Convergence Analysis of Online Algorithms[†]

Yiming Ying

Department of Mathematics, City University of Hong Kong

Kowloon, Hong Kong, CHINA.

mathying@yahoo.com.cn

Abstract

In this paper, we are interested in the analysis of regularized online algorithms associated with reproducing kernel Hilbert spaces. General conditions on the loss function and step sizes are given to ensure convergence. Explicit learning rates are also given for particular step sizes.

Keywords and Phrases: Online learning algorithm, reproducing kernel Hilbert space, regularized sample error, general loss function

AMS Subject Classification Numbers: 68T05, 62J02.

† The author's current address: Department of Computer Science, University College London, Gower Street, London, WC1E, England, UK.

1 Introduction

We study online algorithms associated with a general convex loss function and reproducing kernel Hilbert spaces. To this end, we review necessary background material and established notations for subsequent use. Let (X, d) be a compact metric space and Y a bounded subset of $\mathbb{R}$. We shall learn a function f^* from X to Y from random samples drawn according to a probability measure ρ on the space $Z := X \times Y$. One way to accomplish this is to specify a loss function $V : Z \to \mathbb{R}_+$ and choose the function $f^* := f_\rho^V$ to be a minimizer of *the generalization error* $\mathcal{E}(f) := \int_Z V(y, f(x)) d\rho(x, y)$ that is,

$$f_\rho^V := \arg\min \Big\{ \mathcal{E}(f) : f \text{ is a measurable function from } X \text{ to } Y \Big\}.$$

We can easily identify the target function f_ρ^V. Indeed, let ρ_X as the marginal distribution of ρ on X and $\rho(\cdot|x)$ the conditional distribution for $x \in X$, then we obtain that

$$f_\rho^V(x) = \arg\min_{t \in Y} \int_Y V(y, t) d\rho(y|x), \qquad \text{a.e., } x \in X.$$

Reproducing kernel Hilbert spaces are often used in the design of leaning algorithms. Let $K : X \times X \to \mathbb{R}$ be continuous, symmetric and positive semidefinite, *i.e.*, for any finite set of distinct points $\{x_j : j \in N_\ell\} \subseteq X$, the matrix $(K(x_i, x_j))_{i,j \in N_\ell}$ is positive semidefinite where we set $N_\ell := \{1, \ldots, \ell\}$. Such a kernel is called a *Mercer kernel*. The *Reproducing Kernel Hilbert Space* (RKHS) $\mathcal{H}_K$ associated with the kernel K is defined (see [1]) to be the completion of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in X\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying the requirement that $\langle K_x, K_y \rangle_K = K(x, y)$ for all $x, y \in X$ and

$$\langle K_x, g \rangle_K = g(x), \qquad x \in X, \ g \in \mathcal{H}_K. \tag{1.1}$$

Now, one standard learning algorithm called *offline regularized algorithm* (see [9] e.g.) associated with the Mercer kernel K and a set of random samples $\mathbf{z} := \{z_t = (x_t, y_t) : t \in N_T\} \subseteq Z$ independently drawn according to ρ is given by

$$f_{\mathbf{z},\lambda} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{T} \sum_{t \in N_T} V(y_t, f(x_t)) + \frac{\lambda}{2} \|f\|_K^2 \right\}. \tag{1.2}$$

where $\lambda > 0$ is called the regularization parameter.

The off-line algorithm (1.2) has been extensively studied in the literature. In particular, its error analysis has been well-developed and can be found in [4, 12, 22, 13, 15, 16, 17, 19, 20].

The essential idea of the analysis is to show that $f_{\mathbf{z},\lambda}$ has an asymptotic behavior similar to the *regularization function* $f_\lambda^V \in \mathcal{H}_K$ defined by

$$f_\lambda^V := \arg\inf_{f\in\mathcal{H}_K}\Big\{\mathcal{E}(f)+\frac{\lambda}{2}\|f\|_K^2\Big\}. \tag{1.3}$$

Moreover, if we regard this regularization function $f_\lambda^V \in \mathcal{H}_K$ as a good learner for the target function f_ρ^V, we can then use the classical gradient descent method [3] to learn it step by step. To explain this fact, we introduce the *regularized loss function* $\mathcal{Q}$ defined for $f \in \mathcal{H}_K$ and $z = (x, y) \in Z$ as

$$\Theta(f,z) = \Theta_\lambda(f,z) := V(y, f(x)) + \frac{\lambda}{2}\|f\|_K^2$$

and the *regularized generalization error* $\mathcal{Q}$ defined for $f \in \mathcal{H}_K$

$$\mathcal{Q}(f) = \mathcal{Q}_\lambda(f) := \int_Z \Theta(f,z)d\rho(z) = \mathcal{E}(f) + \frac{\lambda}{2}\|f\|_K^2.$$

For the purpose of presenting our algorithm for minimizing $\mathcal{Q}$ in $\mathcal{H}_K$, we define the function $\partial\Theta$ at $f \in \mathcal{H}_K$ and $z = (x, y) \in Z$ by the formula

$$\partial\Theta(f,z) := V_2'(y, f(x))K_x + \lambda f$$

where $V_2'(y, s)$ means the derivative of V at the point (y, s) with respect to the second variable. The Hilbert space valued random variable $\partial\Theta(f, z)$ plays the role of the gradient of the functional Θ defined above.

The classical gradient descent tells us that the following sequence $\big\{g_t : g_t \in \mathcal{H}_K, t \in N_{T+1}\big\}$ provides an approximation to f_λ^V

$$\begin{cases} g_1 = 0, \\ \text{and for } t \in N_T \\ g_{t+1} = g_t - \eta_t \displaystyle\int_Z \partial\Theta(g_t, z)d\rho(z) = g_t - \eta_t\Big(\int_Z V_2'(y, g_t(x))K_x d\rho + \lambda g_t\Big). \end{cases} \tag{1.4}$$

Unfortunately, the use of this algorithm requires a knowledge of the distribution ρ. However, in practice it is unknown as we only have the random sample $\mathbf{z}$. Hence, we are led to replace the integral above by the random value $V_2'(y_t, f(x_t))K_{x_t}$. This gives us the so-called *Stochastic Gradient Descent* (SGD) online algorithm [5, 10, 14] defined by

$$\begin{cases} f_1 = 0, \\ \text{and for } t \in N_T \\ f_{t+1} = f_t - \eta_t\big(V_2'(y_t, f_t(x_t))K_{x_t} + \lambda f_t\big). \end{cases} \tag{1.5}$$

With this iterative method, we use f_{T+1} to learn f_λ^V and hence we can also learn f_ρ^V by our remarks above. For each $t \in N_{T+1}$, the function f_t is in general dependent on the inputs $\{z_j : j \in N_{t-1}\}$. The algorithm (1.5) produces the *learning sequence* $\{f_t : t \in N_{T+1}\}$ while the offline learning (1.2) uses all the data immediately.

In this paper, we are mainly interested in the expectation over the random samples of *regularized sample error*

$$\|f_{T+1} - f_\lambda^V\|_K \tag{1.6}$$

and we shall provide effective and useful upper bounds for this quantity. We turn our attention to the description of our results.

2 Reducing the online scheme and Main results

Our error analysis of the online scheme (1.5) assumes a regularity condition on the loss function.

Definition 1. *We say that the loss function is admissible if $V : Y \times \mathbb{R} \to \mathbb{R}_+$ is convex, differentiable with respect to the second variable, $|V|_0 := \sup_{y\in Y} |V(y,0)| < \infty$ and $V_2'(y,s)$ locally Lipschitz continuous i.e. for any $R > 0$, there exists $c > 0$ such that*

$$|V_2'(y,s_1) - V_2'(y,s_2)| \le c|s_1 - s_2| \quad \textit{for all } s_1, s_2 \in [-R,R] \textit{ and } \ y \in Y. \tag{2.1}$$

For simplicity, we denote the right derivative of $V_2'(y,s)$ at the point (y,s) with respect to the second variable as $V_+''(y,s)$. When V is admissible, it is easy to see that $V(y,s)$ is absolutely continuous as a function of s for every $y \in Y$ and $V_+''(y,s)$ exists almost everywhere. Therefore, there holds the equation

$$V_2'(y,s_1) - V_2'(y,s_2) = \int_{s_2}^{s_1} V_+''(y,s)ds, \quad \forall\ s_1, s_2 \in \mathbb{R}.$$

We list below several useful examples of admissible loss functions.

(1) SVM q-norm ($2 \le q < \infty$) soft margin classifier with $V(y,t) = (1-yt)_+^q$, see [18, 22, 4];

(2) least square loss $V(y,t) = (y-t)^2/2$, see e.g. [7, 9, 14, 22];

(3) the exponential loss $V(y,t) = e^{-yt}$, see [22, 11];

(4) the logistic regression $V(y,t) = \log(1+e^{-yt})$, see [22].

Now, we turn our attention to the estimation of the expectation over the samples of the norm $\|f_{T+1} - f_\lambda^V\|_K$. To this end, we set $\mathcal{R}_t := f_t - f_\lambda^V$ and use the definition of the SGD online algorithm (1.5), we obtain that

$$\begin{aligned}\mathcal{R}_{t+1} &= \mathcal{R}_t - \eta_t\big(V_2'(y_t, f_t(x_t))K_{x_t} + \lambda f_t\big)\\ &= \mathcal{R}_t - \eta_t\Big\{[V_2'(y_t, f_t(x_t)) - V_2'(y_t, f_\lambda^V(x_t))]K_{x_t} + \lambda(f_t - f_\lambda^V)\Big\} - \eta_t\partial\Theta(f_\lambda^V, z_t)\\ &= (I - \eta_t\mathcal{A}_t)\mathcal{R}_t - \eta_t\partial\Theta(f_\lambda^V, z_t) \qquad (2.2)\end{aligned}$$

where the linear operator $\mathcal{A}_t : \mathcal{H}_K \to \mathcal{H}_K$ is defined for any $g \in \mathcal{H}_k$ as

$$\mathcal{A}_t(g) := \int_0^1 V_+''(y_t, \theta f_t(x_t) + (1-\theta)f_\lambda^V(x_t))d\theta g(x_t)K_{x_t} + \lambda g$$

and $I : \mathcal{H}_K \to \mathcal{H}_K$ is the notation we use for the identity operator.

We shall use this formula to establish the convergence of the SGD online algorithm (1.5). Before we do this, we introduce the *regularization error* $\mathcal{D}$ defined for every $\lambda > 0$ by

$$\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda^V) - \mathcal{E}(f_\rho^V) + \frac{\lambda}{2}\|f_\lambda^V\|_K^2. \qquad (2.3)$$

Lemma 1. *If V is admissible then f_λ^V is the regularization function if and only if*

$$\int_Z \partial\Theta(f_\lambda^V, z)d\rho = \int_Z V_2'(y, f_\lambda^V(x))K_x d\rho + \lambda f_\lambda^V = 0 \qquad (2.4)$$

and the norm of f_λ^V satisfies

$$\|f_\lambda^V\|_K \le \sqrt{2\mathcal{D}(\lambda)/\lambda}. \qquad (2.5)$$

Proof. The bound for f_λ^V can be easily derived from the inequality

$$\mathcal{D}(\lambda) \ge \lambda/2\|f_\lambda^V\|_K^2.$$

For the proof of (2.4), we observe that the functional $\mathcal{Q} : \mathcal{H}_K \to \mathcal{H}_K$ is differentiable and strictly convex. Therefore, it has a unique minimizer which we have called f_λ^V. Moreover, f_λ^V is determined by the fact that the gradient of $\mathcal{Q}$ at f_λ^V is zero. Indeed, it can be verified for any $f, g \in \mathcal{H}_K$ that

$$\lim_{h\to 0}\frac{\mathcal{Q}(f+hg) - \mathcal{Q}(f)}{h} = \langle\int_Z \partial\Theta(f, z)d\rho(z), g\rangle_K.$$

This proves the lemma. □

Before we are in a position to state our main results, we require additional notation. Since $\mathbb{E}\big(\partial\Theta(f_\lambda^V)\big) = 0$ by Lemma 1, the variance is given by $\sigma^2 := \mathbb{E}\big(\|\partial\Theta(f_\lambda^V)\|_K^2\big)$. We also need the quantities $\kappa := \sup_{x\in X}\sqrt{K(x,x)}$, $|V'|_0 := \sup_{y\in Y}|V_2'(y,0)|$,

$$M(\lambda) := \sup\Big\{V_+''(y,s) : y \in Y,\ |s| \le \kappa\max\big\{\sqrt{2\mathcal{D}(\lambda)/\lambda}, \kappa|V'|_0/\lambda\big\}\Big\} \tag{2.6}$$

and

$$\mu_0(\lambda) := \kappa^2 M(\lambda) + \lambda. \tag{2.7}$$

We can easily compute these constants for all admissible loss functions mentioned above.

Theorem 1. *If V is admissible, $\lambda > 0$, $0 < \eta_t \le \frac{1}{\mu_0(\lambda)}$ and*

$$\sum_{t=1}^{\infty}\eta_t = \infty, \qquad \sum_{t=1}^{\infty}\eta_t^2 < \infty \tag{2.8}$$

then

$$\lim_{T\to\infty}\mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_K\big) = 0.$$

For the step size $\eta_t = O(t^{-\theta}), t \to \infty$ we get the following convergence rates.

Theorem 2. *If V is admissible, $\lambda > 0$, $\mu(\lambda) \ge \mu_0(\lambda)$ and $\eta_t = \frac{1}{\mu(\lambda)t^\theta}$ with $1/2 < \theta < 1$ then for $T \ge 4$, $\mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_K\big)$ is bounded by*

$$\frac{8\sigma}{\lambda}\Big(\frac{2}{2\theta-1}\Big)^{1/2}T^{1/2-\theta} + \Big[\big(6C_\theta(T) + \sqrt{2}\big)\frac{\sigma}{\mu(\lambda)}\Big(\frac{1}{2\theta-1}\Big)^{1/2} + \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}\Big)\Big]e^{-D_{\theta,\lambda}T^{1-\theta}} \tag{2.9}$$

where

$$C_\theta(T) = \begin{cases} \frac{2}{4\theta-3} & \text{for } \theta \in (3/4, 1) \\ \ln(\frac{T}{2}) & \text{for } \theta = 3/4 \\ \frac{2}{3-4\theta}T^{3/2-2\theta} & \text{for } \theta \in (1/2, 3/4) \end{cases}$$

and

$$D_{\theta,\lambda} := \frac{\lambda}{(1-\theta)\mu(\lambda)}\big(1 - (1/2)^{1-\theta}\big).$$

Although Theorem 1 allows us to choose $\eta_t = O(t^{-\theta})$ with $\theta = 1$, the resulting convergence rate is unacceptably slow as we present in the following theorem.

Theorem 3. *Under the hypotheses above and $\theta = 1$, then for $T \ge 4$, $\mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_K\big)$ is bounded by*

$$2\Big[\frac{2\sigma}{\mu(\lambda)}\Big(1 + C'(T)\Big) + \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}\Big]\Big(\frac{1}{T}\Big)^{\frac{\lambda}{\mu(\lambda)}} + \frac{2\sigma}{\mu(\lambda)}\frac{1}{\sqrt{T}}.$$

where

$$C'(T) := \begin{cases} \ln(T+1), & \frac{\lambda}{\mu(\lambda)} = 1/2 \\ \frac{2\mu(\lambda)}{\mu(\lambda)-\lambda}\Big[1-(T+1)^{-1/2+\lambda/\mu(\lambda)}\Big], & \frac{\lambda}{\mu(\lambda)} \in (0,1/2)\cup(1/2,1). \end{cases}$$

We prove Theorem 1 in Section 3 and the proofs of Theorem 2 and 3 will be given in Section 4.

The above theorems give us the convergence rate of the regularized sample error $\|f_{T+1} - f_\lambda^V\|_K$ under the specific choices of step sizes. In order to get the whole leaning rate for the *excess generalization error*

$$\mathcal{E}(f_{T+1}) - \mathcal{E}(f_\rho^V), \tag{2.10}$$

we should bias the regularized sample error and the regularization (approxiamtion) error between f_λ^V and f_ρ^V. In the following, let us illustrate this approach for least square loss $V(y,s) = (y-s)^2/2$. In this case, the target function f_ρ^V shall be denoted by f_ρ and is given for $x \in X$ by $f_\rho(x) := \int_Y y d\rho(y|x)$. Often, f_ρ is referred to as *the regression function*.

Consider the least square regression problem with $Y = [-M, M]$ for some $M > 0$. The approximation error between f_λ^V and f_ρ is measured by $\|f_\lambda^V - f_\rho\|_{L^2_{\rho_X}}$ (see [15, 17]). It depends on the approximating property of $\mathcal{H}_K$ which can be characterized by the integral operator $L_K : L^2_{\rho_X}(X) \to L^2_{\rho_X}(X)$ defined for any $f \in L^2_{\rho_X}(X)$ and $x \in X$ as

$$L_K(f)(x) = \int_X K(x,x')f(x')d\rho_X(x').$$

Since K is a Mercer kernel, the operator L_K is positive, compact and symmetric. Therefore the fractional power of the operator denoted by L_K^β is well-defined for any $0 < \beta \le 1$.

Denote $C(X)$ as the space of continuous functions on X with the norm $\|\cdot\|_\infty$. Then the reproducing property (1.1) tells us the following useful inequality

$$\|f\|_\infty \le \kappa \|f\|_K. \tag{2.11}$$

Corollary 1. *If V is the least square loss, f_ρ is in the range of L_K^β for some $0 < \beta \le 1$, $0 < \varepsilon < \frac{\beta}{2\beta+4}$, $\lambda = T^{-\frac{1}{2(\beta+2)}+\frac{\varepsilon}{\beta}}$ and $\eta_t = \frac{1}{\kappa^2+\lambda} t^{\frac{(\beta+1)\varepsilon}{\beta} - \frac{2\beta+3}{2\beta+4}}$ then there exists a constant $c > 0$ such that for all T*

$$\mathbb{E}\Big(\|f_{T+1} - f_\rho\|_{L^2_{\rho_X}}\Big) \le c\Big(\frac{1}{T}\Big)^{\frac{\beta}{2(\beta+2)}-\varepsilon}. \tag{2.12}$$

This corollary follows from Theorem 2. We provide its proof here.

Proof. In order to apply the result of Theorem 2, we should first estimate the constants appearing in the righthand side of the inequality (2.9). We first estimate $\mathcal{D}(\lambda)$ and σ.

Since $Y = [-M, M]$, then $|f_\rho(x)| \leq M$ and

$$\mathcal{D}(\lambda) = \mathcal{E}(f_\lambda^V) - \mathcal{E}(f_\rho) + \frac{\lambda}{2}\|f_\lambda^V\|_K^2 \leq \mathcal{E}(0) \leq \frac{M^2}{2}.$$

Note that $\mathcal{E}(f_\lambda^V) \leq \mathcal{D}(\lambda) + \mathcal{E}(f_\rho), V_2'(y,s)^2 = 2V(y,s)$ and $\|f_\lambda^V\|_K \leq \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}$, then the variance can be estimated as follows

$$\begin{aligned}
\sigma^2 = \mathbb{E}_z(\|\partial\Theta(f_\lambda^V)\|_K^2) &\leq 2\kappa^2 \int_Z V'(y, f_\lambda^V(x))^2 d\rho + 2\lambda^2\|f_\lambda^V\|_K^2 \\
&\leq 4\kappa^2 \int_Z V(y, f_\lambda^V(x)) d\rho + 2\lambda^2\|f_\lambda^V\|_K^2 \\
&\leq 4\kappa^2\big(\mathcal{D}(\lambda) + \mathcal{E}(f_\rho)\big) + 4\lambda\mathcal{D}(\lambda) \\
&\leq 4(\kappa^2 + \lambda)\big(\mathcal{D}(\lambda) + \mathcal{E}(f_\rho)\big) \leq 10M^2(\kappa^2 + \lambda).
\end{aligned}$$

Since V is the least square loss, we can choose $\mu(\lambda) = \kappa^2 + \lambda$. Thus, there exists a constant c_θ such that for all $0 < \lambda < 1$, we have that $D_{\theta,\lambda} \geq c_\theta\lambda$ and $C_\theta(T) \leq T^{3/2}$. Now applying (2.9) of Theorem 2 with $\eta_t = \frac{1}{\mu(\lambda)}t^{-\theta}$ for $\theta \in (1/2, 1)$, we have that there exists a constant c_1 such that for all $0 < \lambda < 1$ and $T \geq 4$

$$\mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_K\big) \leq c_1\Big(\lambda^{-1}T^{1/2-\theta} + T^{3/2}e^{-c_\theta\lambda T^{1-\theta}}\Big).$$

Since f_ρ is in the range of L_K^β, by [17] there exists a positive constant c_β such that for all $\lambda > 0$

$$\|f_\lambda^V - f_\rho\|_{L^2_{\rho_X}} \leq c_\beta\lambda^\beta.$$

Therefore, for any $1/2 < \theta < 1$, there exists a constant c_2 such that

$$\begin{aligned}
\mathbb{E}\big(\|f_{T+1} - f_\rho\|_{L^2_{\rho_X}}\big) &\leq \mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_{L^2_{\rho_X}} + \|f_{T+1} - f_\rho\|_{L^2_{\rho_X}}\big) \\
&\leq \mathbb{E}\big(\kappa(\|f_{T+1} - f_\lambda^V\|_K + \|f_\lambda^V - f_\rho\|_{L^2_{\rho_X}}\big) \\
&\leq c_2\Big(\lambda^{-1}T^{1/2-\theta} + T^{3/2}e^{-c_\theta\lambda T^{1-\theta}} + \lambda^\beta\Big).
\end{aligned}$$

For any $0 < \varepsilon < \frac{\beta}{2\beta+4}$, we select $\theta = \frac{2\beta+3}{2\beta+4} - \frac{(\beta+1)}{\beta}\varepsilon$ and $\lambda = T^{-\frac{1}{2(\beta+2)}+\frac{\varepsilon}{\beta}}$, and conclude that there exists a positive constant c such that

$$\mathbb{E}\Big(\|f_{T+1} - f_\rho\|_{L^2_{\rho_X}}\Big) \leq c\Big(\frac{1}{T}\Big)^{\frac{\beta}{2(\beta+2)}-\varepsilon}.$$

This completes the proof. □

In [14], the authors consider the following stochastic gradient method in Hilbert space $\mathcal{H}_K$. Let the map $A : Z \to SL(\mathcal{H}_K)$ be the vector space of positive definite symmetric linear operators and $B : Z \to \mathcal{H}_K$. They proposed the learning sequence

$$f_{t+1} = f_t - \eta_t\big(A_t(f_t) + B_t\big) \tag{2.13}$$

to learn a stationary point f_* satisfying

$$\mathbb{E}_z(A(z) + B(z))f_* = 0$$

where for each t, $A_t = A(\mathbf{z}_t)$ and $B_t = B(\mathbf{z}_t)$. We also denote $\mathcal{R}_t := f_t - f_*$ and rewrite the above equation

$$\mathcal{R}_{t+1} = (I - \eta_t A_t)\mathcal{R}_t - \eta_t\big(A_t(f_*) + B_t\big).$$

Comparing this equation with Lemma 1, we see that the last term plays the role of $\partial\Theta(f_\lambda^V, z_t)$ in (2.2). Since f_t depends on $\{z_\ell : \ell \in N_{t-1}\}$, the operator $\mathcal{A}_t$ depends on the samples $\{z_\ell : \ell \in N_t\}$. Hence, it cannot in general be written as A_t except for the least square loss function.

In [10], the authors also considered the general regularized online scheme (1.5). When the loss function $V(y, s)$ is convex and uniformly Lipschitz continuous with respect to $s \in \mathbb{R}$ for every $y \in Y$, the step sizes is chosen to be $\eta_t = O(t^{-1/2})$ and $\lambda > 0$, then the authors proved that the average instantaneous risk, $1/T \sum_{t=1}^{T} \Theta(f_t, z_t)$ converges to the regularized generalization error $\mathcal{Q}(f_\lambda^V)$ with rate $O(T^{-1/2})$ as $T \to \infty$.

For classifying loss functions, convergence rates are recently given in [21]. However, the methods used here are quite different than those presented there.

3 General Convergence Results

In this section, we shall prove Theorem 1. We begin with a bound for the learning sequence $\big\{f_t : t \in N_{T+1}\big\}$. In order to do so, we define the quantity

$$\tilde{\mu}_0(\lambda) := \kappa^2 \sup\Big\{V_+''(y, s) : y \in Y,\ |s| \le \kappa^2 |V'|_0/\lambda\Big\}\Big\} + \lambda$$

which is smaller than $\mu(\lambda)$ defined before.

Lemma 2. *If V is admissible and the step size satisfy $\eta_t \cdot \tilde{\mu}_0(\lambda) \leq 1$, $t \in N_{T+1}$ then*

$$\|f_t\|_K \leq \kappa |V'|_0/\lambda.$$

Proof. We prove this inequality by induction on t. Since $f_1 = 0$, the result is true for $t = 1$. We assume that the bound holds for t, and to advance the induction step to $t+1$, we rewrite the iteration (1.5) as follows

$$f_{t+1} = f_t - \eta_t \int_0^1 V_+''(y_t, \theta f_t(x_t)) d\theta f_t(x_t) K_{x_t} - \eta_t \lambda f_t - \eta_t V_2'(y_t, 0) K_{x_t}.$$

We introduce the linear operator $\tilde{\mathcal{A}}_t : \mathcal{H}_K \to \mathcal{H}_K$ defined for any $g \in \mathcal{H}_K$ as

$$\tilde{\mathcal{A}}_t(g) := \int_0^1 V_+''(y_t, \theta f_t(x_t)) d\theta g(x_t) K_{x_t} + \lambda g \tag{3.1}$$

so that

$$f_{t+1} = (I - \eta_t \tilde{\mathcal{A}}_t) f_t - \eta_t V_2'(y_t, 0) K_{x_t}. \tag{3.2}$$

Since V is admissible, we obtain for any $g \in \mathcal{H}_K$ that

$$< \tilde{\mathcal{A}}_t(g), g >_K \geq \lambda \|g\|_K^2.$$

By the bound for f_λ^V in Lemma 1 and the induction hypothesis, we conclude that the operator norm of $\tilde{A}_t$ satisfies the inequality

$$\|\tilde{\mathcal{A}}_t\| \leq \tilde{\mu}_0(\lambda).$$

Consequently, $(I - \eta_t \tilde{\mathcal{A}}_t)$ is positive and self-joint.

To estimate (3.2), we first show for any $g \in \mathcal{H}_K$ that

$$\|(I - \eta_t \tilde{\mathcal{A}}_t)(g)\|_K \leq (1 - \eta_t \lambda) \|g\|_K. \tag{3.3}$$

Indeed, we define the operator

$$\tilde{\mathcal{B}}_t = \tilde{\mathcal{A}}_t - \lambda I$$

and observe that it is positive and its norm is bounded by $\kappa^2 \sup \Big\{ V_+''(y, s) : y \in Y,\ |s| \leq \kappa^2 |V'|_0/\lambda \Big\}$ by the induction hypothesis. Next, we rewrite the quantity $\|(I - \eta_t \tilde{\mathcal{A}}_t)(g)\|_K^2$ as the expression

$$(1 - \eta_t \lambda)^2 \|g\|_K^2 - 2\eta_t \langle \tilde{\mathcal{B}}_t(g), g \rangle_K + 2\eta_t^2 \lambda \langle \tilde{\mathcal{B}}_t(g), g \rangle_K + \eta_t^2 \|\tilde{\mathcal{B}}_t(g)\|_K^2.$$

Note that

$$\|\tilde{\mathcal{B}}_t(g)\|_K^2 = \langle \tilde{\mathcal{B}}_t(g), g \rangle_K \times \int_0^1 V_+''(y_t, \theta f_t(x_t)) d\theta K(x_t, x_t) \leq M(\lambda) \kappa^2 \langle \tilde{\mathcal{B}}_t(g), g \rangle_K$$

which leads us to the inequality

$$\|(I-\eta_t\tilde{\mathcal{A}}_t)(g)\|_K^2 \leq (1-\eta_t\lambda)^2\|g\|_K^2 - 2\eta_t\Big[1-\eta_t\tilde{\mu}(\lambda)\Big]\langle\tilde{\mathcal{B}}_t(g),g\rangle_K. \tag{3.4}$$

Since $\langle\tilde{\mathcal{B}}_t(g),g\rangle_K \geq 0$ and our hypothesis on the step size, we obtain (3.3).

The expression (3.2) together with the estimate (3.3) yields the bound

$$\begin{aligned}\|f_{t+1}\|_K &\leq (1-\eta_t\lambda)\|f_t\|_K + \kappa\eta_t|V'|_0\\ &\leq (1-\eta_t\lambda)\kappa|V'|_0/\lambda + \kappa\eta_t|V'|_0\\ &= \kappa|V'|_0/\lambda\end{aligned}$$

which advances the induction step and proves the lemma . □

Lemma 3 below is proved in a similar fashion.

Lemma 3. *If V is admissible and the step size satisfies $\eta_t\cdot\mu_0(\lambda)\leq 1$ then*

$$\|\mathcal{A}_t\| \leq \mu_0(\lambda), \quad \|(I-\eta_t\mathcal{A}_t)\| \leq (1-\eta_t\lambda).$$

For the next lemma, we define $\mathcal{A}_k^T = \prod_{t=k}^{T}(I-\eta_t\mathcal{A}_t)$ for $k\in N_T$ and $\mathcal{S}_k = \sum_{j=1}^{k}\eta_j\partial\Theta(f_\lambda^V,z_j)$ for all $k\in N$. Also, we set $\mathcal{A}_{T+1}^T = I$ and $\mathcal{S}_0 = 0$.

Lemma 4.

$$\mathcal{R}_{T+1} = \mathcal{A}_1^T\mathcal{R}_1 - \sum_{k=1}^{T-1}\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)\mathcal{S}_k - \mathcal{S}_T \tag{3.5}$$

Proof. Recall that

$$\mathcal{R}_{t+1} = (I-\eta_t\mathcal{A}_t)\mathcal{R}_t - \eta_t\partial\Theta(f_\lambda^V,z_t)$$

and therefore by induction on t that

$$\mathcal{R}_{T+1} = \mathcal{A}_1^T\mathcal{R}_1 - \sum_{k=1}^{T}\eta_k\mathcal{A}_{k+1}^T\partial\Theta(f_\lambda^V,z_k). \tag{3.6}$$

Since for $k\in N$, $\mathcal{S}_k - \mathcal{S}_{k-1} = \eta_k\partial\Theta(f_\lambda^V,z_k)$, we can rewrite the second term of the above equality as

$$\sum_{k=1}^{T}\mathcal{A}_{k+1}^T(\mathcal{S}_k-\mathcal{S}_{k-1}) = \sum_{k=1}^{T-1}(\mathcal{A}_{k+1}^T-\mathcal{A}_{k+2}^T)\mathcal{S}_k + \mathcal{S}_T.$$

This proves the lemma. □

Before we turn to the proof of Theorem 1, we also need the following intermediate lemma.

Lemma 5. *If* $\sum_{t=1}^{\infty} \eta_t^2 < \infty$ *then there exists a* $L^2_\rho(\mathcal{H}_K)$*-valued random variable* $\mathcal{S}_*$ *such that*

$$\lim_{T\to\infty} \mathbb{E}\big(\|\mathcal{S}_T - \mathcal{S}_*\|_K^2\big) = 0.$$

Proof. Recall that Lemma 1 tells us that $\mathbb{E}\big(\partial\Theta(f_\lambda^V)\big) = 0$. Since the samples are selected i.i.d., we have for any $T, T' \in N$ with $T' > T$ that

$$\mathbb{E}\big(\|\mathcal{S}_T - \mathcal{S}_{T'}\|_K^2\big) = \sum_{j,k=T+1}^{T'} \eta_j\eta_k \mathbb{E}\langle\partial\Theta(f_\lambda^V, z_j), \partial\Theta(f_\lambda^V, z_k)\rangle_K \quad = \Big(\sum_{j=T+1}^{T'} \eta_j^2\Big)\mathbb{E}(\|\partial\Theta(f_\lambda^V)\|_K^2).$$

Since the step sizes satisfy $\sum_{j=1}^{\infty} \eta_j^2 < \infty$, this means $\mathcal{S}_T$ is a Cauchy sequence in the Hilbert-valued space $L^2_\rho(\mathcal{H}_K)$ which gives us the lemma. □

With the above preparations, we can give the proof of Theorem 1.

Proof of Theorem 1. By Lemma 4 and Lemma 5, we get that

$$\begin{aligned}\mathcal{R}_{T+1} =& \mathcal{A}_1^T\mathcal{R}_1 - \sum_{k=1}^{T-1}\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*) - (\mathcal{S}_T - \mathcal{S}_*) - \sum_{k=1}^{T-1}\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)\mathcal{S}_* - \mathcal{S}_* \\ =& \mathcal{A}_1^T\mathcal{R}_1 - \sum_{k=1}^{T-1}\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*) - (\mathcal{S}_T - \mathcal{S}_*) - \mathcal{A}_2^T\mathcal{S}_*. \qquad (3.7)\end{aligned}$$

Using Lemma 3, we can bound the expectation of the $\mathcal{H}_K$ norm of the first term in (3.7) by

$$\mathbb{E}\big(\|\mathcal{A}_1^T\mathcal{R}_1\|_K\big) \le \prod_{j=1}^{T}(1-\eta_j\lambda)\|f_\lambda\|_K \le \exp\Big\{-\lambda\sum_{j=1}^{T}\eta_j\Big\}\|f_\lambda\|_K \qquad (3.8)$$

where we have used the fact $1 - x \le e^{-x}$ for all $x > 0$ in the last inequality. Thus the assumption on the step size $\sum_{j=1}^{\infty}\eta_j = \infty$ tells us that the upper bound in (3.8) tends to zero when $T \to \infty$.

The expectation of the norm of the fourth term in (3.7) can be estimated as above, namely

$$\mathbb{E}\big(\|\mathcal{A}_2^T\mathcal{S}_*\|_K\big) \le \exp\big\{-\lambda\sum_{j=2}^{T}\eta_j\big\}\mathbb{E}(\|\mathcal{S}_*\|_K) \qquad (3.9)$$

which goes to zero as $T \to \infty$.

By Lemma 5, the expectation o f the norm of the third term $\mathbb{E}(\|\mathcal{S}_T - \mathcal{S}_*\|_K)$ tends to zero as $T \to \infty$.

Hence, it remains to estimate the second term. By Lemma 5, we know that for any $\varepsilon > 0$, there exists a positive integer T_1 such that, for all $k \geq T_1$ there holds

$$\mathbb{E}\big(\|\mathcal{S}_k - \mathcal{S}_*\|_K\big) \leq \varepsilon. \tag{3.10}$$

Hence, we can decompose the expectation of the norm of the second term in (3.7) into two parts

$$\begin{aligned}\mathbb{E}\big(\|\sum_{k=1}^{T-1}\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*)\|_K\big) \leq \sum_{k=1}^{T_1}\mathbb{E}(\|\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*)\|) \\ + \sum_{k=T_1+1}^{T-1}\mathbb{E}(\|\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*)\|).\end{aligned} \tag{3.11}$$

Using Lemma 3, we know that

$$\begin{aligned}\|\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*)\|_K &= \|\eta_{k+1}\mathcal{A}_{k+1}\mathcal{A}_{k+2}^T(\mathcal{S}_k - \mathcal{S}_*)\|_K \\ &\leq \eta_{k+1}\mu_0(\lambda)\prod_{j=k+2}^{T}(1-\eta_j\lambda)\|\mathcal{S}_k - \mathcal{S}_*\|_K.\end{aligned} \tag{3.12}$$

It implies that the first part of the righthand side of (3.11) is bounded by

$$\mu_0(\lambda)\exp\big\{-\lambda\sum_{j=T_1+2}^{T}\eta_j\big\}\sum_{k=1}^{T_1}\eta_{k+1}\mathbb{E}\big(\|\mathcal{S}_k - \mathcal{S}_*\|_K\big) \tag{3.13}$$

which tends to zero as $T \to \infty$.

We treat the second part in the upper bound of (3.11) by the observation that

$$\sum_{k=T_1+1}^{T-1}\mathbb{E}(\|\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*)\|) \leq \varepsilon\mu_0(\lambda)\big[\sum_{k=T_1+1}^{T-2}\prod_{j=k+2}^{T}(1-\eta_j\lambda)\eta_{k+1} + \eta_T\big].$$

Hence, the upper bound in (3.11) will tend to zero as $T \to \infty$ if we prove that

$$\sup_T\sum_{k=1}^{T-2}\prod_{j=k+2}^{T}(1-\eta_j\lambda)\eta_{k+1} < \infty. \tag{3.14}$$

In order to do so, we first note that $\lambda\eta_j \leq 1$ and $\eta_{k+1}\lambda = 1 - (1-\eta_{k+1}\lambda)$. Therefore, (3.14)

is dominated by

$$\frac{1}{\lambda}\sum_{k=1}^{T-2}\prod_{j=k+2}^{T}(1-\eta_j\lambda)\eta_{k+1}\lambda = \frac{1}{\lambda}\sum_{k=1}^{T-2}\prod_{j=k+2}^{T}(1-\eta_j\lambda)\big[1-(1-\eta_{k+1}\lambda)\big]$$
$$= \frac{1}{\lambda}\sum_{k=1}^{T-2}\Big[\prod_{j=k+2}^{T}(1-\eta_j\lambda) - \prod_{j=k+1}^{T}(1-\eta_j\lambda)\Big]$$
$$= \frac{1}{\lambda}\big[(1-\lambda\eta_T) - \prod_{j=2}^{T}(1-\eta_j\lambda)\big] \leq \frac{1}{\lambda}.$$

This completes the proof. □

4 Convergence Rate Analysis

In this section, we give the explicit convergence rate of $\|f_{T+1} - f_\lambda^V\|_K$ for specific step sizes $\eta_t = O(t^{-\theta})$. Let us first give the proof of Theorem 2. Specially, we prove the following facts.

Lemma 6. *If V is admissible, $\eta_t = \frac{1}{\mu(\lambda)t^\theta}$ with $1/2 < \theta \leq 1$ and $\mathcal{S}_*$ as stated in Lemma 5 then there holds*

$$\mathbb{E}\big(\|\mathcal{S}_*\|_K^2\big) \leq \frac{\sigma^2}{\mu^2(\lambda)}\Big(\frac{2\theta}{2\theta-1}\Big) \tag{4.1}$$

and

$$\mathbb{E}\big(\|\mathcal{S}_k - \mathcal{S}_*\|_K^2\big) \leq \frac{\sigma^2}{\mu^2(\lambda)}\Big(\frac{1}{2\theta-1}\Big)k^{1-2\theta}. \tag{4.2}$$

Proof. Since we have for any $l > k$ that

$$\mathbb{E}\big(\|\mathcal{S}_k - \mathcal{S}_l\|_K^2\big) = \sum_{j,j'=k+1}^{l}\eta_j\eta_{j'}\mathbb{E}\langle\partial\Theta(f_\lambda^V, z_j), \partial\Theta(f_\lambda^V, z_j')\rangle$$
$$= \sigma^2\sum_{j=k+1}^{l}\eta_j^2 \leq \frac{\sigma^2}{\mu^2(\lambda)}\big(\frac{1}{2\theta-1}\big)k^{1-2\theta},$$

the inequality (4.2) follows from the observation $\mathbb{E}\big(\|\mathcal{S}_k - \mathcal{S}_*\|_K^2\big) \leq \lim_{l\to+\infty}\mathbb{E}\big(\|\mathcal{S}_k - \mathcal{S}_l\|_K^2\big)$.

Note that

$$\mathbb{E}\big(\|\mathcal{S}_l\|_K^2\big) \leq \frac{\sigma^2}{\mu^2(\lambda)}\sum_{j=1}^{l}j^{-2\theta} \leq \frac{\sigma^2}{\mu^2(\lambda)}(1+\int_1^l x^{-2\theta}dx) \leq \frac{\sigma^2}{\mu^2(\lambda)}(\frac{2\theta}{2\theta-1}).$$

This in connection with $\mathbb{E}\big(\|\mathcal{S}_*\|_K^2\big) \le \lim\limits_{l\to+\infty} \mathbb{E}\big(\|\mathcal{S}_l\|_K^2\big)$ implies the first inequality (4.1). This proves the lemma. □

We shall also use the following estimation. A modified form was given in [14].

Lemma 7. *If* $0<\nu<1$ *then the quantity* $\sum\limits_{k=1}^{T-2}\exp\Big\{-\nu\sum\limits_{j=k+2}^{T} j^{-\theta}\Big\}(k+1)^{-\theta}k^{1/2-\theta}$ *is bounded by*

$$\begin{cases} \frac{6\sqrt{2}}{\nu}T^{-\theta+1/2}+6C_\theta(T)\exp\Big\{-\frac{\nu(1-2^{\theta-1})}{1-\theta}(T+1)^{1-\theta}\Big\}, & \theta\in(1/2,1)\\ 4C_\nu(T)T^{-\nu}, & \theta=1 \end{cases}$$

where

$$C_\theta(T):=\begin{cases} \frac{2}{4\theta-3} & \theta\in(3/4,1)\\ \ln(\frac{T}{2}) & \theta=3/4\\ \frac{2}{3-4\theta}T^{3/2-2\theta} & \theta\in(1/2,3/4) \end{cases}$$

and

$$C_\nu(T):=\int_1^{T+1}x^{-3/2+\nu}dx=\begin{cases}\ln(T+1), & \nu=1/2\\ \frac{2}{1-2\nu}\Big[1-(T+1)^{-1/2+\nu}\Big], & \nu\in(0,1/2)\cup(1/2,1).\end{cases}$$

Proof. Denote

$$I=\sum_{k=1}^{T-2}\exp\Big\{-\nu\sum_{j=k+2}^{T}j^{-\theta}\Big\}(k+1)^{-\theta}k^{1/2-\theta}. \tag{4.3}$$

For any $\theta\in(1/2,1)$, observe that for any $k\ge 0$, there holds

$$\sum_{j=k+2}^{T}j^{-\theta}\ge\int_{k+2}^{T+1}x^{-\theta}dx=\Big(\frac{1}{1-\theta}\Big)\Big[(T+1)^{1-\theta}-(k+2)^{1-\theta}\Big]. \tag{4.4}$$

Since $(k+1)^{-\theta}k^{1/2-\theta}\le\frac{3\sqrt{3}}{2}(k+2)^{1/2-2\theta}$ for any $k\ge1$ and $1/2<\theta\le1$, we have

$$I\le\frac{3\sqrt{3}}{2}\exp\Big\{-\frac{\nu}{1-\theta}(T+1)^{1-\theta}\Big\}\sum_{k=2}^{T}\exp\Big\{\frac{\nu}{1-\theta}k^{1-\theta}\Big\}k^{1/2-2\theta}.$$

For $x\in[k,k+1]$ with $k\ge2$, we have $k^{1/2-2\theta}\le2x^{1/2-2\theta}$ and $\exp\big\{\frac{\nu}{1-\theta}k^{1-\theta}\big\}\le\exp\big\{\frac{\nu}{1-\theta}x^{1-\theta}\big\}$. Then

$$I\le6\exp\Big\{-\frac{\nu}{1-\theta}(T+1)^{1-\theta}\Big\}\int_2^{T+1}f(x)dx \tag{4.5}$$

where we set $f(x):=\exp\Big\{\frac{\nu}{1-\theta}x^{1-\theta}\Big\}x^{1/2-2\theta}$.

To estimate the integral, we decompose it into two parts

$$\int_2^{T+1} f(x)dx = \int_{T/2}^{T+1} f(x)dx + \int_2^{T/2} f(x)dx. \tag{4.6}$$

Since

$$\begin{aligned}\int_{T/2}^{T+1} f(x)dx &\le \sqrt{2}T^{1/2-\theta}\int_{T/2}^{T+1} x^{-\theta}\exp\Big\{\frac{\nu}{1-\theta}x^{1-\theta}\Big\}dx\\ &\le \frac{\sqrt{2}}{\nu}T^{1/2-\theta}\exp\Big\{\frac{\nu}{1-\theta}(T+1)^{1-\theta}\Big\}\end{aligned} \tag{4.7}$$

and

$$\begin{aligned}\int_2^{T/2} f(x)dx &\le \Big(\int_1^{T/2} x^{1/2-2\theta}dx\Big)\exp\Big\{\frac{\nu}{1-\theta}(T/2)^{1-\theta}\Big\}\\ &\le C_\theta(T)\exp\Big\{\frac{\nu}{1-\theta}(T/2)^{1-\theta}\Big\}\end{aligned} \tag{4.8}$$

where

$$C_\theta(T) := \begin{cases}\frac{2}{4\theta-3} & \text{for } \theta\in(3/4,1)\\ \ln(\frac{T}{2}) & \text{for } \theta = 3/4\\ \frac{2}{3-4\theta}T^{3/2-2\theta} & \text{for } \theta\in(1/2,3/4).\end{cases}$$

Putting the estimates (4.5), (4.6), (4.7) and (4.8) together, we get the first assertion of Lemma 7.

For $\theta = 1$ and $\nu\in(0,1)$, we have for any $0\le k\le T-2$ that

$$\sum_{j=k+2}^{T} j^{-1} \ge \int_{k+2}^{T+1} x^{-1}dx = \ln(T+1) - \ln(k+2). \tag{4.9}$$

Hence, we see that

$$I \le \frac{3\sqrt{3}}{2}\Big(T+1\Big)^{-\nu}\sum_{k=1}^{T-2}\big(k+2\big)^{-3/2+\nu} \le \frac{3\sqrt{3}}{2}\Big(T+1\Big)^{-\nu}\int_1^{T+1} x^{-3/2+\nu}dx.$$

The observation that

$$C_\nu(T) := \int_1^{T+1} x^{-3/2+\nu}dx = \begin{cases}\ln(T+1), & \nu = 1/2\\ \frac{2}{1-2\nu}\Big[1-(T+1)^{-1/2+\nu}\Big], & \nu\in(0,1/2)\cup(1/2,1)\end{cases}$$

completes the proof of the lemma. □

To this end, we are in a position to prove Theorem 2.

Proof of Theorem 2. To estimate the explicit rate, we will follow the proof of Theorem 1. Recall the equality (3.7)

$$\mathcal{R}_{T+1} = \mathcal{A}_1^T\mathcal{R}_1 - \sum_{k=1}^{T-1}\big(\mathcal{A}_{k+1}^T - \mathcal{A}_{k+2}^T\big)(\mathcal{S}_k - \mathcal{S}_*) - (\mathcal{S}_T - \mathcal{S}_*) - \mathcal{A}_2^T\mathcal{S}_* := I_1 + I_2 + I_3 + I_4. \quad (4.10)$$

We shall estimate the four terms on the righthand side of (4.10) one by one.

Applying (4.4) with $k = 0$, we know for $T \geq 4$ that

$$\sum_{j=2}^{T} j^{-\theta} \geq \frac{1 - 2^{\theta-1}}{1-\theta} T^{1-\theta}. \quad (4.11)$$

Since $\mathbb{E}\big(\|\mathcal{S}_*\|_K\big) \leq \frac{\sigma}{\mu(\lambda)}\big(\frac{2\theta}{2\theta-1}\big)^{1/2}$ by (4.1) and $\|f_\lambda\|_K \leq \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}$, we know from the estimates (3.8) and (3.9) that

$$\begin{aligned}\mathbb{E}(\|I_1\|_K) + \mathbb{E}(\|I_4\|_K) &\leq \Big(\|f_\lambda\|_K + \mathbb{E}\big(\|\mathcal{S}_*\|_K\big)\Big)\exp\Big\{-\frac{\lambda(1-2^{\theta-1})}{\mu(\lambda)(1-\theta)}T^{1-\theta}\Big\} \\ &\leq \bigg(\frac{\sigma}{\mu(\lambda)}\Big(\frac{2\theta}{2\theta-1}\Big)^{1/2} + \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}\bigg)e^{-D_{\theta,\lambda}T^{1-\theta}} \quad (4.12)\end{aligned}$$

where we set $D_{\theta,\lambda} := \frac{\lambda(1-2^{\theta-1})}{\mu(\lambda)(1-\theta)}$.

The expectation of the norm of the third term in (4.10) is immediately from Lemma 6 by Cauchy-Schwarz inequality

$$\mathbb{E}(\|I_3\|_K) \leq \frac{\sigma}{\mu(\lambda)}\Big(\frac{1}{2\theta-1}\Big)^{1/2}T^{1/2-\theta}. \quad (4.13)$$

We shall use lemma 7 to estimate the expectation of the norm of the third term in (4.10). Indeed, it is bounded by

$$\mathbb{E}\big(\|I_2\|_K\big) \leq \mathbb{E}\Big(\sum_{k=1}^{T-2}\eta_{k+1}\|\mathcal{A}_{k+1}\|\|\mathcal{A}_{k+2}^T\|\|\mathcal{S}_k - \mathcal{S}_*\|_K\Big) + \mu(\lambda)\eta_T\mathbb{E}\Big(\|\mathcal{S}_{T-1} - \mathcal{S}_*\|_K\Big). \quad (4.14)$$

By Lemma 3 and Lemma 6, for any $\theta \in (1/2, 1]$, the above upper bound is dominated by

$$\begin{aligned}\mathbb{E}(\|I_2\|_K) \leq{}& \frac{\sigma}{\mu(\lambda)}\Big(\frac{1}{2\theta-1}\Big)^{1/2}\sum_{k=1}^{T-2}\exp\Big\{-\frac{\lambda}{\mu(\lambda)}\sum_{j=k+2}^{T} j^{-\theta}\Big\}(k+1)^{-\theta}k^{1/2-\theta} \\ &+ \frac{\sigma}{\mu(\lambda)}\Big(\frac{2}{2\theta-1}\Big)^{1/2}T^{1/2-\theta}. \quad (4.15)\end{aligned}$$

Since $\lambda < \mu(\lambda)$, we apply Lemma 7 with $\theta \in (1/2, 1)$ and $\nu = \frac{\lambda}{\mu(\lambda)}$ to the righthand side of the above inequality which implies that

$$\mathbb{E}\big(\|I_2\|_K\big) \leq \frac{7\sigma}{\lambda}\Big(\frac{2}{2\theta-1}\Big)^{1/2} \ T^{1/2-\theta} \ + \frac{6\sigma}{\mu(\lambda)}\Big(\frac{1}{2\theta-1}\Big)^{1/2} C_\theta(T) \ e^{-D_{\theta,\lambda}T^{1-\theta}}. \tag{4.16}$$

Combining with the bounds (4.12), (4.13) and (4.16), we conclude that for any $\theta \in (1/2, 1)$, $\mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_K\big)$ is bounded by

$$\frac{8\sigma}{\lambda}\Big(\frac{2}{2\theta-1}\Big)^{1/2} T^{1/2-\theta} + \bigg[(6C_\theta(T) + \sqrt{2})\frac{\sigma}{\mu(\lambda)}\Big(\frac{1}{2\theta-1}\Big)^{1/2} + \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}\bigg] e^{-D_{\theta,\lambda}T^{1-\theta}}.$$

This completes Theorem 2. □

We turn our attention to the proof of Theorem 3.

Proof of Theorem 3. Let us estimate the four terms on the righthand side of the equality (4.10) one by one.

Using (4.9) with $k = 0$, then for $T \geq 2$, there holds

$$\sum_{j=2}^{T} j^{-1} \geq \ln\Big(\frac{T+1}{2}\Big).$$

Therefore, the estimates (3.8), (3.9) in connection with the bounds of (2.5) and (4.1) imply that

$$\mathbb{E}\Big(\|I_1\|_K + \|I_4\|_K\Big) \leq \Big[\frac{\sqrt{2}\sigma}{\mu(\lambda)} + \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}\Big]\Big(\frac{2}{T+1}\Big)^{\frac{\lambda}{\mu(\lambda)}}.$$

Lemma 6 gives us the estimate for the expectation of the norm of the third term in (4.10) as follows

$$\mathbb{E}\Big(\|I_3\|_K\Big) \leq \frac{\sigma}{\mu(\lambda)} T^{-1/2}.$$

By Lemma 7 with $\theta = 1$ and $\nu = \frac{\lambda}{\mu(\lambda)}$, it yields from (4.15) that

$$\begin{aligned}\mathbb{E}\Big(\|I_2\|_K\Big) &\leq \frac{\sqrt{2}\sigma}{\mu(\lambda)} T^{-1/2} + \frac{4}{\mu(\lambda)\sigma}\sum_{k=1}^{T-2} \exp\Big\{-\frac{\lambda}{\mu(\lambda)}\sum_{j=k+2}^{T} j^{-1}\Big\}(k+1)^{-1}k^{-1/2} \\ &\leq \frac{\sqrt{2}\sigma}{\mu(\lambda)} T^{-1/2} + \frac{4\sigma C'(T)}{\mu(\lambda)}\Big(\frac{1}{T}\Big)^{\frac{\lambda}{\mu(\lambda)}}.\end{aligned}$$

where

$$C'(T) := \begin{cases} \ln(T+1), & \frac{\lambda}{\mu(\lambda)} = 1/2 \\ \frac{2\mu(\lambda)}{\mu(\lambda)-2\lambda}\Big[1 - (T+1)^{-1/2+\lambda/\mu(\lambda)}\Big], & \frac{\lambda}{\mu(\lambda)} \in (0, 1/2) \cup (1/2, 1). \end{cases}$$

Combining with all the above estimates, we see that

$$\mathbb{E}\big(\|f_{T+1} - f_\lambda^V\|_K\big) \le 2\Big[\frac{2\sigma}{\mu(\lambda)}\big(1 + C'(T)\big) + \sqrt{\frac{2\mathcal{D}(\lambda)}{\lambda}}\Big]\Big(\frac{1}{T}\Big)^{\frac{\lambda}{\mu(\lambda)}} + \frac{3\sigma}{\mu(\lambda)}\frac{1}{\sqrt{T}}$$

This completes the proof of Theorem 3. □

Acknowledgements: The author would like to thank Prof. D. X. Zhou and Prof. Charles A. Micchelli for their constructive suggestions and comments.

References

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

[2] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, Convexity, classification, and risk bounds, Preprint, Department of Statistics, University of California Berkeley, 2003.

[3] S.Boyd and L.Vandenberghe, Convex optimization, Cambridge, 2004.

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

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

[6] N.Cristianini and J. Shawe-Taylor, An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000.

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

[8] L.Devroye, L.Györfi and G.Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1997.

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

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

[11] B. Blanchard, G. Lugosi and N. Vayatis, On the rate of convergence of regularized boosting classifiers, J. Mach. Learning Res. **4** (2003), 861-894.

[12] P. Niyogi and F. Girosi, On the relationships between generalization error, hypothesis complexity and sample complexity for radial basis functions, Neural Computation, **8** (1996), 819-842.

[13] C. Scovel and I. Steinwart, Fast rates for support vector machines, Los Alamos National Laboratory Technical Report, 2005.

[14] S.Smale and Y.Yao, Online learning algorithms, Preprint, Department of Mathematics, University of California Berkeley, 2004.

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

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

[17] S.Smale and D.X.Zhou, Shannon sampling II: Connection to learning theory, Preprint, 2004.

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

[19] Q. Wu, Y.Ying and D. X. Zhou, Multi-kernel Regularized Classifiers, Submitted to J. of Complexity, Department of Mathematics, City University of Hong Kong, 2004.

[20] Y.Ying and D.X.Zhou, Learnability of Gaussians with flexible variances, Preprint, Department of Mathematics, City University of Hong Kong, 2004.

[21] Y.Ying and D.X.Zhou, Online regularized classification algorithms, Preprint, 2005.

[22] T.Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Annals Statis. **32** (2004), 56-85.

[23] D. X. Zhou, The covering number in learning theory, J. Complexity **18** (2002), 739-767.

[24] D. X. Zhou, Capacity of reproducing kernel spaces in learning theory, IEEE Trans. Inform. Theory **49** (2003), 1743-1752.