitly in the proof).

We also need the following elementary inequalities, which, for completeness, will be proved in the Appendix using a similar approach as that in [28].

Lemma 5: For any $q^* \geq 0$, there holds

$$\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \leq 2T^{-\min(1,q^*)} \log(eT).$$

Furthermore, if $q^* > 1$, then

$$\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \leq 2\left(2^{q^*} + \frac{q^*}{q^*-1}\right) T^{-1}.$$

B. Deriving Convergence From Averages

An essential tool in our error analysis is to derive the convergence of a sequence $\{u_t\}_t$ from its averages of the form $(1/T)\sum_{j=1}^T u_j$ and $(1/k)\sum_{j=T-k+1}^T u_j$. Lemma 6 is elementary for sequences and the idea is from [7]. We provide a proof in the Appendix.

Lemma 6: Let $\{u_t\}_t$ be a real-valued sequence. We have

$$u_T = \frac{1}{T}\sum_{j=1}^T u_j + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{j=T-k+1}^{T} (u_j - u_{T-k}). \quad (21)$$

From Lemma 6, we see that if the average $(1/T)\sum_{j=1}^T u_j$ tends to some u^* and the moving average $\sum_{k=1}^{T-1} 1/(k(k+1)) \sum_{j=T-k+1}^T (u_j - u_{T-k})$ tends to zero, then u_T tends to u^* as well.

Recall that our goal is to derive upper bounds for the expected excess generalization error $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$. We can easily bound the weighted average $(1/T)\sum_{t=1} 2\eta_t \mathbb{E}_{z_1,\ldots,z_{T-1}}[\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)]$ from (14) [or (19)]. This, together with Lemma 6, demonstrates how to bound the weighted excess generalization error $2\eta_T \mathbb{E}_{z_1,\ldots,z_{T-1}}[\mathcal{E}(f_T) - \mathcal{E}(f_\rho^V)]$ in terms of the weighted average and the moving weighted average. Interestingly, the bounds on the weighted average and the moving weighted average are essentially the same, as shown in Sections IV-D and IV-E.

C. Error Decomposition

Our proofs rely on a novel error decomposition derived from Lemma 6. In what follows, we shall use the notation $\mathbb{E}$ for $\mathbb{E}_{z_1,\ldots,z_{T-1}}$. Choosing $u_t = 2\eta_t \mathbb{E}\{\mathcal{E}(f_t) - \mathcal{E}(f_\rho^V)\}$ in Lemma 6, we get

$$\begin{aligned}
2\eta_T \mathbb{E}\big\{\mathcal{E}(f_T) - \mathcal{E}\big(f_\rho^V\big)\big\}
&= \frac{1}{T}\sum_{j=1}^T 2\eta_j \mathbb{E}\big\{\mathcal{E}(f_j) - \mathcal{E}\big(f_\rho^V\big)\big\} \\
&\quad + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{j=T-k+1}^{T} \big(2\eta_j \mathbb{E}\big\{\mathcal{E}(f_j) - \mathcal{E}\big(f_\rho^V\big)\big\} \\
&\quad - 2\eta_{T-k} \mathbb{E}\big\{\mathcal{E}(f_{T-k}) - \mathcal{E}\big(f_\rho^V\big)\big\}\big)
\end{aligned}$$

which can be rewritten as

$$
\begin{aligned}
&2\eta_T\mathbb{E}\{\mathcal{E}(f_T)-\mathcal{E}(f_\rho^V)\}\\
&\quad=\frac{1}{T}\sum_{t=1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\}\\
&\qquad+\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}\\
&\qquad+\sum_{k=1}^{T-1}\frac{1}{k+1}\left[\frac{2}{k}\sum_{t=T-k+1}^{T}\eta_t-\eta_{T-k}\right]\\
&\qquad\times\mathbb{E}\{\mathcal{E}(f_{T-k})-\mathcal{E}(f_\rho^V)\}. \qquad (22)
\end{aligned}
$$

Since, $\mathcal{E}(f_{T-k})-\mathcal{E}(f_\rho^V)\geq 0$ and that $\{\eta_t\}_{t\in\mathbb{N}}$ is a nonincreasing sequence, we know that the last term of (22) is at most zero. Therefore, we get

$$
\begin{aligned}
&2\eta_T\mathbb{E}\{\mathcal{E}(f_T)-\mathcal{E}(f_\rho^V)\}\\
&\quad\leq\frac{1}{T}\sum_{t=1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\}\\
&\qquad+\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}. \qquad (23)
\end{aligned}
$$

D. Proof of Theorem 1

In this section, we prove Theorem 1. We first prove the following general result, from which we can derive Theorem 1.

Theorem 5: Under Assumption 1.a, let $\eta_t=\eta_1 t^{-\theta}$ with $\max((1/2),q/(q+1))<\theta<1$ and η_1 satisfying (9). Then, for any fixed $f_*\in\mathcal{H}_K$

$$
\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_{T-1}}\{\mathcal{E}(f_T)-\mathcal{E}(f_\rho^V)\}\\
&\quad\leq\bar{C}_1\mathcal{A}(f_*)+\bar{C}_2\|f_*\|_K^2T^{-1+\theta}+\bar{C}_3T^{-1+\theta} \qquad (24)
\end{aligned}
$$

where $\bar{C}_1,\bar{C}_2$, and $\bar{C}_3$ are positive constants depending on η_1,q,κ, and θ (independent of T or f_* and given explicitly in the proof).

Proof: Let us first bound the average error, the first term of (23). Choosing $f=f_*$ in (14), taking expectation on both sides, and noting that f_t depends only on $z_1,z_2,\ldots,z_{t-1}$, we have

$$
\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_t}\big[\|f_{t+1}-f_*\|_K^2\big]\\
&\quad\leq\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[\|f_t-f_*\|_K^2\big]+\eta_t^2\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[G_t^2\big]\\
&\qquad+2\eta_t\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[\mathcal{E}(f_*)-\mathcal{E}(f_t)\big]\\
&\quad=\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[\|f_t-f_*\|_K^2\big]+\eta_t^2\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[G_t^2\big]\\
&\qquad+2\eta_t\mathcal{A}(f_*)-2\eta_t\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\big] \qquad (25)
\end{aligned}
$$

which implies

$$
\begin{aligned}
&2\eta_t\mathbb{E}\big[\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\big]\\
&\quad\leq\mathbb{E}\big[\|f_t-f_*\|_K^2\big]-\mathbb{E}\big[\|f_{t+1}-f_*\|_K^2\big]\\
&\qquad+2\eta_t\mathcal{A}(f_*)+\eta_t^2\mathbb{E}\big[G_t^2\big].
\end{aligned}
$$

Summing over $t=1,\ldots,T$, with $f_1=0$ and $\eta_t=\eta_1 t^{-\theta}$

$$
\begin{aligned}
&\sum_{t=1}^{T}2\eta_t\mathbb{E}\big[\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\big]\\
&\quad\leq\|f_*\|_K^2+2\eta_1\mathcal{A}(f_*)\sum_{t=1}^{T}t^{-\theta}+\sum_{t=1}^{T}\eta_t^2\mathbb{E}\big[G_t^2\big].
\end{aligned}
$$

This together with (18) yields

$$
\begin{aligned}
&\sum_{t=1}^{T}2\eta_t\mathbb{E}\big[\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\big]\\
&\quad\leq\|f_*\|_K^2+2\eta_1\mathcal{A}(f_*)\sum_{t=1}^{T}t^{-\theta}\\
&\qquad+\big(3\|f_*\|_K^2+2\mathcal{A}(f_*)T^{1-\theta}+3\big)\sum_{t=1}^{T}t^{-q^*}.
\end{aligned}
$$

Applying the elementary inequalities

$$
\sum_{j=1}^{t}j^{-\theta'}\leq 1+\int_1^t u^{-\theta'}du\leq\begin{cases}\dfrac{t^{1-\theta'}}{1-\theta'}, & \text{when } \theta'<1\\ \log(et), & \text{when } \theta'=1\\ \dfrac{\theta'}{\theta'-1}, & \text{when } \theta'>1\end{cases} \qquad (26)
$$

with $\theta'=\theta$ and $q^*>1$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}2\eta_t\mathbb{E}\big[\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\big]\\
&\quad\leq\left(\frac{2\eta_1}{1-\theta}+\frac{2q^*}{q^*-1}\right)\mathcal{A}(f_*)T^{1-\theta}+\big(4\|f_*\|_K^2+3\big)\frac{q^*}{q^*-1}.
\end{aligned}
$$

Dividing both sides by T, we get a bound for the first term of (23) as

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}2\eta_t\mathbb{E}\big[\mathcal{E}(f_t)-\mathcal{E}(f_\rho^V)\big]\\
&\quad\leq\left(\frac{2\eta_1}{1-\theta}+\frac{2q^*}{q^*-1}\right)\mathcal{A}(f_*)T^{-\theta}\\
&\qquad+\big(4\|f_*\|_K^2+3\big)\frac{q^*}{q^*-1}T^{-1}. \qquad (27)
\end{aligned}
$$

Then, we turn to the moving average error, the second term of (23). Let $k\in\{1,\ldots,T-1\}$. Note that f_{T-k} depends only on $z_1,\ldots,z_{T-k-1}$. Taking expectation on both sides of (14), and rearranging terms, we have that for $t\geq T-k$

$$
\begin{aligned}
&2\eta_t\mathbb{E}[\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})]\\
&\quad\leq\mathbb{E}\big[\|f_t-f_{T-k}\|_K^2\big]-\mathbb{E}\big[\|f_{t+1}-f_{T-k}\|_K^2\big]+\eta_t^2\mathbb{E}\big[G_t^2\big].
\end{aligned}
$$

Using this inequality repeatedly for $t=T-k,\ldots,T$, we have

$$
\begin{aligned}
&\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}\\
&\quad\leq\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}\eta_t^2\mathbb{E}\big[G_t^2\big].
\end{aligned}
$$

Combining this with (18) implies

$$\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}$$
$$\leq \left(3\|f_*\|_K^2+2\mathcal{A}(f_*)T^{1-\theta}+3\right)\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k}^{T}t^{-q^*}.$$

Applying Lemma 5, we have

$$\sum_{k=1}^{T-1}\frac{1}{k(k+1)}\sum_{t=T-k+1}^{T}2\eta_t\mathbb{E}\{\mathcal{E}(f_t)-\mathcal{E}(f_{T-k})\}$$
$$\leq 2\left(2^{q^*}+\frac{q^*}{q^*-1}\right)\left(3\|f_*\|_K^2+2\mathcal{A}(f_*)T^{1-\theta}+3\right)T^{-1}. \tag{28}$$

Finally, putting (27) and (28) into the error decomposition (23), and then dividing both sides by $2\eta_T=2\eta_1T^{-\theta}$, by a direct calculation, we get our desired bound (24) with

$$\bar{C}_1=\frac{1}{1-\theta}+\frac{3q^*}{\eta_1(q^*-1)}+\frac{2^{q^*+1}}{\eta_1}$$
$$\bar{C}_2=\frac{5q^*}{\eta_1(q^*-1)}+\frac{3\cdot 2^{q^*}}{\eta_1}$$

and

$$\bar{C}_3=\frac{9q^*}{2\eta_1(q^*-1)}+\frac{3\cdot 2^{q^*}}{\eta_1}.$$

The proof is complete. □

We are in a position to prove Theorem 1.

Proof of Theorem 1: By Theorem 5, we have

$$\mathbb{E}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\}$$
$$\leq(\bar{C}_1+\bar{C}_2)\left\{\mathcal{E}(f_*)-\mathcal{E}\left(f_\rho^V\right)+\|f_*\|_K^2T^{\theta-1}\right\}+\bar{C}_3T^{\theta-1}.$$

Since the constants $\bar{C}_1,\bar{C}_2$, and $\bar{C}_3$ are independent of $f_*\in\mathcal{H}_K$, we take the infimum over $f_*\in\mathcal{H}_K$ on both sides, and conclude that

$$\mathbb{E}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\}\leq(\bar{C}_1+\bar{C}_2)\mathcal{D}(T^{\theta-1})+\bar{C}_3T^{\theta-1}.$$

The proof of Theorem 1 is complete by taking $\widetilde{C}=\bar{C}_1+\bar{C}_2+\bar{C}_3$.

E. Proof of Theorem 3

In this section, we give the proof of Theorem 3. It follows from the following more general theorem, as shown in the proof of Theorem 1.

Theorem 6: Under Assumption 1.b, let $\eta_t=\eta_1t^{-\theta}$ with $0<\theta<1$ and η_1 satisfying (11). Then, for any fixed $f_*\in\mathcal{H}_K$

$$\mathbb{E}_{z_1,\ldots,z_{T-1}}\left\{\mathcal{E}(f_T)-\mathcal{E}\left(f_\rho^V\right)\right\}$$
$$\leq\left(2\mathcal{A}(f_*)+(2\eta_1)^{-1}\|f_*\|_K^2T^{-1+\theta}+\bar{B}_1T^{-\min(\theta,1-\theta)}\right)\log T \tag{29}$$

where $\bar{B}_1$ is a positive constant depending only on η_1,a_V,b_V,κ, and θ (independent of T or f_* and given explicitly in the proof).

Proof: The proof parallels to that of Theorem 5. Note that we have the error decomposition (23). We only need to estimate the last two terms of (23).

To bound the first term of the right-hand side of (23), we first apply Lemma 3 with a fixed $f\in\mathcal{H}_K$ and subsequently take the expectation on both sides of (19) to get

$$\mathbb{E}\left[\|f_{l+1}-f\|_K^2\right]$$
$$\leq\mathbb{E}\left[\|f_l-f\|_K^2\right]$$
$$+\eta_l^2\kappa^2(a_V\mathbb{E}[\mathcal{E}(f_l)]+b_V)+2\eta_l\mathbb{E}(\mathcal{E}(f)-\mathcal{E}(f_l)). \tag{30}$$

By Lemma 4, we have (20). Introducing (20) into (30) with $f=f_*$, and rearranging terms

$$2\eta_l\mathbb{E}\left(\mathcal{E}(f_l)-\mathcal{E}\left(f_\rho^V\right)\right)\leq\mathbb{E}\left[\|f_l-f_*\|_K^2-\|f_{l+1}-f_*\|_K^2\right]$$
$$+2\eta_l\mathcal{A}(f_*)+\eta_l^2\kappa^2(a_V\tilde{B}+b_V).$$

Summing up over $l=1,\ldots,T$, rearranging terms, and then dividing both sides by T, we get

$$\frac{1}{T}\sum_{l=1}^{T}2\eta_l\mathbb{E}(\mathcal{E}(f_l)-\mathcal{E}(f_*))$$
$$\leq\frac{\|f_*\|_K^2}{T}+\frac{2\eta_1}{T}\mathcal{A}(f_*)\sum_{t=1}^{T}t^{-\theta}+\eta_1^2\kappa^2(a_V\tilde{B}+b_V)\frac{1}{T}\sum_{l=1}^{T}l^{-2\theta}.$$

By using the elementary inequality with $q\geq 0, T\geq 3$

$$\sum_{t=1}^{T}t^{-q}\leq T^{\max(1-q,0)}\sum_{t=1}^{T}t^{-1}\leq 2T^{\max(1-q,0)}\log T$$

one can get

$$\frac{1}{T}\sum_{l=1}^{T}2\eta_l\mathbb{E}(\mathcal{E}(f_l)-\mathcal{E}(f_*))$$
$$\leq\frac{\|f_*\|_K^2}{T}+4\eta_1\mathcal{A}(f_*)T^{-\theta}\log T$$
$$+\eta_1^2 2\kappa^2(a_V\tilde{B}+b_V)T^{-\min(2\theta,1)}\log T. \tag{31}$$

To bound the last term of (23), we let $1\leq k\leq t-1$ and $i\in\{t-k,\ldots,t\}$. Note that f_i depends only on $z_1,\ldots,z_{i-1}$ when $i>1$. We apply Lemma 3 with $f=f_{t-k}$, and then take the expectation on both sides of (19) to derive

$$2\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]$$
$$\leq\mathbb{E}\left[\|f_i-f_{t-k}\|_K^2-\|f_{i+1}-f_{t-k}\|_K^2\right]$$
$$+\eta_i^2\kappa^2(a_V\mathbb{E}[\mathcal{E}(f_i)]+b_V).$$

Summing up over $i=t-k,\ldots,t$

$$\sum_{i=t-k}^{t}2\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]\leq\kappa^2\sum_{i=t-k}^{t}\eta_i^2(a_V\mathbb{E}[\mathcal{E}(f_i)]+b_V).$$

Note that the left-hand side is exactly $\sum_{i=t-k+1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]$. We thus know that

$$\begin{aligned}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}&\sum_{i=t-k+1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]\\ &\le \frac{\kappa^2}{2}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t}\eta_i^2(a_V\mathbb{E}[\mathcal{E}(f_i)]+b_V)\\ &\le \frac{\kappa^2}{2}\big(a_V\sup_{1\le i\le t}\mathbb{E}[\mathcal{E}(f_i)]+b_V\big)\sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k}^{t}\eta_i^2.\end{aligned}$$

With $\eta_t=\eta_1 t^{-\theta}$, by using Lemma 5, this can be relaxed as

$$\begin{aligned}\sum_{k=1}^{t-1}\frac{1}{k(k+1)}&\sum_{i=t-k+1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)-\mathcal{E}(f_{t-k})]\\ &\le \eta_1^2\kappa^2 t^{-\min(2\theta,1)}\log(et)(a_V\sup_{1\le i\le t}\mathbb{E}[\mathcal{E}(f_i)]+b_V).\quad(32)\end{aligned}$$

Introducing (31) and (32) into (23), plugging with (20), and dividing both sides by $2\eta_T=2\eta_1 T^{-\theta}$, one can prove the desired result with $\bar{B}_1=2\eta_1\kappa^2(a_V\tilde{B}+b_V)$. □

V. Numerical Simulations

The simplest case to implement online learning algorithm (3) is when $X=\mathbb{R}^d$ for some $d\in\mathbb{N}$ and K is the linear kernel given by $K(x,w)=w^Tx$. In this case, it is straightforward to see that $f_{t+1}(x)=w_{t+1}^\top x$ with $w_1=0\in\mathbb{R}^d$ and

$$w_{t+1}=w_t-\eta_t V'_-(y_t,w_t^\top x_t)x_t,\quad t=1,\ldots,T.$$

For a general kernel, by induction, it is easy to see that $f_{t+1}(x)=\sum_{j=1}^{T}c_{t+1}^jK(x,x_j)$ with

$$c_{t+1}=c_t-\eta_t V'_-\left(y_t,\sum_{j=1}^{T}c_t^jK(x_t,x_j)\right)\mathbf{e}_t,\quad t=1,\ldots,T$$

for $c_1=0\in\mathbb{R}^T$. Here, $c_t=(c_t^1,\ldots,c_t^T)^\top$ for $1\le t\le T$, and $\{\mathbf{e}_1,\ldots,\mathbf{e}_T\}$ is a standard basis of $\mathbb{R}^T$. Indeed, it is straightforward to check by induction that

$$\begin{aligned}f_{t+1}&=\sum_{j=1}^{T}c_t^jK_{x_j}-\eta_tV'_-(y_t,f_t(x_t))K_{x_t}\\ &=\sum_{j=1}^{T}K_{x_j}\big(c_t^j-\eta_tV'_-(y_j,f_t(x_j))\mathbf{e}_t^j\big).\end{aligned}$$

To see how the step-size decaying rate indexed by θ affects the performance of the studied algorithm, we carry out simple numerical simulations on the *Adult*[1] data set with the hinge loss and the Gaussian kernel with kernel width $\sigma=4$. We consider a subset of *Adult* with $T=1000$, and run the algorithm for different θ values with $\eta_1=1/4$. The test and training errors (with respect to the hinge loss) for different θ values are shown in Fig. 1. We see that the minimal test error (with respect to the hinge loss) is achieved at some $\theta^*<1/2$,

[1]The data set can be downloaded from `archive.ics.uci.edu/ml` and `www.csie.ntu.edu.tw/cjlin/libsvmtools/`

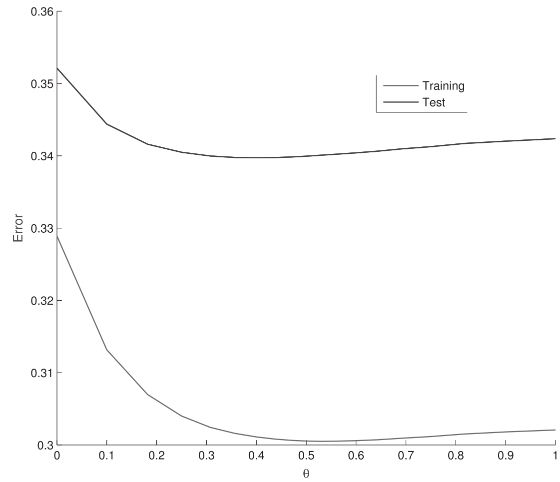

Fig. 1. Test and training errors for online learning with different θ values on *Adult* ($T=1000$).

TABLE I
Comparison of Online Learning Using Cross Validation With LIBSVM

Algorithm	test classification error	training time
online learning	$16.2\pm0.2\%$	5.4 ± 0.3
LIBSVM	$18.7\pm0.0\%$	5.8 ± 0.5

which complements our obtained results. We also compare the performance of online learning algorithm (3) in terms of test error and training time with that of LIBSVM, a state-of-the-art batch learning algorithm for classification [30]. The test classification error and training time, for the online learning algorithm using cross validation (for choosing the best θ) and LIBSVM, are summarized in Table I, from which we see that the online learning algorithm is comparable to LIBSVM on both test error and running time.

Appendix

In this appendix, we prove the lemmas stated before.

Proof of Lemma 1: Since f_{t+1} is given by (3), by expanding the inner product, we have

$$\begin{aligned}\|f_{t+1}-f\|_K^2=&\|f_t-f\|_K^2+\eta_t^2\|V'_-(y_t,f_t(x_t))K_{x_t}\|_K^2\\ &+2\eta_tV'_-(y_t,f_t(x_t))\langle K_{x_t},f-f_t\rangle_K.\end{aligned}$$

Observe that $\|K_{x_t}\|_K=(K(x_t,x_t))^{1/2}\le\kappa$ and that

$$\|f\|_\infty\le\kappa\|f\|_K,\quad\forall f\in\mathcal{H}_K.$$

These together with the incremental condition (5) yield

$$\begin{aligned}&\|V'_-(y_t,f_t(x_t))K_{x_t}\|_K\\ &\quad\le\kappa|V'_-(y_t,f_t(x_t))|\\ &\quad\le\kappa c_q(1+|f_t(x_t)|^q)\le\kappa c_q\big(1+\kappa^q\|f_t\|_K^q\big).\end{aligned}$$

Therefore, $\|f_{t+1}-f\|_K^2$ is bounded by

$$\|f_t-f\|_K^2+\eta_t^2G_t^2+2\eta_tV'_-(y_t,f_t(x_t))\langle K_{x_t},f-f_t\rangle_K.$$

Using the reproducing property, we get

$$\begin{aligned}\|f_{t+1}-f\|_K^2\le&\|f_t-f\|_K^2+\eta_t^2G_t^2\\ &+2\eta_tV'_-(y_t,f_t(x_t))(f(x_t)-f_t(x_t)).\quad(33)\end{aligned}$$

Since $V(y_t, \cdot)$ is a convex function, we have

$$V'_-(y_t, a)(b-a) \le V(y_t, b) - V(y_t, a), \quad \forall a, b \in \mathbb{R}.$$

Using this relation to (33), we get our desired result.

In order to prove Lemma 2, we first bound the learning sequence uniformly as follows.

Lemma 7: Under Assumption 1.a, let $0 \le \theta < 1$ satisfy $\theta \ge \frac{q}{q+1}$ and $\eta_t = \eta_1 t^{-\theta}$ with η_1 satisfying

$$0 < \eta_1 \le \min\left\{\frac{\sqrt{1-\theta}}{\sqrt{8}c_q(\kappa+1)^{q+1}}, \frac{1-\theta}{4|V|_0}\right\}. \tag{34}$$

Then, for $t = 1, \dots, T-1$

$$\|f_{t+1}\|_K \le t^{\frac{1-\theta}{2}}. \tag{35}$$

Proof: We prove our statement by induction.

Taking $f = 0$ in Lemma 1, we know that

$$\begin{aligned}\|f_{t+1}\|_K^2 &\le \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t[V(y_t, 0) - V(y_t, f_t(x_t))] \\ &\le \|f_t\|_K^2 + \eta_t^2 G_t^2 + 2\eta_t |V|_0. \end{aligned} \tag{36}$$

Since $f_1 = 0$, G_1 is given by (15) and by (34), $\eta_1^2 c_q^2 \kappa^2 + 2\eta_1 |V|_0 \le 1$, we thus get (35) for the case $t = 1$.

Now, assume $\|f_t\|_K \le (t-1)^{(1-\theta)/2}$ with $t \ge 2$. Then

$$\begin{aligned}G_t^2 &\le \kappa^2 c_q^2 (1+\kappa^q)^2 \max\left(1, \|f_t\|_K^{2q}\right) \\ &\le 4c_q^2(\kappa+1)^{2q+2}(t-1)^{(1-\theta)q} \end{aligned} \tag{37}$$

where for the last inequality, we used $\kappa \le \kappa + 1$ and $1 + \kappa^q \le 2(\kappa+1)^q$. Hence, using (36)

$$\begin{aligned}&\|f_{t+1}\|_K^2 \\ &\le (t-1)^{1-\theta} + \eta_1^2 t^{-2\theta} 4c_q^2(\kappa+1)^{2q+2} t^{(1-\theta)q} + 2\eta_1 t^{-\theta}|V|_0 \\ &= t^{1-\theta}\left\{\left(1 - \frac{1}{t}\right)^{1-\theta} + \frac{\eta_1^2 4c_q^2(\kappa+1)^{2q+2}}{t^{(q+1)\theta+1-q}} + \frac{2\eta_1|V|_0}{t}\right\}.\end{aligned}$$

Since $(1-(1/t))^{1-\theta} \le 1 - (1-\theta)/t$ and the condition $\theta \ge q/(q+1)$ implies $(q+1)\theta + 1 - q \ge 1$, we see that $\|f_{t+1}\|_K^2$ is bounded by

$$t^{1-\theta}\left\{1 - \frac{1-\theta}{t} + \frac{\eta_1^2 4c_q^2(\kappa+1)^{2q+2}}{t} + \frac{2\eta_1|V|_0}{t}\right\}.$$

Finally, we use the restriction (34) for η_1 and find $\|f_{t+1}\|_K^2 \le t^{1-\theta}$. This completes the induction procedure and proves our conclusion. □

Now, we are ready to prove Lemma 2.

Proof of Lemma 2: Recall an iterative relation (25) of error terms in the proof of Theorem 5. It follows from $\mathcal{E}(f_t) \ge \mathcal{E}(f_\rho^V)$ that

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \le \mathbb{E}_{z_1,\dots,z_{t-1}}&\left[\|f_t - f_*\|_K^2\right] \\ &+ \eta_t^2 \mathbb{E}_{z_1,\dots,z_{t-1}}\left[G_t^2\right] + 2\eta_t \mathcal{A}(f_*).\end{aligned} \tag{38}$$

Since G_t is given by (15), applying Schwarz's inequality

$$\mathbb{E}_{z_1,\dots,z_{t-1}}\left[G_t^2\right] \le 2\kappa^2 c_q^2\left(1 + \kappa^{2q}\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^{2q}\right]\right).$$

If $q \le 1$, using Hölder's inequality

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^{2q}\right] &\le \left(\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^2\right]\right)^q \\ &\le 1 + \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^2\right].\end{aligned}$$

If $q > 1$, noting that (9) implies (34), we have (35) and thus

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^{2q}\right] &\le \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^2\right] t^{(q-1)(1-\theta)} \\ &= \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^2\right] t^{2\theta - q^*}.\end{aligned}$$

Combining the above-mentioned two cases yields

$$\begin{aligned}&\eta_t^2 \mathbb{E}_{z_1,\dots,z_{t-1}}\left[G_t^2\right] \\ &\quad\le 2\kappa^2 c_q^2 \eta_t^2\left(1 + \kappa^{2q}\left(1 + \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t\|_K^2\right]\right) t^{2\theta - q^*}\right) \\ &\quad\le 2\kappa^2 c_q^2 \eta_t^2\left(1 + \kappa^{2q} t^{2\theta - q^*}\right. \\ &\qquad\qquad \left.\cdot \left(1 + 2\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t - f^*\|_K^2\right] + 2\|f_*\|_K^2\right)\right) \\ &\quad\le C_1\left(1 + \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t - f^*\|_K^2\right] + \|f_*\|_K^2\right) t^{-q^*} \end{aligned} \tag{39}$$

where

$$C_1 = 4\eta_1^2 c_q^2 (1+\kappa)^{2q+2}. \tag{40}$$

Putting (39) into (38) yields

$$\begin{aligned}&\mathbb{E}_{z_1,\dots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\ &\quad\le \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t - f_*\|_K^2\right] + 2\eta_1 t^{-\theta}\mathcal{A}(f_*) \\ &\qquad + C_1\left(1 + \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|f_t - f^*\|_K^2\right] + \|f_*\|_K^2\right) t^{-q^*}.\end{aligned}$$

Applying this inequality iteratively, with $f_1 = 0$, we derive

$$\begin{aligned}&\mathbb{E}_{z_1,\dots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\ &\quad\le \|f_*\|_K^2 + 2\eta_1\mathcal{A}(f_*)\sum_{j=1}^t j^{-\theta} \\ &\qquad + C_1\left(1 + \|f_*\|_K^2\right. \\ &\qquad\qquad \left.+ \max_{j=1,\dots,t}\mathbb{E}_{z_1,\dots,z_{j-1}}\left[\|f_j - f^*\|_K^2\right]\right)\sum_{j=1}^t j^{-q^*}.\end{aligned}$$

Note that $\theta \in (1/2, 1)$ and that from the restriction on θ, $q^* > 1$. Applying the elementary inequality (26) to bound $\sum_{j=1}^t j^{-q^*}$ and $\sum_{j=1}^t j^{-\theta}$, we get

$$\begin{aligned}&\mathbb{E}_{z_1,\dots,z_t}\left[\|f_{t+1} - f_*\|_K^2\right] \\ &\le \|f_*\|_K^2 + \frac{2\eta_1}{1-\theta}\mathcal{A}(f_*)t^{1-\theta} \\ &\quad + \frac{C_1 q^*}{q^*-1}\left(1 + \|f_*\|_K^2 + \max_{j=1,\dots,t}\mathbb{E}_{z_1,\dots,z_{j-1}}\left[\|f_j - f^*\|_K^2\right]\right).\end{aligned}$$

Now, we derive upper bounds for $\mathbb{E}_{z_1,\dots,z_t}[\|f_{t+1} - f_*\|_K^2]$ by induction for $t = 1, \dots, T-1$. Assume that $\mathbb{E}_{z_1,\dots,z_{j-1}}[\|f_j - f_*\|_K^2] \le 2(\|f_*\|_K^2 + \mathcal{A}(f_*)(j-1)^{1-\theta} + 1)$ holds for

$j = 1, \ldots, t$. Then

$$\begin{aligned}
&\mathbb{E}_{z_1,\ldots,z_t}\big[\|f_{t+1} - f_*\|_K^2\big] \\
&\quad\le \|f_*\|_K^2 + \frac{C_1 q^*}{q^*-1}(3 + 3\|f_*\|_K^2 + 2\mathcal{A}(f_*)t^{1-\theta}]) \\
&\qquad + \frac{2\eta_1}{1-\theta}\mathcal{A}(f_*)t^{1-\theta} \\
&\quad\le \left(1 + \frac{3C_1 q^*}{q^*-1}\right)(1 + \|f_*\|_K^2) \\
&\qquad + \left(\frac{2C_1 q^*}{q^*-1} + \frac{2\eta_1}{1-\theta}\right)\mathcal{A}(f_*)t^{1-\theta}.
\end{aligned}$$

Recall that C_1 is given by (40). We see from (9) that $3C_1 q^*/(q^*-1) \le 1-\theta \le 1$ and $2\eta_1/(1-\theta) \le 1$. It follows that

$$\mathbb{E}_{z_1,\ldots,z_t}\big[\|f_{t+1} - f_*\|_K^2\big] \le 2\big(\|f_*\|_K^2 + \mathcal{A}(f_*)t^{1-\theta} + 1\big). \quad (41)$$

From the above-mentioned induction procedure, we conclude that for $t = 1, \ldots, T-1$, the bound (41) holds, which leads to the desired bound (17) using $\|f_t\|_K^2 \le 2\|f_t - f_*\|_K^2 + 2\|f_*\|_K^2$. Applying (41) into (39), and noting that $C_1 \le 1$ by the restriction (9), we get the other desired bound (18). The proof is complete.

Proof of Lemma 3: Following the proof of Lemma 1, we have:

$$\begin{aligned}
\|f_{t+1} - f\|_K^2 &\le \|f_t - f\|_K^2 + \eta_t^2\kappa^2|V_-(y_t, f_t(x_t))|^2 \\
&\quad + 2\eta_t\left[V(y_t, f(x_t)) - V(y_t, f_t(x_t))\right].
\end{aligned}$$

Applying Assumption 1.b to the above, we get the desired result.

Proof of Lemma 4: The proof is divided into several steps.

Basic Decomposition: We choose $\mu_t = \eta_t\mathbb{E}[\mathcal{E}(f_t)]$ in Lemma 6 to get

$$\begin{aligned}
&\eta_t\mathbb{E}[\mathcal{E}(f_t)] \\
&\quad = \frac{1}{t}\sum_{i=1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)] \\
&\qquad + \sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k+1}^{t}(\eta_i\mathbb{E}[\mathcal{E}(f_i)] - \eta_{t-k}\mathbb{E}[\mathcal{E}(f_{t-k})]).
\end{aligned}$$

Since $\{\eta_t\}_t$ is decreasing and $\mathbb{E}[\mathcal{E}(f_{t-k})]$ is nonnegative, the above can be relaxed as

$$\begin{aligned}
\eta_t\mathbb{E}[\mathcal{E}(f_t)] &\le \frac{1}{t}\sum_{i=1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i)] \\
&\quad + \sum_{k=1}^{t-1}\frac{1}{k(k+1)}\sum_{i=t-k+1}^{t}\eta_i\mathbb{E}[\mathcal{E}(f_i) - \mathcal{E}(f_{t-k})].
\end{aligned} \quad (42)$$

In the rest of the proof, we will bound the last two terms in the above-mentioned estimate.

Bounding the Average: To bound the first term on the right-hand side of (42), we apply (30) with $f = 0$ to get

$$\begin{aligned}
\mathbb{E}\big[\|f_{l+1}\|_K^2\big] &\le \mathbb{E}\big[\|f_l\|_K^2\big] + \eta_l^2\kappa^2(a_V\mathbb{E}[\mathcal{E}(f_l)] + b_V) \\
&\quad + 2\eta_l\mathbb{E}(\mathcal{E}(0) - \mathcal{E}(f_l)).
\end{aligned}$$

Rearranging terms, and using the fact that $\mathcal{E}(0) \le |V|_0$

$$\begin{aligned}
&\eta_l(2 - a_V\eta_l\kappa^2)\mathbb{E}[\mathcal{E}(f_l)] \\
&\quad \le \mathbb{E}[\|f_l\|_K^2 - \|f_{l+1}\|_K^2] + b_V\eta_l^2\kappa^2 + 2\eta_l|V|_0.
\end{aligned}$$

It thus follows from $a_V\eta_l\kappa^2 \le 1$, implied by (11), that

$$\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \mathbb{E}\big[\|f_l\|_K^2 - \|f_{l+1}\|_K^2\big] + b_V\eta_l^2\kappa^2 + 2\eta_l|V|_0. \quad (43)$$

Summing up over $l = 1, \ldots, t$, introducing $f_1 = 0$, $\|f_{t+1}\|_K^2 \ge 0$, and then multiplying both sides by $1/t$, we get

$$\frac{1}{t}\sum_{l=1}^{t}\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \frac{1}{t}\sum_{l=1}^{t}\big(b_V\eta_l^2\kappa^2 + 2\eta_l|V|_0\big).$$

Since $\eta_t = \eta_1 t^{-\theta}$, we have

$$\frac{1}{t}\sum_{l=1}^{t}\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \big(b_V\eta_1^2\kappa^2 + 2\eta_1|V|_0\big)\frac{1}{t}\sum_{l=1}^{t}l^{-\theta}.$$

Using (26), we get

$$\frac{1}{t}\sum_{l=1}^{t}\eta_l\mathbb{E}[\mathcal{E}(f_l)] \le \frac{b_V\eta_1^2\kappa^2 + 2\eta_1|V|_0}{1-\theta}t^{-\theta}. \quad (44)$$

Bounding the Moving Average: To bound the last term of (42), we let $1 \le k \le t-1$ and $i \in \{t-k, \ldots, t\}$. Recall the inequality (32) in the proof of Theorem 6. Applying the basic inequality $e^{-x} \le (ex)^{-1}, x > 0$, which implies $t^{-\min(\theta,1-\theta)}\log(et) \le (1/\min(\theta, 1-\theta))$, we see that the last term of (42) can be upper bounded by

$$\frac{\eta_1^2\kappa^2}{\min(\theta, 1-\theta)}t^{-\theta}\left(a_V\sup_{1\le i\le t}\mathbb{E}[\mathcal{E}(f_i)] + b_V\right).$$

Induction: Introducing (32) and (44) into the decomposition (42), and then dividing both sides by $\eta_t = \eta_1 t^{-\theta}$, we get

$$\mathbb{E}[\mathcal{E}(f_t)] \le A\sup_{1\le i\le t}\mathbb{E}[\mathcal{E}(f_i)] + B \quad (45)$$

where we set $A = (\eta_1 a_V\kappa^2/\min(\theta, 1-\theta))$ and

$$B = \frac{b_V\eta_1\kappa^2 + 2|V|_0}{1-\theta} + \frac{\eta_1 b_V\kappa^2}{\min(\theta, 1-\theta)}.$$

The restriction (11) on η_1 tells us that $A \le 1/2$. Then, using (45) with an inductive argument, we find that for all $t \le T$

$$\mathbb{E}[\mathcal{E}(f_t)] \le 2B \quad (46)$$

which leads to the desired result with $\tilde{B} = 2B$. In fact, the case $t = 2$ can be verified directly from (43), by plugging with $f_1 = 0$. Now, assume that (46) holds for any $k \le t-1$, where $t \ge 3$. Under this hypothesis condition, if $\mathbb{E}[\mathcal{E}(f_t)] \le \sup_{1\le i\le t-1}\mathbb{E}[\mathcal{E}(f_i)]$, then using the hypothesis condition, we know that $\mathbb{E}[\mathcal{E}(f_t)] \le 2B$. If $\mathbb{E}[\mathcal{E}(f_t)] \ge \sup_{1\le i\le t-1}\mathbb{E}[\mathcal{E}(f_i)]$, we use (45) to get

$$\mathbb{E}[\mathcal{E}(f_t)] \le A\mathbb{E}[\mathcal{E}(f_t)] + B \le \mathbb{E}[\mathcal{E}(f_t)]/2 + B$$

which implies $\mathbb{E}[\mathcal{E}(f_t)] \le 2B$. The proof is thus complete.

Proof of Lemma 5: Exchanging the order in the sum, we have

$$
\begin{aligned}
&\sum_{k=1}^{T-1} \frac{1}{k(k+1)} \sum_{t=T-k}^{T} t^{-q^*} \\
&\quad = \sum_{t=1}^{T-1} \sum_{k=T-t}^{T-1} \frac{1}{k(k+1)} t^{-q^*} + \sum_{k=1}^{T-1} \frac{1}{k(k+1)} T^{-q^*} \\
&\quad = \sum_{t=1}^{T-1} \left(\frac{1}{T-t} - \frac{1}{T}\right) t^{-q^*} + \left(1 - \frac{1}{T}\right) T^{-q^*} \\
&\quad \leq \sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*}.
\end{aligned}
$$

What remains is to estimate the term $\sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*}$. Note that

$$
\sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*} = \sum_{t=1}^{T-1} \frac{t^{1-q^*}}{(T-t)t} \leq T^{\max(1-q^*,0)} \sum_{t=1}^{T-1} \frac{1}{(T-t)t}
$$

and that by (26)

$$
\begin{aligned}
\sum_{t=1}^{T-1} \frac{1}{(T-t)t} &= \frac{1}{T} \sum_{t=1}^{T-1} \left(\frac{1}{T-t} + \frac{1}{t}\right) \\
&= \frac{2}{T} \sum_{t=1}^{T-1} \frac{1}{t} \leq \frac{2}{T} \log(eT).
\end{aligned}
$$

From the above-mentioned analysis, we see the first statement of the lemma.

To prove the second part of the lemma, we split the term $\sum_{t=1}^{T-1} 1/(T-t) t^{-q^*}$ into two parts

$$
\begin{aligned}
&\sum_{t=1}^{T-1} \frac{1}{T-t} t^{-q^*} \\
&\quad = \sum_{T/2 \leq t \leq T-1} \frac{1}{T-t} t^{-q^*} + \sum_{1 \leq t < T/2} \frac{1}{T-t} t^{-q^*} \\
&\quad \leq 2^{q^*} T^{-q^*} \sum_{T/2 \leq t \leq T-1} \frac{1}{T-t} + 2T^{-1} \sum_{1 \leq t < T/2} t^{-q^*} \\
&\quad = 2^{q^*} T^{-q^*} \sum_{1 \leq t \leq T/2} t^{-1} + 2T^{-1} \sum_{1 \leq t < T/2} t^{-q^*}.
\end{aligned}
$$

Applying (26) to the above and then using $T^{-q^*+1} \log T \leq 1/(2(q^* - 1))$, we see the second statement of Lemma 5.

Proof of Lemma 6: For $k = 1, \ldots, T-1$

$$
\begin{aligned}
&\frac{1}{k} \sum_{j=T-k+1}^{T} u_j - \frac{1}{k+1} \sum_{j=T-k}^{T} u_j \\
&\quad = \frac{1}{k(k+1)} \left\{ (k+1) \sum_{j=T-k+1}^{T} u_j - k \sum_{j=T-k}^{T} u_j \right\} \\
&\quad = \frac{1}{k(k+1)} \sum_{j=T-k+1}^{T} (u_j - u_{T-k}).
\end{aligned}
$$

Summing over $k = 1, \ldots, T-1$, and rearranging terms, we get (21).

Acknowledgment

The authors would like to thank the reviewers and editors for their constructive comments, and Dr. Raffaello Camoriano and Prof. Lorenzo Rosasco for sharing codes for our simulations.

References

[1] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[2] I. Steinwart and A. Christmann, *Support Vector Machines*. New York, NY, USA: Springer, 2008.

[3] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of on-line learning algorithms," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2050–2057, Sep. 2004.

[4] S. Smale and Y. Yao, "Online learning algorithms," *Found. Comput. Math.*, vol. 6, no. 2, pp. 145–170, 2006.

[5] Y. Ying and D. X. Zhou, "Online regularized classification algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 4775–4788, Nov. 2006.

[6] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 773–781.

[7] O. Shamir and T. Zhang, "Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 71–79.

[8] G. B. Ye and D. X. Zhou, "Fully online classification by regularization," *Appl. Comput. Harmon. Anal.*, vol. 23, no. 2, pp. 198–214, 2007.

[9] P. Tarres and Y. Yao, "Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5716–5735, Sep. 2014.

[10] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar, "Information-theoretic lower bounds on the oracle complexity of convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1–9.

[11] T. Hu, J. Fan, Q. Wu, and D. X. Zhou, "Regularization schemes for minimum error entropy principle," *Anal. Appl.*, vol. 13, no. 4, pp. 437–455, 2015.

[12] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 817–824.

[13] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri, "Some properties of regularized kernel methods," *J. Mach. Learn. Res.*, vol. 5, pp. 1363–1390, Oct. 2004.

[14] Y. Ying and D. X. Zhou, "Unregularized online learning algorithms with general loss functions," *Appl. Comput. Harmon. Anal.*, vol. 42, pp. 224–244, Aug. 2017.

[15] T. Zhang, "Statistical behavior and consistency of classification methods based on convex risk minimization," *Ann. Statist.*, vol. 32, no. 1, pp. 56–85, 2004.

[16] N. Srebro, K. Sridharan, and A. Tewari. (Sep. 2010). "Optimistic rates for learning with a smooth loss." [Online]. Available: https://arxiv.org/abs/1009.3896

[17] S. Smale and D. X. Zhou, "Estimating the approximation error in learning theory," *Anal. Appl.*, vol. 1, no. 1, pp. 17–41, 2003.

[18] Y. Ying and M. Pontil, "Online gradient descent learning algorithms," *Found. Comput. Math.*, vol. 8, no. 5, pp. 561–596, 2008.

[19] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[20] D. R. Chen, Q. Wu, Y. Ying, and D. X. Zhou, "Support vector machine soft margin classifiers: Error analysis," *J. Mach. Learn. Res.*, vol. 5, pp. 1143–1175, 2004.

[21] G. Wahba, *Spline Models for Observational Data*, vol. 59. Philadelphia, PA, USA: SIAM, 1990.

[22] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Inf. Comput.*, vol. 132, no. 1, pp. 1–63, 1997.

[23] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 928–936.

[24] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: A meta-algorithm and applications.," *Theory Comput.*, vol. 8, no. 1, pp. 121–164, 2012.

[25] S. Smale and D.-X. Zhou, "Online learning with Markov sampling," *Anal. Appl.*, vol. 7, no. 1, pp. 87–113, 2009.

[26] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu, "The generalization ability of online SVM classification based on Markov sampling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 628–639, Mar. 2015.

[27] Z. Hu, M. Lin, and C. Zhang, "Dependent online kernel learning with constant number of random Fourier features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2464–2476, Oct. 2015.

[28] J. Lin, L. Rosasco, and D.-X. Zhou, "Iterative regularization for learning with convex loss functions," *J. Mach. Learn. Res.*, vol. 17, no. 77, pp. 1–38, 2016.

[29] J. Lin and D.-X. Zhou, "Learning theory of randomized Kaczmarz algorithm," *J. Mach. Learn. Res.*, vol. 16, pp. 3341–3365, Jan. 2015.

[30] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.

Junhong Lin received the Ph.D. degree in applied mathematics from Zhejiang University, Hangzhou, China, in 2013.

From 2013 to 2015, he was a Post-Doctoral Fellow with the Department of Mathematics, City University of Hong Kong, Hong Kong. He is currently with the Laboratory for Computational and Statistical Learning, Massachusetts Institute of Technology, Cambridge, MA, USA, and also with the Istituto Italiano di Tecnologia, Genoa, Italy. His current research interests include compressed sensing and learning theory.

Ding-Xuan Zhou received the B.Sc. and Ph.D. degrees in mathematics from Zhejiang University, Hangzhou, China, in 1988 and 1991, respectively.

He joined the Faculty of the City University of Hong Kong, Hong Kong, in 1996, where he is currently a Chair Professor of the Department of Mathematics. He has authored over 100 research papers. His current research interests include learning theory, data science, wavelet analysis, and approximation theory.

Dr. Zhou is serving on the Editorial Board of over 10 international journals, and is an Editor-in-Chief of the *Journal Analysis and Application*. He received a Joint Research Fund from the National Science Fund of China for Distinguished Young Scholars in 2005 and the Humboldt Research Fellowship in 1993, and was rated in 2014, 2015, and 2016 by Thomson Reuters as a Highly-cited Researcher. He has co-organized over 20 international conferences and conducted over 20 research grants.