Error analysis for online gradient descent algorithms in reproducing kernel Hilbert spaces[†]

Massimiliano Pontil, Yiming Ying

Department of Computer Science, University College London
Gower Street, London, WC1E 6BT, England, UK
{m.pontil, y.ying}@cs.ucl.ac.uk

Ding-Xuan Zhou

Department of Mathematics, City University of Hong Kong
Kowloon, Hong Kong, CHINA.
mazhou@cityu.edu.hk

Abstract

We consider online gradient descent algorithms with general convex loss functions in reproducing kernel Hilbert spaces (RKHS). These algorithms offer an advantageous way for learning from large training sets. We provide general conditions ensuring convergence of the algorithm in the RKHS norm. Explicit generalization error rates for q-norm ε-insensitive regression loss are given by choosing the step sizes and the regularization parameter appropriately.

Keywords and Phrases: Online learning, regularization, reproducing kernel Hilbert space, gradient descent, general loss function, error analysis.

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

[†]Corresponding author: Yiming Ying. Tel: +44(0)20 7679 0374; Fax: +44(0)20 7387 1397.

1 Introduction

Let X be a compact subset of Euclidean space $\mathbb{R}^d$, Y a bounded subset in $\mathbb{R}$ and define $\mathbb{N}_\ell := \{1, \ldots, \ell\}$ for any $\ell \in \mathbb{N}$. We consider the supervised learning problem of learning a function or predictor $f : X \to Y$ on the base of a finite training set $\mathbf{z} = \{z_t = (x_t, y_t) : t \in \mathbb{N}_T\} \subseteq Z$ of input/output points sampled *i.i.d.* from a fixed but unknown distribution ρ. The quality of the predictor f is measured by the *generalization error*

$$\mathcal{E}(f) := \int_Z V(y, f(x)) d\rho(x, y), \tag{1.1}$$

where $V : Y \times \mathbb{R} \to \mathbb{R}_+$ is a prescribed loss function.

In this paper we restrict our attention to learning algorithms which compute a predictor in a prescribed reproducing kernel Hilbert spaces (RKHS). Let $K : X \times X \to \mathbb{R}$ be a *Mercer kernel*, that is, a continuous, symmetric and positive semi-definite kernel, see e.g. [7]. The RKHS $\mathcal{H}_K$ associated with kernel K is defined [1] to be the completion of the linear span of the set of functions $\{K_x(\cdot) = K(x, \cdot) : x \in X\}$ with inner product satisfying, for all $x \in X$ and $g \in \mathcal{H}_K$, the *reproducing property*

$$\langle K_x, g \rangle_K = g(x). \tag{1.2}$$

A common approach is to compute f by minimizing the regularization functional, that is, we consider the learning algorithm

$$f_{\mathbf{z},\lambda} := \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t \in \mathbb{N}_T} V(y_t, f(x_t)) + \frac{\lambda}{2} \|f\|_K^2 \right\} \tag{1.3}$$

where $\|f\|_K^2 := \langle f, f \rangle_K$ and $\lambda > 0$ is a positive parameter. We usually require the loss function $V(y, \cdot)$ is convex in $\mathbb{R}$ for every $y \in Y$ in which case (1.3) is a convex optimization problem. The off-line algorithm (1.3) is referred to as a Tikhonov regularization scheme for learning, see e.g. [9] and references therein. This method has been extensively studied in the literature. In particular, its error analysis is well developed due to many results, see e.g. [4, 9, 16, 17, 21]. A central step in these papers has been to show that $f_{\mathbf{z},\lambda}$ is close to the *regularizing function* defined by

$$f_\lambda = \arg \inf_{f \in \mathcal{H}_K} \left\{ \mathcal{E}(f) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \tag{1.4}$$

This expected similarity between $f_{\mathbf{z},\lambda}$ and f_λ is motivated by the theory of uniform convergence, see e.g. [2, 18, 21].

When we are presented with a large number of training samples, solving the convex optimization (1.3) may be practically challenging. Online gradient descent algorithms provide an alternative and efficient way to compute a predictor. To introduce this class of algorithms, we define the *regularized loss function* $\mathcal{V}_\lambda$ for $f \in \mathcal{H}_K$ and $z = (x, y) \in Z$ as $\mathcal{V}_\lambda(z, f) := V(y, f(x)) + \frac{\lambda}{2}\|f\|_K^2$ and use the notation $V_2'(y, \cdot)$ for the left derivative of $V(y, \cdot)$. Since $V(y, \cdot)$ is convex in $\mathrm{I\!R}$ for every $y \in Y$, $V_2'(y, s)$ exists and equals $\sup_{\delta<0}(V(y, s+\delta) - V(y, s))/\delta$ for every $y \in Y, s \in \mathrm{I\!R}$. For brevity's sake, we introduce the Hilbert-valued function

$$\partial\mathcal{V}_\lambda(z, f)(\cdot) := V_2'(y, f(x))K_x(\cdot) + \lambda f(\cdot). \tag{1.5}$$

The random variable $\partial\mathcal{V}_\lambda(z, f)$ plays the role of the gradient of the functional $\mathcal{V}_\lambda(z, \cdot) : \mathcal{H}_K \to \mathrm{I\!R}$ defined above and leads us to the online *stochastic gradient descent* (SGD) algorithm [5, 11, 14, 22] given by

$$\begin{cases} f_1 = 0, \\ f_{t+1} = f_t - \eta_t \partial\mathcal{V}_\lambda(z_t, f_t), \ \text{for } t \in \mathrm{I\!N}_T \end{cases} \tag{1.6}$$

where η_t is called the *step size* and the sequence $\{f_t : t \in \mathrm{I\!N}_{T+1}\}$ the *learning sequence.*

The main focus of the paper is to study the behavior of the random variable $\|f_{T+1} - f_\lambda\|_K$ and provide conditions which ensure its convergence in probability as well as almost surely. This problem is not only interesting in its own right but allows us to derive, for the first time, generalization error rates for the online gradient descent algorithm given in (1.6). Our analysis applies to a general class of loss functions which we call α-admissible (see below). This class of loss functions includes commonly used loss functions for regression and classification. Our main results are presented in Section 2 and proofs are given in Section 3. In Section 4 we give explicit bounds for $\|f_{T+1} - f_\lambda\|_K$ for particular choices of the step sizes as well as for the generalization error of the algorithm (1.6).

2 Main results

In this section, we describe the main results for the online gradient descent algorithm.

2.1 Learning sequence and loss functions

Before the description of our main results for $\|f_{T+1} - f_\lambda\|_K$, it is useful to formulate some notations and observations related to the leaning sequence $\{f_t : t \in \mathbb{N}_{T+1}\}$ and the regularized loss function $\mathcal{V}_\lambda$. In this discussion, we always assume that $V_0 := \sup\big\{|V(y,0)| : y \in Y\big\}$ and $V_0' := \sup\big\{|V_2'(y,0)| : y \in Y\big\}$ are finite. In addition, we introduce the quantity $\kappa := \sup_{x\in X} \sqrt{K(x,x)}$.

By the definition of f_λ, we have $\lambda/2\|f_\lambda\|_K^2 \leq \mathcal{E}(f_\lambda)+\lambda/2\|f_\lambda\|_K^2 \leq \mathcal{E}(0) \leq V_0$ which yields inequality $\|f_\lambda\|_K \leq \sqrt{2V_0/\lambda}$. Thus, it is intuitively reasonable that the learning sequence $\{f_t : t \in \mathbb{N}_{T+1}\}$ is uniformly bounded by a constant depending on λ, since it is used to approximate the bounded function f_λ. Below, we shall verify this intuition for a general class of admissible loss functions.

Definition 1. *Let $V(y,\cdot)$ be convex in $\mathbb{R}$ for every $y \in Y$. We say that V is α-admissible with $0 \leq \alpha < 1$ if*

$$M_\alpha := \sup\Big\{|V_2'(y,s) - V_2'(y,0)|/|s|^\alpha : \ y \in Y, s \in \mathbb{R}\Big\} < \infty$$

and V is 1-admissible if, for all $\lambda > 0$ there holds

$$M_1(\lambda) := \sup\Big\{|V_2'(y,s) - V_2'(y,0)|/|s| : y \in Y, |s| \leq \frac{2\kappa^2 V_0'}{\lambda}\Big\} < \infty.$$

It is noteworthy that we only require local Lipschitz continuity at zero for $\alpha = 1$ in contrast to the uniform Hölder continuity for the case $0 \leq \alpha < 1$. We list below frequently used loss functions which are all α-admissible for some $0 \leq \alpha \leq 1$.

1. Binary classification $Y = \{1, -1\}$:

   (a) q-norm support vector machine (SVM): $V(y,s) = (1-ys)_+^q$ with $q \geq 1$, see [4, 9, 17, 18, 21];

   (b) Least square: $V(y,s) = (1-ys)^2$, see [9, 21];

   (c) Exponential: $V(y,s) = e^{-ys}$, see [12, 21];

   (d) Logistic regression: $V(y,s) = \log(1+e^{-ys})$, see [12, 21].

   The above loss functions are all 1-admissible.

2. Regression $Y = [-M, M]$ for some $M > 0$:

   (a) Least square: $V(y,s) = (y-s)^2$, see [7, 8, 9, 21];

(b) q-norm $\varepsilon-$insensitive regression: $V(y,s) := (|y-s|-\varepsilon)_+^q$ for $q \geq 1$ and $\varepsilon \geq 0$, see [9, 18].

The loss function $V(y,s) := (|y-s|-\varepsilon)_+^q$ is α-admissible with $\alpha = \min\{1, q-1\}$ and is not 1-admissible for $q \in [1,2)$. The least square loss is a special case of $q=2$ and $\varepsilon = 0$.

Now we are ready to present our observation for the learning sequence $\big\{f_t : t \in \mathbb{N}_{T+1}\big\}$. For this purpose, we introduce additional quantities,

$$\mu_\alpha(\lambda) := \begin{cases} \kappa^2 + \lambda, & \text{for } 0 \leq \alpha < 1; \\ \kappa^2 M_1(\lambda) + \lambda, & \text{for } \alpha = 1. \end{cases} \tag{2.1}$$

and

$$C_\alpha(\lambda) = \begin{cases} \frac{2\kappa V_0'}{\lambda} + \frac{2(1-\alpha)}{\lambda}\Big[M_\alpha \kappa \Big(\frac{2\kappa}{\lambda}\Big)^\alpha\Big]^{\frac{1}{1-\alpha}}, & \text{for } 0 \leq \alpha < 1; \\ \frac{2\kappa V_0'}{\lambda}, & \text{for } \alpha = 1. \end{cases} \tag{2.2}$$

One can easily compute the quantities $\mu_\alpha(\lambda), C_\alpha(\lambda)$ for all admissible loss functions mentioned above.

Proposition 1. *If V is α-admissible for some $0 \leq \alpha \leq 1$ and $\eta_t \mu_\alpha(\lambda) \leq 1$ for $t \in \mathbb{N}_T$ then, for any $t \in \mathbb{N}_{T+1}$ there holds*

$$\|f_t\|_K \leq C_\alpha(\lambda). \tag{2.3}$$

For general convex loss functions, we have the following useful observation.

Proposition 2. *If $\lambda > 0$ and $V(y,\cdot)$ is convex in $\mathbb{R}$ for every $y \in Y$ then for any $f, g \in \mathcal{H}_K$, there holds*

(a) $\frac{\lambda}{2}\|f-g\|_K^2 \leq \mathcal{V}_\lambda(z,f) - \mathcal{V}_\lambda(z,g) - \langle \partial \mathcal{V}_\lambda(z,g), f-g\rangle_K$;

(b) $\lambda/2\|f - f_\lambda\|_K^2 \leq \big\{\mathcal{E}(f) + \lambda/2\|f\|_K^2\big\} - \big\{\mathcal{E}(f_\lambda) + \lambda/2\|f_\lambda\|_K^2\big\}$.

We postpone the proofs of these two propositions to the appendix.

2.2 Convergence results and generalization error rates

Equipped with the above observations, we can state our main results. The first result addresses the convergence of $\|f_{T+1} - f_\lambda\|_K$ in probability.

Theorem 1. *Let $\lambda > 0$ and V be α-admissible for some $\alpha \in [0, 1]$. If the step sizes $\{\eta_t : t \in \mathbb{N}\}$ satisfy the inequality $\eta_t \mu_\alpha(\lambda) \leq 1$ for all $t \in \mathbb{N}$ and*

$$\sum_{t \in \mathbb{N}} \eta_t = \infty, \qquad \lim_{t \to \infty} \eta_t = 0, \tag{2.4}$$

then we have weak convergence in probability, that is, for any $\varepsilon > 0$ we have that

$$\lim_{T \to \infty} Prob_{\mathbf{z} \in Z^T} \Big\{ \ \|f_{T+1} - f_\lambda\|_K \geq \varepsilon \Big\} = 0. \tag{2.5}$$

We can improve our convergence result above if additional assumptions on the step sizes are made.

Theorem 2. *Suppose the assumptions in Theorem 3 hold. Moreover, if the function $V(y, \cdot)$ is differentiable in $\mathbb{R}$ for each $y \in Y$ and the step sizes also satisfy $\sum_{t \in \mathbb{N}} \eta_t^2 < \infty$ then the almost surely convergence holds true, that is, for any $\varepsilon > 0$ there holds*

$$\lim_{\ell \to \infty} Prob \Big\{ \sup_{T \geq \ell} \|f_{T+1} - f_\lambda\|_K \geq \varepsilon \Big\} = 0. \tag{2.6}$$

Note that, if the step sizes have the form $\eta_t = O(t^{-\theta}), t \to \infty$ then the hypothesis in Theorem 1 allows the choice $\theta \in (0, 1]$ while Theorem 2 requires $\theta \in (1/2, 1]$.

We shall give proofs of Theorems 1 and 2 in Section 3. In Section 4.1, we shall derive error bounds for $\|f_{T+1} - f_\lambda\|_K$ when the step sizes decay in $\eta_t = O(t^{-\theta}), t \to \infty$ with $0 < \theta \leq 1$. It should be emphasized that the convergence in $\mathcal{H}_K$ yields convergence in $\mathcal{C}^k(X)$ under some conditions on K, where $\mathcal{C}^k$ denotes the space of all functions whose derivatives up to order k are continuous, see [16].

Bounds for the error $\|f_{T+1} - f_\lambda\|_K$ also lead to generalization error rates, as done extensively in the offline literature [15, 16, 17, 21]. Here the goal is to bound the *excess generalization error*

$$\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho^V) \tag{2.7}$$

where $f_\rho^V := \arg\inf \big\{ \mathcal{E}(f) : \ f \text{ is measurable on } X \big\}$. We only demonstrate two examples in the regression problem $Y := [-M, M]$ for some $M > 0$. These examples illustrate how the strong error $\|f_{T+1} - f_\lambda\|_K$ gives rise to the excess generalization error rates by selecting $\lambda = \lambda(T)$ appropriately.

The first example concerns with the least-square loss function. Denote the marginal and conditional distributions of ρ by $\rho_X(\cdot)$ and $\rho(\cdot|x)$ respectively and by $\mathcal{L}^2_{\rho_X}$ the Hilbert space of square integrable functions with norm $\|f\|_\rho := \big(\int_X |f(x)|^2 d\rho_X(x)\big)^{1/2}$. In this case, we have that $f^V_\rho(x) = f_\rho(x) := \int_Y y d\rho(y|x)$ and $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_\rho$. Moreover, we obtain that

$$\|f_{T+1} - f_\rho\|^2_\rho \leq 2\kappa\|f_{T+1} - f_\lambda\|^2_K + 2\|f_\lambda - f_\rho\|^2_\rho. \tag{2.8}$$

The first term on the right hand side of (2.8) has been discussed above. To estimate the second one on the right hand side of the inequality above, we introduce the integral operator defined as $L_K(f) := \int_X K(\cdot, x')f(x')d\rho_X(x')$, $f \in \mathcal{L}^2_{\rho_X}$. Since K is a Mercer kernel, L_K is compact and self-adjoint. Therefore, the fractional power operator L^β_K is well-defined for any $\beta > 0$. We indicate its range space by $L^\beta_K(\mathcal{L}^2_{\rho_X}) := \big\{L^\beta_K(f) : f \in \mathcal{L}^2_{\rho_X}\big\}$. In fact [7], $L^{1/2}_K(\mathcal{L}^2_{\rho_X}) = \mathcal{H}_K$ and $L^\beta_K(\mathcal{L}^2_{\rho_X}) \subseteq \mathcal{H}_K$ for $\beta > 1/2$. It was shown in [16] that $\|f_\lambda - f_\rho\|^2_\rho \leq \lambda^{2\beta}\|L^{-\beta}_K(f_\rho)\|^2_\rho$ if $f_\rho \in L^\beta_K(\mathcal{L}^2_{\rho_X})$ with some $0 < \beta \leq 1$. Now our rate for the least-square online learning algorithm (1.6) reads as follows.

Example 1. Let $V(y, s) = (y - s)^2$ and $f_\rho \in L^\beta_K(\mathcal{L}^2_{\rho_X})$ with some $0 < \beta \leq 1$. For any $0 < \delta < \frac{\beta}{\beta+1}$, choosing $\lambda = T^{\frac{\delta}{2\beta} - \frac{1}{2(\beta+1)}}$ and $\eta_t = \frac{1}{18\kappa^2+5} t^{\frac{(2\beta+1)\delta}{2\beta} - \frac{2\beta+1}{2(\beta+1)}}$, then we have that

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

In addition, if $1/2 < \beta \leq 1$ then, for any $0 < \delta < \frac{2\beta-1}{2\beta+1}$, by selecting $\lambda = T^{-\frac{1}{2\beta+1} + \frac{\delta}{2\beta-1}}$ and $\eta_t = \frac{1}{18\kappa^2+5} t^{\frac{2\beta\delta}{2\beta-1} - \frac{2\beta}{2\beta+1}}$ we have that

$$\mathbb{E}_{\mathbf{z}\in Z^T}\Big[\|f_{T+1} - f_\rho\|^2_K\Big] = O\big(T^{\delta - \frac{2\beta-1}{2\beta+1}}\big). \tag{2.10}$$

It is interesting to point out that our rate (2.10) is almost the same as the rate $O\big(T^{-\frac{2\beta-1}{2\beta+1}}\big)$ established recently by [16] for the least-square offline regularization scheme (1.3).

To get the error rates of (2.7) with general loss functions, we recall the *regularization error* [15, 16] defined as $\mathcal{D}(\lambda) := \inf_{f\in\mathcal{H}_K}\big\{\mathcal{E}(f) - \mathcal{E}(f^V_\rho) + \lambda/2\|f\|^2_K\big\} = \mathcal{E}(f_\lambda) - \mathcal{E}(f^V_\rho) + \lambda/2\|f_\lambda\|^2_K$ which describes the approximation property of $\mathcal{H}_K$. With this at hand, we have the standard *error decomposition*

$$\mathcal{E}(f_{T+1}) - \mathcal{E}(f^V_\rho) \leq \big\{\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)\big\} + \mathcal{D}(\lambda). \tag{2.11}$$

The first term on the right hand side of (2.11) is usually referred to as the *sample error*, which can be estimated by $\|f_{T+1} - f_\lambda\|_K$ as we shall show in

Section 4. For the second term $\mathcal{D}(\lambda)$, in order to get meaningful error rates, we generally assume that it has some polynomial decay $\mathcal{D}(\lambda) = O(\lambda^\beta), \lambda \to 0+$ for some $0 < \beta \le 1$. Next, we only provide the error rates for q-norm ε-*insensitive regression loss* : $V(y,s) = (|y-s|-\varepsilon)_+^q$. Similar results can be derived for other loss functions as well.

Example 2. Let $V(y,s) = (|y-s|-\varepsilon)_+^q$ with $q \ge 1$ and $\varepsilon \ge 0$. Assume that $\mathcal{D}(\lambda) \le c_\beta \lambda^\beta$ for some $0 < \beta \le 1$. Set $\mu(\lambda) := (\kappa+1)^2 \big[q^2 2^q (1 + M + 2\kappa^2 M^{q-1})^q + 4q^2\big] \lambda^{-\max\{q-2,0\}} + \lambda$.

1. For $1 \le q \le 2$ and any $0 < \delta < \frac{\beta}{(q+1)+(3-q)\beta}$, choosing $\lambda = T^{\frac{\delta}{\beta} - \frac{1}{(q+1)+(3-q)\beta}}$ and $\eta_t = \frac{1}{\mu(\lambda)} t^{\frac{(q+(3-q)\beta)\delta}{\beta} - \frac{q+(3-q)\beta}{(q+1)+(3-q)\beta}}$, we have that
$$\mathbb{E}_{\mathbf{z}\in Z^T}\Big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho^V)\Big] = O\Big(T^{\delta - \frac{\beta}{(q+1)+(3-q)\beta}}\Big). \tag{2.12}$$

2. For $q > 2$ and any $0 < \delta < \frac{\beta}{4q+2\beta-2}$, selecting $\lambda := T^{-\frac{1}{4q+2\beta-2}+\frac{\delta}{\beta}}$ $\eta_t = \frac{1}{\mu(\lambda)} t^{-\frac{3q+2\beta-1}{4q+2\beta-2}+\frac{(3q+2\beta-1)\delta}{\beta}}$, we have that
$$\mathbb{E}_{\mathbf{z}\in Z^T}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho^V)\big] = O\big(T^{\delta - \frac{\beta}{4q+2\beta-2}}\big).$$

We shall give the proofs of these two examples in Section 4.

2.3 Related work

There is a vast literature on online gradient descent learning. Let us mention some papers relating to ours. The cumulative loss $1/T \sum_{t=1}^T V(y_t, f_t(x_t))$ for online algorithms more general than (1.6) has been well studied in the literature, see e.g. [3, 5, 6, 10, 22] and references therein. A general regularized online learning scheme (1.6) is introduced and analyzed in [11].

In contrast, in this paper we are interested in the statistical behavior of the last output f_{T+1} of the online algorithm (1.6). In [14], an online gradient descent method in $\mathcal{H}_K$ is considered which mainly focused on least-square regression and specific step sizes $\eta_t = O(t^{-\theta})$ with $\theta \in (1/2, 1)$. For the online gradient descent algorithm (1.6) associated with uniformly Lipschitz loss functions and linear kernels, if $\lambda = 0, \eta_t = \eta$ for any $t \in \mathbb{N}$ then the generalization error bounds were established in [22]. We [20] presented general convergence in probability and error bounds for all commonly used loss functions $V(y,s) = \phi(ys)$ (only 1-admissible) in classification. The main difference between our results here and [20] is that we not only allow α-admissible loss functions for any $0 \leq \alpha \leq 1$, but also obtain almost surely convergence and refine $\mathcal{H}_K$ error rates in some cases (see Corollary 1 in Section 4).

3 Proofs of Theorems 1 and 2

In this section, we prove our main results. We first introduce some notations. Set $\omega_k^T(\lambda) = \prod_{j=k+1}^T (1-\eta_j\lambda)$ for $k \in \mathbb{N}_{T-1}$ where $\omega_T^T(\lambda) = 1$. Using (2.3) in Proposition 1, if α-admissible loss function and the step sizes satisfy $\eta_t \mu_\alpha(\lambda) \leq 1$ for $t \in \mathbb{N}_T$ then, for $t \in \mathbb{N}_{T+1}$ we have

$$\|\partial \mathcal{V}_\lambda(z_t, f_t)\|_K \leq |V_2'(y, f_t(x_t))| \sqrt{K(x_t, x_t)} + \lambda \|f_t\|_K \leq \nu_\alpha(\lambda) \tag{3.1}$$

with $\nu_\alpha(\lambda) := \kappa \sup\big\{|V_2'(y,t)| : |t| \leq \kappa C_\alpha(\lambda),\ y \in Y\big\} + \lambda C_\alpha(\lambda)$. Denote the expectation $\mathbb{E}_{z_1,\ldots,z_t}$ as $\mathbb{E}_{Z^t}$ and $\mathcal{R}_t = f_t - f_\lambda$ for $t \in \mathbb{N}$, we have the following intermediate lemma.

Lemma 1. *If $\lambda > 0$, V is α-admissible for some $\alpha \in [0,1]$ and $\eta_t \mu_\alpha(\lambda) \leq 1$ for $t \in \mathbb{N}_T$ then there holds:*

(a) $\mathbb{E}_{\mathbf{z} \in Z^T}\big[\|\mathcal{R}_{T+1}\|_K^2\big] \leq \frac{2\mathcal{D}(\lambda)}{\lambda} \omega_0^T(\lambda) + \big(\nu_\alpha(\lambda)\big)^2 \sum_{k=1}^T \eta_k^2 \omega_k^T(\lambda)$;

(b) For any $\ell \leq T$, $\sup_{\ell \leq t \leq T} \|\mathcal{R}_{t+1}\|_K^2$ is bounded by

$$\|\mathcal{R}_\ell\|_K^2 + \big(\nu_\alpha(\lambda)\big)^2 \sum_{t=\ell}^\infty \eta_t^2 + 2 \sup_{\ell \leq k \leq T} \Big| \sum_{t=\ell}^k \eta_t \langle \partial \mathcal{V}_\lambda(z_t, f_\lambda), \mathcal{R}_t \rangle_K \Big|.$$

Proof. The definition of the online algorithm (1.6) tells us that $\mathcal{R}_{t+1} = \mathcal{R}_t - \eta_t \partial \mathcal{V}_\lambda(z_t, f_t)$. Hence,

$$\|\mathcal{R}_{t+1}\|_K^2 = \|\mathcal{R}_t\|_K^2 + \eta_t^2 \|\partial \mathcal{V}_\lambda(z_t, f_t)\|_K^2 + 2\eta_t \langle \partial \mathcal{V}_\lambda(z_t, f_t), f_\lambda - f_t \rangle_K. \tag{3.2}$$

By the reproducing property (1.2) and the convexity of $V(y_t, \cdot)$, we have that

$$\begin{aligned} \langle V_2'(y_t, f_t(x_t)) K_{x_t}, f_\lambda - f_t \rangle_K &= V_2'(y_t, f_t(x_t))(f_\lambda(x_t) - f_t(x_t)) \\ &\leq V(y_t, f_\lambda(x_t)) - V(y_t, f_t(x_t)). \end{aligned} \tag{3.3}$$

Observe also that $\lambda \langle f_t, f_\lambda - f_t \rangle_K \leq \lambda/2 \|f_\lambda\|_K^2 - \lambda/2 \|f_t\|_K^2$. Putting this and (3.3) together, from the definition (1.5) of $\partial \mathcal{V}_\lambda$ we have

$$\langle \partial \mathcal{V}_\lambda(z_t, f_t), f_\lambda - f_t \rangle_K \leq \mathcal{V}_\lambda(z_t, f_\lambda) - \mathcal{V}_\lambda(z_t, f_t). \tag{3.4}$$

Since f_t depends on $\{z_1, z_2, \cdots, z_{t-1}\}$ but not on z_t, it follows that

$$\begin{aligned}&\mathbb{E}_{Z^t}\big[\langle \partial\mathcal{V}_\lambda(z_t, f_t), f_\lambda - f_t\rangle_K\big] \leq \mathbb{E}_{Z^{t-1}}\big[\mathbb{E}_{z_t}\big[\mathcal{V}_\lambda(z_t, f_\lambda) - \mathcal{V}_\lambda(z_t, f_t)\big]\big] \\ &= \big(\mathcal{E}(f_\lambda) + \tfrac{\lambda}{2}\|f_\lambda\|_K^2\big) - \mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t) + \tfrac{\lambda}{2}\|f_t\|_K^2\big].\end{aligned}$$

Putting this into (3.2) yields

$$\begin{aligned}\mathbb{E}_{Z^t}\big[\|\mathcal{R}_{t+1}\|_K^2\big] &\leq \mathbb{E}_{Z^{t-1}}\big[\|\mathcal{R}_t\|_K^2\big] + \eta_t^2\mathbb{E}_{Z^t}\big[\|\partial\mathcal{V}_\lambda(z_t, f_t)\|_K^2\big] \\ &+2\eta_t\mathbb{E}_{Z^{t-1}}\big[\big(\mathcal{E}(f_\lambda) + \tfrac{\lambda}{2}\|f_\lambda\|_K^2\big) - \big(\mathcal{E}(f_t) + \tfrac{\lambda}{2}\|f_t\|_K^2\big)\big].\end{aligned}$$

Substituting property (b) of Proposition 2 with $f = f_t$ and the bound (3.1) into (3.5), we have $\mathbb{E}_{Z^t}\big[\|\mathcal{R}_{t+1}\|_K^2\big] \leq (1 - \eta_t\lambda)\mathbb{E}_{Z^{t-1}}\big[\|\mathcal{R}_t\|_K^2\big] + \eta_t^2\big(\nu_\alpha(\lambda)\big)^2$. By induction, we obtain the first argument by noting $\lambda/2\|f_\lambda\|_K^2 \leq \mathcal{D}(\lambda)$.

We now prove property (b). Applying property (a) of Proposition 2 with $g = f_\lambda, f = f_t$ to the right hand side of (3.4), we know that $\langle\partial\mathcal{V}_\lambda(z_t, f_t), f_\lambda - f_t\rangle_K \leq \langle\partial\mathcal{V}_\lambda(z_t, f_\lambda), f_\lambda - f_t\rangle_K - \lambda/2\|f_t - f_\lambda\|_K^2$. Substituting this and (3.1) into (3.2), we obtain $\|\mathcal{R}_{t+1}\|_K^2 \leq \|\mathcal{R}_t\|_K^2 + \eta_t^2\big(\nu_\alpha(\lambda)\big)^2 + 2\eta_t\langle\partial\mathcal{V}_\lambda(z_t, f_\lambda), f_\lambda - f_t\rangle_K$. By induction, we know for any $\ell \leq k \leq T$ that

$$\begin{aligned}\|\mathcal{R}_{k+1}\|_K^2 - \|\mathcal{R}_\ell\|_K^2 &\leq \big(\nu_\alpha(\lambda)\big)^2\sum_{t=\ell}^{k}\eta_t^2 - 2\sum_{t=\ell}^{k}\eta_t\langle\partial\mathcal{V}_\lambda(z_t, f_\lambda), \mathcal{R}_t\rangle_K \\ &\leq \big(\nu_\alpha(\lambda)\big)^2\sum_{t=\ell}^{\infty}\eta_t^2 + 2\sup_{\ell\leq k\leq T}\Big|\sum_{t=\ell}^{k}\eta_t\langle\partial\mathcal{V}_\lambda(z_t, f_\lambda), \mathcal{R}_t\rangle_K\Big|.\end{aligned}$$

Since $k(\ell \leq k \leq T)$ is arbitrary, the above inequality yields the desired result. □

Now we turn our attention to the proofs of our theorems using Lemma 1.

Proof of Theorem 1. Since $\eta_t\mu_\alpha(\lambda) \leq 1$ for all $t \in \mathbb{N}$, Proposition 1 and property (a) in Lemma 1 hold true for every $T \in \mathbb{N}$. Hence, we only need to show $\lim_{T\to\infty}\omega_0^T(\lambda) = 0$ and $\lim_{T\to\infty}\sum_{k=1}^T\eta_k^2\omega_k^T(\lambda)$ under the assumption (2.4). This is exactly included in the proof Theorem 1 in [20] for 1-admissible loss functions in classification. □

To prove Theorem 2, we need additional lemmas. First we need a property of the regularization function f_λ.

Lemma 2. *If $V(y,\cdot)$ is differentiable and convex for every $y \in Y$ then there holds*

$$\int_Z \partial\mathcal{V}_\lambda(z, f_\lambda)d\rho(z) = \int_Z V_2'(y, f_\lambda(x))K_x d\rho + \lambda f_\lambda = 0. \qquad (3.5)$$

Proof. For $\lambda > 0$, we introduce the functional $\mathcal{Q} : \mathcal{H}_K \to \mathbb{R}$ defined as $\mathcal{Q}(f) := \mathcal{E}(f) + \frac{\lambda}{2}\|f\|_K^2$ for any $f \in \mathcal{H}_K$. Since $V(y, \cdot)$ is differentiable and convex for every $y \in Y$, we know that $\mathcal{Q}$ is differentiable and strictly convex. Therefore, it has a unique minimizer which we have called f_λ. Moreover, f_λ is determined by the fact that the gradient of $\mathcal{Q}$ at f_λ is zero. Indeed, it can be verified for any $f, g \in \mathcal{H}_K$ that

$$\begin{aligned}\lim_{h\to 0}\frac{\mathcal{Q}(f+hg)-\mathcal{Q}(f)}{h} &= \langle \int_Z V_2'(y, f(x))K_x + \lambda f, g\rangle_K d\rho(z)\\ &= \langle \int_X V_2'(y, f(x))K_x d\rho_X(x) + \lambda f, g\rangle_K\\ &= \langle \partial\mathcal{V}_\lambda(z, f), g\rangle_K.\end{aligned}$$

This proves the equality (3.5). □

We also need a probabilistic inequality. Denote $Y_t = Y(z_1, \ldots, z_t) : Z^t \to \mathbb{R}$ for $t \in \mathbb{N}_T$ with $a_t \leq Y_t \leq b_t$. The sequence $\{Y_t : t \in \mathbb{N}_T\}$ is called a real-valued *martingale difference* sequence if $\mathbb{E}_{z_t}\Big[Y_t\big|z_1, \ldots, z_{t-1}\Big] = \int_Z Y_t d\rho(z_t) = 0$. Then, the probability inequality (3.30) of [13] tells us that, for any $\epsilon > 0$ there holds

$$\text{Prob}_{\mathbf{z}\in Z^T}\left\{\sup_{1\leq t\leq T}\Big|\sum_{k=1}^t Y_k\Big| \geq \epsilon\right\} \leq 2\exp\left\{-\frac{2\epsilon^2}{\sum_{t=1}^T (b_t - a_t)^2}\right\}. \tag{3.6}$$

We now can present the proof of Theorem 2.

Proof of Theorem 2. Since $\sum_{t=1}^\infty \eta_t^2 < \infty$, then there exists $\ell_1(\varepsilon) \in \mathbb{N}$ such that, for all $\ell \geq \ell_1(\varepsilon)$ we have that $\big(\nu_\alpha(\lambda)\big)^2 \sum_{t=\ell}^\infty \eta_t^2 \leq \varepsilon$. By property (b) in Lemma 1, we have that

$$\begin{aligned}\text{Prob}\Big\{ \quad \sup\nolimits_{\ell\leq t\leq T}\|\mathcal{R}_{t+1}\|_K^2 \geq 4\varepsilon\Big\} \leq\ & \text{Prob}\Big\{\ \|\mathcal{R}_\ell\|_K^2 \geq \varepsilon\Big\}\\ &+ \text{Prob}\Big\{\ \sup\nolimits_{\ell\leq k\leq T}\Big|\sum_{t=\ell}^k \eta_t\langle\partial\mathcal{V}_\lambda(z_t, f_\lambda), \mathcal{R}_t\rangle_K\Big| \geq \varepsilon\Big\}.\end{aligned}$$

By Theorem 1, the first term tends to zero as $\ell \to \infty$.

Next, we use (3.6) to estimate the second term on the right hand side of (3.7). To this end, we set the sequence $Y_t = \langle\partial\mathcal{V}_\lambda(z_t, f_\lambda), \mathcal{R}_t\rangle_K, \ell \leq t \leq T$ and zero for $1 \leq t < \ell$. Since f_t is only dependent on $\{z_k : k \in \mathbb{N}_{t-1}\}$, so is $\mathcal{R}_t$. Combining this fact with the assumption that $V(y, \cdot)$ is differentiable, from (3.5) we know that $\mathbb{E}_{z_t}[Y_t | z_1, \ldots, z_{t-1}] = \langle \int_Z \partial \mathcal{V}_\lambda(z_t, f_\lambda) d\rho(z_t), \mathcal{R}_t \rangle_K = 0$ for all $1 \leq t \leq T$. This implies that $\{Y_t : t \in \mathbb{N}_T\}$ is a martingale difference sequence. Also, since $\eta_t \mu_\alpha(\lambda) \leq 1$ for any $t \in \mathbb{N}$, Proposition 1 and the fact that $\lambda/2 \|f_\lambda\|_K^2 \leq \mathcal{E}(0) \leq V_0$ tell us that $\|\mathcal{R}_t\|_K \leq \|f_t\|_K + \|f_\lambda\|_K \leq \widetilde{\nu}_\alpha(\lambda) := C_\alpha(\lambda) + \sqrt{2V_0/\lambda}$ for any $t \in \mathbb{N}$. This in connection with (3.1) yields $|\langle \partial \mathcal{V}_\lambda(z_t, f_t), \mathcal{R}_t \rangle_K| \leq \nu_\alpha(\lambda) \widetilde{\nu}_\alpha(\lambda)$. Therefore, applying (3.6) with $b_t = -a_t = \eta_t \nu_\alpha(\lambda) \widetilde{\nu}_\alpha(\lambda)$ for $t \geq \ell$ and $b_t = a_t = 0$ for $t < \ell$, we have that

$$\begin{aligned} \text{Prob}\Big\{ \sup_{\ell \leq k \leq T} \Big| \sum_{t=\ell}^{k} \eta_t \langle \partial \mathcal{V}_\lambda(z_t, f_\lambda), \mathcal{R}_t \rangle_K \Big| \geq \varepsilon \Big\} \\ \leq 2 \exp\Big\{ \frac{-\varepsilon^2}{2\big(\nu_\alpha(\lambda) \widetilde{\nu}_\alpha(\lambda)\big)^2 \sum_{t=\ell}^{\infty} \eta_t^2} \Big\}. \end{aligned} \tag{3.7}$$

Noting $\Big\{ \sup_{t \geq \ell} \|\mathcal{R}_{t+1}\|_K^2 \geq \varepsilon \Big\} = \bigcup_{T \geq \ell} \Big\{ \sup_{\ell \leq t \leq T} \|\mathcal{R}_{t+1}\|_K^2 \geq \varepsilon \Big\}$, we have that $\text{Prob} \Big\{ \sup_{t \geq \ell} \|\mathcal{R}_{t+1}\|_K^2 \geq \varepsilon \Big\} = \lim_{T \to \infty} \text{Prob} \Big\{ \sup_{\ell \leq t \leq T} \|\mathcal{R}_{t+1}\|_K^2 \geq \varepsilon \Big\}$. Hence, the result follows from (3.7) and $\lim_{\ell \to \infty} \sum_{t=\ell}^{\infty} \eta_t^2 = 0$. □

4 Error rates

In this section, we derive error rates for $\|f_{T+1} - f_\lambda\|_K$ for the specific step sizes $\eta_t = \frac{1}{\mu_\alpha(\lambda)} t^{-\theta}$ with $\theta \in (0, 1]$ and apply them to get the error rates for the excess generalization error.

4.1 Strong error rates in RKHS norm

In order to apply Lemma 1 to get error bounds for $\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big]$, we need to estimate the summation $\sum_{k=1}^{T} \eta_k^2 \omega_k^T(\lambda)$ in property (a) of Lemma 1. This leads to the following lemma whose proof can be found in [20].

Lemma 3. *Let $0 < \nu \leq 1$. Then we can bound* $\displaystyle\sum_{t=1}^{T-1} \frac{1}{t^{2\theta}} \exp\Big\{ -\nu \sum_{j=t+1}^{T} j^{-\theta} \Big\}$ *by*

$$\begin{cases} \frac{18}{\nu T^\theta} + \frac{9 T^{1-\theta}}{(1-\theta) 2^{1-\theta}} \exp\Big\{ -\frac{\nu(1-2^{\theta-1})}{1-\theta} (T+1)^{1-\theta} \Big\}, & 0 < \theta < 1; \\ \frac{8}{1-\nu} (T+1)^{-\nu}, & \theta = 1. \end{cases}$$

Now we present general bounds for $\|f_{T+1} - f_\lambda\|_K$.

Theorem 3. *Let $\lambda > 0$, V be α-admissible, $\alpha \in [0,1]$ and $\eta_t = \frac{1}{\mu(\lambda)} t^{-\theta}$ with $\mu(\lambda) \geq \mu_\alpha(\lambda)$ and $\theta \in (0,1]$. Define $\{f_t : t \in \mathbb{N}_{T+1}\}$ by (1.6) and $\nu_\alpha(\lambda)$ by (3.1).*

1. For $0 < \theta < 1$, $\mathbb{E}_{\mathbf{z}\in Z^T}\left[\|f_{T+1} - f_\lambda\|_K^2\right]$ is bounded by

$$\Big(\frac{2\mathcal{D}(\lambda)}{\lambda} + \frac{9(\nu_\alpha(\lambda))^2 T^{1-\theta}}{(1-\theta)\mu^2(\lambda)}\Big) \exp\Big\{-\frac{\lambda(1-2^{\theta-1})}{\mu(\lambda)(1-\theta)}(T+1)^{1-\theta}\Big\} + \frac{19\big(\nu_\alpha(\lambda)\big)^2}{\lambda\mu(\lambda)T^\theta}. \tag{4.1}$$

2. For $\theta = 1$, then we get

$$\mathbb{E}_{\mathbf{z}\in Z^T}\left[\|f_{T+1} - f_\lambda\|_K^2\right] \leq \Big(\frac{2\mathcal{D}(\lambda)}{\lambda} + \frac{9\nu_\alpha^2(\lambda)}{\mu(\lambda)(\mu(\lambda)-\lambda)}\Big) T^{-\frac{\lambda}{\mu(\lambda)}}. \tag{4.2}$$

Proof. Note $1 - x \leq e^{-x}$ for all $x \geq 0$. Then, property (a) in Lemma 1 and the bound $\|f_\lambda\|_K^2 \leq 2\mathcal{D}(\lambda)/\lambda$ imply $\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big] \leq I_1 + I_2$, where

$$I_1 = \frac{2\mathcal{D}(\lambda)}{\lambda} \exp\Big\{-\sum_{t=1}^T \frac{\lambda}{\mu(\lambda)t^\theta}\Big\}, \quad I_2 = \Big(\frac{\nu_\alpha(\lambda)}{\mu(\lambda)}\Big)^2 \sum_{t=1}^T \frac{1}{t^{2\theta}} \exp\Big\{-\frac{\lambda}{\mu(\lambda)} \sum_{j=t+1}^T j^{-\theta}\Big\}.$$

The first quantity can be estimated as

$$I_1 \leq \begin{cases} \frac{2\mathcal{D}(\lambda)}{\lambda} \exp\Big\{-\frac{(1-2^{\theta-1})\lambda}{(1-\theta)\mu(\lambda)}(T+1)^{1-\theta}\Big\}, & \text{if } 0 < \theta < 1, \\ \frac{2\mathcal{D}(\lambda)}{\lambda}(T+1)^{-\frac{\lambda}{\mu(\lambda)}}, & \text{if } \theta = 1. \end{cases}$$

Applying Lemma 3 with $\nu = \frac{\lambda}{\mu(\lambda)}$, the second term I_2 can estimated as

$$\begin{cases} \Big(\frac{\nu_\alpha(\lambda)}{\mu(\lambda)}\Big)^2 \Big(\frac{18\mu(\lambda)/\lambda}{T^\theta} + \frac{9T^{1-\theta}}{(1-\theta)} \exp\big\{-\frac{(1-2^{\theta-1})\lambda}{(1-\theta)\mu(\lambda)}(T+1)^{1-\theta}\big\} + \frac{1}{T^{2\theta}}\Big), & \text{if } \theta \in (0,1), \\ \Big(\frac{\nu_\alpha(\lambda)}{\mu(\lambda)}\Big)^2 \Big(\frac{8}{1-\lambda/\mu(\lambda)}(T+1)^{-\frac{\lambda}{\mu(\lambda)}} + \frac{1}{T^2}\Big), & \text{if } \theta = 1. \end{cases}$$

The proof is completed. □

Although Theorem 1 ensures that $\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big]$ converges to zero for $\eta_t = O(t^{-1})$ for fixed $\lambda > 0$, Theorem 3 tells us the rate (4.2) is unacceptably slow since λ is usually very small. Hence, we will not consider this degenerate case here.

In some cases, using the special nature of the loss function, we can refine the $\mathcal{H}_K$ rates by reducing $(\nu_\alpha(\lambda))^2$ in the last term of (4.1) to an absolute constant independent of λ. Loss functions like $V(y,s) = (|y-s| - \varepsilon)_+^q$ for

$1 \le q \le 2$, share a common property that there exist $a \ge 0$ and $b \ge 0$ such that

$$(V_2'(y,s))^2 \le a + bV(y,s), \qquad \text{for any } y \in Y, s \in \mathbb{R}. \tag{4.3}$$

Corollary 1. *Let $0 < \lambda \le 1$, V be α-admissible for some $0 \le \alpha \le 1$ and satisfy (4.3). Assume the step size $\eta_t = \frac{1}{\mu(\lambda)} t^{-\theta}$ with some $\mu(\lambda) \ge \mu_\alpha(\lambda) + 4(b\kappa^2+1)$ and $\theta \in (0,1)$. Then $\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big]$ is bounded by*

$$\Big(\frac{2\mathcal{D}(\lambda)}{\lambda} + \frac{18c_1 T^{1-\theta}}{(1-\theta)\mu^2(\lambda)}\Big)\exp\Big\{-\frac{\lambda(1-2^{\theta-1})}{2\mu(\lambda)(1-\theta)}(T+1)^{1-\theta}\Big\} + \frac{74c_1}{\mu(\lambda)\lambda T^\theta} \tag{4.4}$$

where $c_1 = a\kappa^2 + 2(b\kappa^2+1)$.

Proof. Recall $\mathcal{R}_t = f_t - f_\lambda$ and argue as in the proof of Lemma 1, we already have (3.5):

$$\begin{aligned}\mathbb{E}_{Z^t}\big[\|\mathcal{R}_{t+1}\|_K^2\big] &\le \mathbb{E}_{Z^{t-1}}\big[\|\mathcal{R}_t\|_K^2\big] + \eta_t^2 \mathbb{E}_{Z^t}\big[\|\partial\mathcal{V}_\lambda(z_t,f_t)\|_K^2\big] \\ &+2\eta_t \mathbb{E}_{Z^{t-1}}\Big[\big(\mathcal{E}(f_\lambda) + \tfrac{\lambda}{2}\|f_\lambda\|_K^2\big) - \big(\mathcal{E}(f_t) + \lambda/2\|f_t\|_K^2\big)\Big].\end{aligned}$$

We estimate $\mathbb{E}_{Z^t}\big[\|\partial\mathcal{V}_\lambda(z_t,f_t)\|_K^2\big]$ by the special feature (4.3) rather than using (2.3) and (3.1) directly.

By the definition of $\partial\mathcal{V}_\lambda(z_t,f_t)$, we note from (4.3) that $\|\partial\mathcal{V}_\lambda(z_t,f_t)\|_K^2 \le 2\kappa^2|V_2'(y_t,f_t(x_t))|^2 + 2\lambda^2\|f_t\|_K^2 \le 2\kappa^2(a + bV(y_t,f_t(x_t))) + 2\lambda\|f_t\|_K^2$. This in connection with the fact that f_t is independent of z_t implies

$$\begin{aligned}\mathbb{E}_{Z^t}\big[\|\partial\mathcal{V}_\lambda(z_t,f_t)\|_K^2\big] &\le 2a\kappa^2 + 2(b\kappa^2+1)\mathbb{E}_{Z^t}\big[V(y_t,f_t(x_t)) + \lambda\|f_t\|_K^2\big] \\ &= 2a\kappa^2 + 2(b\kappa^2+1)\mathbb{E}_{Z^{t-1}}\big[\mathcal{E}(f_t) + \lambda\|f_t\|_K^2\big] \\ &\le 2a\kappa^2 + 4(b\kappa^2+1)\mathbb{E}_{Z^{t-1}}\big[\big(\mathcal{E}(f_t) + \lambda/2\|f_t\|_K^2\big) \\ &\quad -\big(\mathcal{E}(f_\lambda) + \lambda/2\|f_\lambda\|_K^2\big)\big] + 4(b\kappa^2+1)V_0\end{aligned}$$

where we used $\mathcal{E}(f_\lambda) + \lambda/2\|f_\lambda\|_K^2 \le \mathcal{E}(0) \le V_0$ in the last inequality.

Substituting this into (4.5), we know

$$\begin{aligned}\mathbb{E}_{Z^t}\big[\|\mathcal{R}_{t+1}\|_K^2\big] &\le \mathbb{E}_{Z^{t-1}}\big[\|\mathcal{R}_t\|_K^2\big] + 2(a\kappa^2 + 2(b\kappa^2+1)V_0)\eta_t^2 \\ &+2\eta_t(1-2(b\kappa^2+1)\eta_t)\mathbb{E}_{Z^{t-1}}\Big[\big(\mathcal{E}(f_\lambda) + \tfrac{\lambda}{2}\|f_\lambda\|_K^2\big) - \big(\mathcal{E}(f_t) + \lambda/2\|f_t\|_K^2\big)\Big].\end{aligned}$$

Also, $\eta_t = \frac{1}{\mu(\lambda)}t^{-\theta} \le \frac{1}{\mu_\alpha(\lambda)+4(b\kappa^2+1)}t^{-\theta}$ implies $1 - 2(b\kappa^2+1)\eta_t \ge 1/2$. Applying property (b) of Proposition 2 with $f = f_t$, we get $\mathbb{E}_{Z^t}\big[\|\mathcal{R}_{t+1}\|_K^2\big] \le (1 - \frac{\eta_t\lambda}{2})\mathbb{E}_{Z^{t-1}}\big[\|\mathcal{R}_t\|_K^2\big] + 2c_1\eta_t^2$ where we set $c_1 = a\kappa^2 + 2(b\kappa^2+1)V_0$. Recall $\omega_k^T(\lambda/2) = \prod_{j=k+1}^T(1-\eta_j\lambda/2)$, by induction we have

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\in Z^T}\big[\|\mathcal{R}_{T+1}\|_K^2\big] &\le \omega_0^T\big(\tfrac{\lambda}{2}\big)\|f_\lambda\|_K^2 + 2c_1\textstyle\sum_{k=1}^T \eta_k^2\omega_k^T\big(\tfrac{\lambda}{2}\big) \\ &\le \tfrac{2\mathcal{D}(\lambda)}{\lambda}\omega_0^T\big(\tfrac{\lambda}{2}\big) + 2c_1\textstyle\sum_{k=1}^T \eta_k^2\omega_k^T\big(\tfrac{\lambda}{2}\big).\end{aligned}$$

Using lemma 3 with $\nu = \frac{\lambda}{2\mu(\lambda)}$ and arguing exactly as in the proof of Theorem 3 give us the desired corollary. □

Equipped with the above results we can now derive the rates of the excess generalization error (2.7) for the online gradient descent algorithm (1.6).

4.2 Example 1: least-square regression

We turn our attention to the least-square regression.

Proof of Example 1. We apply Corollary 1. Note that $V_0' = 2M$, V is 1-admissible and $M_1(\lambda) = 2, \mu_1(\lambda) = 2\kappa^2 + \lambda$. Also, (4.3) holds with $a = 0, b = 4$. From (2.8), we know that $\|f_{T+1} - f_\rho\|_\rho^2 \leq 2\kappa \|f_{T+1} - f_\lambda\|_K^2 + 2\|f_\lambda - f_\rho\|_\rho^2$. We estimate the righthand side of the above inequality as follows. First, note, for all $\epsilon > 0, s > 0$ and $c > 0$ the asymptotic behavior holds

$$\exp\{-cT^\epsilon\} = O(T^{-s}). \tag{4.5}$$

Applying Corollary 1 with $\mu(\lambda) = 18\kappa^2 + 5 \geq \mu_1(\lambda) + 4(b\kappa^2 + 1)$ and $\lambda = T^{-\gamma}$ with $0 < \gamma < \min\{1 - \theta, \theta\}$, we know from (4.5) that the first term of (4.4) decays in the form of $O(T^{-s})$ for any large $s > 0$. However, the second term of (4.4) is bounded by $O(T^{-(\theta-\gamma)})$. Consequently,

$$\mathbb{E}_{\mathbf{z}\in Z^T}\big[\|f_{T+1} - f_\lambda\|_K^2\big] \leq O\big(T^{-(\theta-\gamma)}\big). \tag{4.6}$$

Second, we know from Lemma 3 in [16] that if $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ for some $0 < \beta \leq 1$ then the second term of (2.8) can be bounded as $\|f_\lambda - f_\rho\|_\rho^2 \leq \lambda^{2\beta}\|L_K^{-\beta} f_\rho\|_\rho^2$. Now, putting this and (4.6) into (2.8), with $\lambda = T^{-\gamma}$ and $\gamma < \min\{1 - \theta, \theta\}$, yields $\mathbb{E}_{\mathbf{z}\in Z^T}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] \leq O\big(T^{-(\theta-\gamma)} + T^{-2\beta\gamma}\big)$. Choosing $0 < \gamma = \theta/(2\beta + 1) = \frac{1}{2(\beta+1)} - \frac{\delta}{2\beta}(< \min\{1 - \theta, \theta\})$, we obtain that the desired bound $\mathbb{E}\big[\|f_{T+1} - f_\rho\|_\rho^2\big] = O\big(T^{\delta - \frac{\beta}{\beta+1}}\big)$.

For the rate in the $\mathcal{H}_k$ norm, we note that $\|f_{T+1} - f_\rho\|_K^2 \leq 2\|f_{T+1} - f_\lambda\|_K^2 + 2\|f_\lambda - f_\rho\|_K^2$. If the regression function $f_\rho \in L_K^\beta(\mathcal{L}^2_{\rho_X})$ for $1/2 < \beta \leq 1$, then [16] $\|f_\lambda - f_\rho\|_K^2 \leq \lambda^{2\beta-1}\|L_K^{-\beta} f_\rho\|_\rho^2$. Combining this inequality with (4.6), by selecting $\lambda = T^{-\gamma}$ with $0 < \gamma < \min\{1 - \theta, \theta\}$, we have that $\mathbb{E}_{\mathbf{z}\in Z^T}\big[\|f_{T+1} - f_\rho\|_K^2\big] \leq O\big(T^{-(\theta-\gamma)} + T^{-(2\beta-1)\gamma}\big)$. Selecting $0 < \gamma = \frac{\theta}{2\beta} = \frac{1}{2\beta+1} - \frac{\delta}{2\beta-1}$ yields the desired result. □

4.3 Example 2: ε-insensitive regression

To present the rate for ε-insensitive regression loss function, we need to estimate the sample error $\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)$ using the error $\|f_{T+1} - f_\lambda\|_K$.

Lemma 4. *If $Y = [-M, M]$ for some $M > 0$, $V(y,s) = (|y-s| - \varepsilon)_+^q$ with some $q \geq 1, \varepsilon \geq 0$ and $f \in \mathcal{H}_K$, then $|\mathcal{E}(f) - \mathcal{E}(f_\lambda)|$ is bounded by*

$$q\kappa[2(M+\kappa)]^{q-1}\Bigg[(1+\sqrt{2\mathcal{D}(\lambda)/\lambda})^{q-1}\|f - f_\lambda\|_K + \|f - f_\lambda\|_K^q\Bigg]. \tag{4.7}$$

Proof. Since $|y| \leq M$, we have the inequality $|(|y-u|-\varepsilon)_+^q - (|y-v|-\varepsilon)_+^q| \leq q(M + \max(|u|,|v|))^{q-1}|u-v| \leq q(M + |u-v| + |v|)^{q-1}|u-v|$. By taking $u = f(x), v = f_\lambda(x)$, $|\mathcal{E}(f) - \mathcal{E}(f_\lambda)|$ can be bounded by

$$\begin{aligned}
& q\kappa\Bigg(\big[M + \kappa\|f - f_\lambda\|_K + \kappa\|f_\lambda\|_K\big]^{q-1}\|f - f_\lambda\|_K\Bigg) \\
& \leq q\kappa\Bigg(\big[M + \kappa\|f - f_\lambda\|_K + \kappa\sqrt{2\mathcal{D}(\lambda)/\lambda}\big]^{q-1}\|f - f_\lambda\|_K\Bigg) \\
& \leq q\kappa[2(M+\kappa)]^{q-1}\Bigg[\big(1+\sqrt{2\mathcal{D}(\lambda)/\lambda}\big)^{q-1}\|f - f_\lambda\|_K + \|f - f_\lambda\|_K^q\Bigg]
\end{aligned}$$

where we used the bound $\|f_\lambda\|_K^2 \leq 2\mathcal{D}(\lambda)/\lambda$ in the first inequality. □

Proof of Example 2. The loss function $V(y,s) = (|y-s| - \varepsilon)_+^q$ is $\alpha = \min\{q-1, 1\}$-admissible. We can calculate the constants $V_0' \leq qM^{q-1}$ for $q \geq 1$, $M_\alpha = q$ for $q \leq 2$ and $M_1(\lambda) \leq \sup\{|V''(y,s)| : y \in Y, |s| \leq \frac{2\kappa V_0'}{\lambda}\} \leq q^2 2^q(1 + M + 2\kappa^2 M^{q-1})^q \lambda^{-(q-2)}$ for $q > 2$ and $0 < \lambda \leq 1$. By the definition (2.1) of $\mu_\alpha(\lambda)$, these give us $\mu_\alpha(\lambda) + 4(q^2\kappa^2 + 1) \leq \mu(\lambda) := (\kappa+1)^2(q^2 2^q(1 + M + 2\kappa^2 M^{q-1})^q + 4q^2)\lambda^{-\max\{q-2,0\}} + \lambda$ for all $q \geq 1$ and $0 < \lambda \leq 1$.

For $1 \leq q \leq 2$, we apply Corollary 1 with $a = b = q^2$ and note $\mu(\lambda) \geq \mu_\alpha(\lambda) + 4(q^2\kappa^2+1)$ for $0 < \lambda \leq 1$. By (4.5), with $\lambda = T^{-\gamma}$ and $\gamma < \min\{1-\theta, \theta\}$, the bound (4.4) yields $\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K\big] = O\big(T^{-(\theta-\gamma)/2}\big)$. From (4.7), we note that $\mathbb{E}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)\big] = O\big(\lambda^{\frac{(\beta-1)(q-1)}{2}}\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K\big] + \mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^q\big]\big) \leq O\big(\lambda^{\frac{(\beta-1)(q-1)}{2}}\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K\big] + \big(\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K^2\big]\big)^{q/2}\big)$ where we used Hölder inequality for $1 \leq q < 2$. Consequently, $\mathbb{E}_{\mathbf{z}\in Z^T}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)\big] = O\big(T^{\frac{\gamma(q+(1-q)\beta)-\theta}{2}}\big)$. Combining this with the regularization rate $\mathcal{D}(\lambda) = O\big(T^{-\beta\gamma}\big)$, from (2.11) we have $\mathbb{E}_{\mathbf{z}\in Z^T}\big[\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho^V)\big] = O\big(T^{\frac{\gamma(q+(1-q)\beta)-\theta}{2}} + T^{-\beta\gamma}\big)$. Thus, choosing $0 < \gamma = \frac{\theta}{q+(3-q)\beta} = \frac{1}{(q+1)+(3-q)\beta} - \frac{\delta}{\beta} (< \min\{1-\theta, \theta\})$ yields the desired results for $1 \leq q \leq 2$.

For the case $q > 2$, the loss function does not satisfy property (4.3). We apply (4.1) in Theorem 3 with $C_\alpha(\lambda) = O(\lambda^{-1})$ and $\nu_\alpha(\lambda) = O(\lambda^{-(q-1)})$ and

$\mu(\lambda) = O(\lambda^{-(q-2)})$. Choosing $\lambda = T^{-\gamma}$ with $0 < \gamma < \min\{\frac{1-\theta}{q-1}, \frac{\theta}{q+1}\}$ and noting the fact (4.5), from (4.1) we have that $\mathbb{E}\big[\|f_{T+1} - f_\lambda\|_K\big] = O\big(T^{\frac{(q+1)\gamma-\theta}{2}}\big)$. Also, $\|f_\lambda\|_K^2 \leq 2\mathcal{D}(\lambda)/\lambda = O\big(\lambda^{-(1-\beta)}\big)$ and $\|f_{T+1}\|_K \leq C_1(\lambda) = O\big(\lambda^{-1}\big)$, from (4.7) we know that $\mathbb{E}\big[|\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\lambda)|\big] = O(\mathbb{E}\big[\lambda^{-(1-\beta)(q-1)/2}\|f_{T+1} - f_\lambda\|_K\big] + \mathbb{E}\big[\lambda^{-(q-1)}\|f_{T+1} - f_\lambda\|_K\big]) = O\big(T^{\frac{((1-q)\beta+2q)\gamma-\theta}{2}} + T^{\frac{(3q-1)\gamma-\theta}{2}}\big) = O\big(T^{\frac{(3q-1)\gamma-\theta}{2}}\big)$. Trading off this quantity and the regularization error $\mathcal{D}(\lambda) = O(T^{-\beta\gamma})$ in (2.11), the choice $0 < \gamma = \frac{\theta}{3q+2\beta-1} = \frac{1}{4q+2\beta-2} - \frac{\delta}{\beta}(< \min\{\frac{1-\theta}{q-1}, \frac{\theta}{q+1}\})$ gives rise to the desired result. $\square$

Acknowledgement. We would like to thank Prof. Charles A. Micchelli for his helpful comments.

References

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

[2] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations,* Cambridge University Press, 1999.

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

[4] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, Convexity, classification, and risk bounds, *Journal of the American Statistical Association*, 2005. To appear.

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

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

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

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

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

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

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

[12] G. Lugosi and N. Vayatis, On the Bayes-risk consistency of regularized boosting methods, *Ann. Stat.* **32** (2004), 30–55.

[13] C. McDiarmid. Concentration, in *Probabilistic Methods for Algorithmic Discrete Mathematics,* Springer-Verlag, Berlin, (1998) 195–248.

[14] S. Smale and Y. Yao, Online learning algorithms, *Found. Comp. Math.*, 2005. To appear.

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

[16] S. Smale and D.X. Zhou, Learning theory estimates via integral operators and their applications, *Constr. Approx.*, 2005. To appear.

[17] I. Steinwart and C. Scovel, Fast rates for support vector machines, *Proceedings of the 18th Conference on Learning Theory,* 2005.

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

[19] G. Wahba, *Spline Models for Observational Data,* SIAM, 1990.

[20] Y.Ying and D. X. Zhou, Online regularized classification algorithms, Submitted, 2005.

[21] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, *Ann. Stat.* **32** (2004), 56–85.

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

Appendix

In this appendix, we give proofs of some necessary observations used before.

Proof of Proposition 1. We prove the results by induction on t. Since $f_1 = 0$, the result is true for $t = 1$. We assume that the bound holds for $t \in \mathbb{N}_T$, and to advance the induction step to $t+1$, we rewrite the iteration (1.6) as follows

$$f_{t+1} = (1-\eta_t\lambda)f_t - \eta_t\big(V_2'(y_t, f_t(x_t)) - V_2'(y_t, 0)\big)K_{x_t} - \eta_t V_2'(y_t, 0)K_{x_t}. \quad (4.8)$$

For $0 \leq \alpha < 1$, we write

$$\|f_{t+1}\|_K \leq (1-\eta_t\lambda)\|f_t\|_K + M_\alpha \eta_t \kappa^{\alpha+1}\|f_t\|_K^\alpha + \eta_t\kappa|V_2'(y_t, 0)|. \quad (4.9)$$

To estimate the second term on the right hand side of (4.9), we discuss the cases $\alpha = 0$ and $0 < \alpha < 1$ respectively.

If $\alpha = 0$, then

$$\begin{aligned}\|f_{t+1}\|_K &\leq (1-\eta_t\lambda)\|f_t\|_K + \eta_t\big(M_0 + \|V_2'(\cdot, 0)\|_\infty\big)\kappa \\ &\leq (1-\eta_t\lambda)\tfrac{2\kappa(M_0+V_0')}{\lambda} + \eta_t\big(M_0 + V_0'\big)\kappa \leq C_0(\lambda) = \tfrac{2\kappa(M_0+V_0')}{\lambda}.\end{aligned}$$

If $0 < \alpha < 1$, we can use Young inequality $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$ for $a, b > 0$ and $1 < p, q < \infty$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Actually, applying the above inequality with $a = \big(\frac{\lambda}{2}\|f_t\|_K\big)^\alpha$, $b = M_\alpha\kappa^{\alpha+1}\big(\frac{2}{\lambda}\big)^\alpha, p = \frac{1}{\alpha}$ and $q = \frac{1}{1-\alpha}$, we get $M_\alpha\kappa^{\alpha+1}\|f_t\|_K^\alpha \leq \frac{\lambda}{2}\|f_t\|_K + (1-\alpha)\Big[M_\alpha\kappa\big(\frac{2\kappa}{\lambda}\big)^\alpha\Big]^{\frac{1}{1-\alpha}}$. Therefore, (4.9) is dominated by

$$\|f_{t+1}\|_K \leq \Big(1 - \frac{\eta_t\lambda}{2}\Big)\|f_t\|_K + \eta_t\bigg((1-\alpha)\Big[M_\alpha\kappa\Big(\frac{2\kappa}{\lambda}\Big)^\alpha\Big]^{\frac{1}{1-\alpha}} + \kappa V_0'\bigg).$$

Since we have assumed $\|f_t\|_K \leq \frac{2\kappa V_0'}{\lambda} + \frac{2(1-\alpha)}{\lambda}\Big[M_\alpha\kappa\Big(\frac{2\kappa}{\lambda}\Big)^\alpha\Big]^{\frac{1}{1-\alpha}} = C_\alpha(\lambda)$ in the induction step t, it follows $\|f_{t+1}\|_K \leq \Big(1 - \frac{\eta_t\lambda}{2}\Big)C_\alpha(\lambda) + \eta_t\lambda C_\alpha(\lambda)/2 = C_\alpha(\lambda)$ which advances the induction step and gives the desired estimate for $0 \leq \alpha < 1$.

The second case $\alpha = 1$ has been implied by [20]. We only sketch its idea. Let $L_t : \mathcal{H}_K \to \mathcal{H}_K$ be an linear operator defined as for any $f \in \mathcal{H}_k$

$$L_t(f) := \frac{V_2'(y_t, f_t(x_t)) - V_2'(y_t, 0)}{f_t(x_t)} f(x_t)K_{x_t} + \lambda f.$$

Thus, (4.8) implies that

$$f_{t+1} = (I - \eta_t L_t)(f_t) - \eta_t V_2'(y_t, 0)K_{x_t} \quad (4.10)$$

where I is the identity operator. Observe V is 1-admissible and $\|f_t\|_\infty \leq \kappa\|f_t\|_K \leq \kappa C_1(\lambda)$ by the induction assumption, we have $\lambda \leq \|L_t\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq$

$\mu_1(\lambda)$. By the assumption $\mu_1(\lambda)\eta_t \leq 1$, we know $I - \eta_t L_t$ is a positive, linear operator which finally implies $\|I - \eta_t L_t\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq (1 - \eta_t \lambda)$. In connection with (4.10), it follows $\|f_{t+1}\|_K \leq (1 - \eta_t\lambda)\|f_t\|_K + \eta_t \kappa V_0' \leq (1 - \eta_t\lambda)\frac{2\kappa V_0'}{\lambda} + \eta_t \kappa V_0' = C_1(\lambda)$. This completes the induction and the proposition. □

In the following, we give the proof of Proposition 2.

Proof of Proposition 2. We first prove property (a). For any $f, g \in \mathcal{H}_K$, we estimate $\mathcal{V}_\lambda(z, (1-\theta)f + \theta g) - \mathcal{V}_\lambda(z, g)$. Since $V(y, \cdot)$ is convex, we know $V(y, (1-\theta)f(x) + \theta g(x)) - V(y, g(x)) \geq V_2'(y, g(x))(1-\theta)(f(x) - g(x))$. Note also $\|(1-\theta)f + \theta g\|_K^2 - \|g\|_K^2 = 2(1-\theta)\langle g, f - g\rangle_K + (\theta - 1)^2\|f - g\|_K^2$. Putting these estimates together, we have

$$\begin{aligned}\mathcal{V}_\lambda(z, (1-\theta)f + \theta g) - \ & \mathcal{V}_\lambda(z, g) \geq (1-\theta)\langle V_2'(y, g(x))K_x, f - g\rangle_K \\ & + (1-\theta)\lambda\langle g, f - g\rangle_K + \lambda(1-\theta)^2\|f - g\|_K^2/2.\end{aligned}$$

Letting $\theta \to 0_+$ completes the desired property (a).

For property (b), when $V(y, \cdot)$ is differentiable, (3.5) implies $\int_Z \langle \partial\mathcal{V}_\lambda(z, f_\lambda), f - f_\lambda\rangle_K d\rho(z) = \langle \int_Z \partial\mathcal{V}_\lambda(z, f_\lambda) d\rho(z), f - f_\lambda\rangle_K = 0$. Applying property (a) with $g = f_\lambda$ proves property (b) for $V(y, \cdot)$ being differentiable. For a non-differentiable and convex loss function $V(y, \cdot)$, it can always be approximated by a sequence of differentiable loss functions such as $V_\delta(y, \cdot) = \int_0^1 V(y, \cdot - \theta\delta) d\theta,\ 0 < \delta \leq 1$. We shall not dwell on this modified technique and refer the reader to Section 4 of [20] for more details. □