Strong Convex Loss Can Increase the Learning Rates of Online Learning

Baohuai Sheng
Department of Mathematical Sciences, Shaoxing College of Arts and Sciences, Shaoxing, China

Liqin Duan
Mathematics & Science College, Shanghai Normal University, Shanghai, China

Peixin Ye
School of Mathematics and LPMC, Nankai University, Tianjin, China
yepx@nankai.edu.cn

***Abstract*—It is known that kernel regularized online learning has the advantages of low complexity and simple calculations, and thus is accompanied with slow convergence and low accuracy. Often the algorithm are designed with the help of gradient of the loss function, the complexity of the loss may influence the convergence. In this paper, we show, at some extent, the strong convexity can increase the learning rates.**

***Index Terms*—Online learning, Strong convex loss, Learning rates**

I. Introduction

Online learning is an important research area of machine learning (see [1,2,3,4]). In addition to the novel learning theory questions that they arise, online algorithms are also attractive in processing large data sets since they process one example at a time and can be more efficient than that of the batch algorithms (see [5,6,7,8,9]).

Let $K: X\times X\to R$ be a function of continuous, symmetric and positive semi-definite, i.e., for any finite set of distinct points $\{x_1,x_2,\ldots,x_T\}\subset X$, the matrix $\left(K(x_i,x_j)\right)_{i,j=1}^{T}$ is positive semi-definite. Such a kernel is called a Mercer kernel. The RKHS H_K associated with the kernel K is defined to be the closure 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 $\langle K_x, K_t\rangle_K = K(x,t)$. The reproducing property is given by

$$\langle K_x, f\rangle_K = f(x), \forall x\in X, f\in H_K.$$

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

$$\|f\|_\infty \le k\|f\|_K, k=\sqrt{\sup_x K(x,x)}, \forall f\in H_K.$$

Let X be a compact subset of R^n and $Y=\{-1,1\}$. The relation between the input $x\in X$ and the output $y\in Y$ is described by a probability distribution $\rho(x,y)=\rho(y\mid x)\rho_X(x)$ on $Z=X\times Y$, where $\rho(y\mid x)$ is the conditional probability of y given x and $\rho_X(x)$ is the marginal probability of x. The distribution ρ is known only through a set of samples $Z=\{z_i\}_{i=1}^T=\{(x_i,y_i)\}_{i=1}^T$ independently drawn according to ρ.

Classification algorithms produce binary classifiers $C: X\to Y$. The misclassification error is used to measure the prediction power of a classifier C. If ρ is a probability distribution on $Z=X\times Y$, then the misclassification error of C is defined by

$$\Re(C) := \mathrm{Prob}\{C(x)\neq y\} = \int_X P(y\neq C(x)\mid x)d\rho_X.$$

Here $P(y\mid x)$ is the conditional probability at $x\in X$. The classifier minimizing the misclassification error is called the Bayes rule f_c and is given by

$$f_c=\begin{cases}1, & P(y=1\mid x)\ge P(y=-1\mid x),\\ -1, & otherwise.\end{cases}$$

The performance of a classifier C can be measured by the excess misclassification error $\Re(C)-\Re(f_c)$.

The binary classifiers $C: X\to Y$ may be induced from real functions $f: X\to R$ by $C_f=\mathrm{sgn}(f)$ which is defined by $\mathrm{sgn}(f)(x)=1$ if $f(x)\ge 0$ and $\mathrm{sgn}(f)(x)=-1$ otherwise.

A loss function $V: R\to R_+$ is often used for the real-valued function f to measure the local error suffered

Corresponding author: qinandfeng@163.com;
Manuscript received January 1, 2013; revised June 1, 2014; accepted July 1, 2014.
This work was supported by the National Science Foundation of China (Grant No. 11201104,11271199).

doi:10.4304/jcp.9.7.1606-1611

from the use of $\mathrm{sgn}(f)$ as a model for the process producing y at $x \in X$.

The batch learning algorithm for a classification associated with RKHS H_K, the sample and a classification loss function $V(t)$ is

$$f_{z,\lambda} := \arg\min_{f \in H_K} \left\{ \frac{1}{T} \sum_{t=1}^{T} V(y_t f(x_t)) + \frac{\lambda}{2} \| f \|_K^2 \right\},$$

where we call $V(t)$ a normal classification loss function if it is a convex function on R^1, $V'(0) < 0$ and 1 is its minimal zero. The most usual classification loss is the least square loss $V(t) = (1-t)^2$.

It is easy to see that $f_{z,\lambda}$ has behaviors similar to regularization function $f_\lambda^V \in H_K$ defined by

$$f_\lambda^V := \arg\min_{f \in H_K} \left\{ \boldsymbol{\mathcal{E}}(f) + \frac{\lambda}{2} \| f \|_K^2 \right\}.$$

where

$$\boldsymbol{\mathcal{E}}(f) := \int_Z V(yf(x)) d\rho.$$

We call (1) the batch learning scheme since it uses all the samples up once, which makes the computations more complexity when the number (i.e., the T) of samples is large. On the contrary, online learning algorithms operate by repetitively drawing random examples, one at a time, and adjusting the learning variables using simple calculations that are usually based on the single example only. Of course, the low computational complexity (per iteration) of online algorithms is often associated with their slow convergence and low accuracy in solving the underlying optimization problems. Therefore, the investigation on the problem of what cause will influence the performance is needed. Many papers have devoted to this field (see e.g.[10,11,12,13]). Among the researches, [14] defined a kind of general classification learning algorithm associating with convex loss and reproducing kernel spaces and showed the convergence rates. The algorithm is improved in as the fully online learning algorithms. On this basis, [16] defined a kind of online classification learning algorithm with the generalized gradient of the loss function. The new online algorithm needs only less additional assumption on the loss and derives a strong convergence rate in case of convex loss (the algorithm are redesigned in [17,18] basing on the strong convexity of the loss).

Definition 1. The generalized gradient descent online algorithm is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t \left[G(y_t f_t(x_t)) y_t K_{x_t} + \lambda f_t \right], \quad t = 1, \ldots, T \qquad (2)$$

where $G(t) \in \partial V(t)$ and $\partial V(t)$ is the generalized gradient (see the Appendix) of $V(t)$ at t, $\eta_t > 0$ is the step size.

The problem that we are most interested in is whether the classifiers $\mathrm{sgn}(f_{T+1})$ will converge to f_c. The aim of theory analysis for the classification algorithm (2) is to bound the excess misclassification error

$$\Re(\mathrm{sgn}(f_{T+1})) - \Re(f_c). \qquad (3)$$

By [16] we know that if $V(t)$ is a convex loss and satisfies some differentiable assumptions, then there is a constant C_V such that for any measurable function f

$$\Re(f) - \Re(f_c) \le C_V (\boldsymbol{\mathcal{E}}(f) - \boldsymbol{\mathcal{E}}(f_\rho^V))^{1/2},$$

where

$$f_\rho^V = \arg\inf \{ \boldsymbol{\mathcal{E}}(f) : f \ is \quad measurable \quad on \quad X \}.$$

Moreover, there are

$$\begin{aligned} \boldsymbol{\mathcal{E}}(f) - \boldsymbol{\mathcal{E}}(f_\rho^V) &= \boldsymbol{\mathcal{E}}(f) - \boldsymbol{\mathcal{E}}(f_\lambda) + \boldsymbol{\mathcal{E}}(f_\lambda) - \boldsymbol{\mathcal{E}}(f_\rho^V) \\ &\le \boldsymbol{\mathcal{E}}(f) - \boldsymbol{\mathcal{E}}(f_\lambda) + D(\lambda), \end{aligned}$$

where

$$D(\lambda) = \inf_{f \in H_K} \left\{ \boldsymbol{\mathcal{E}}(f) - \boldsymbol{\mathcal{E}}(f_\rho^V) + \frac{\lambda}{2} \| f \|_K^2 \right\}$$

is called the regularization error which measures the approximation ability of the space H_K. If H_K is dense in $C(X)$, then we know $\lim_{\lambda \to 0} D(\lambda) = 0$. We usually assume that there is a constant $A > 0$ such that $D(\lambda) \le A\lambda^\beta$ (see [14,19]).

Also, by [16] we know there exists a constant C_{λ, f_λ} such that

$$\boldsymbol{\mathcal{E}}(f) - \boldsymbol{\mathcal{E}}(f_\lambda) \le C_{\lambda, f_\lambda} \| f - f_\lambda \|_K.$$

Then, to bound (3) we need to estimate $\| f_{T+1} - f_\lambda \|_K$. When $V(t)$ is a normal convex classification loss, [16] shows that if we choose the step as $\eta_t = \frac{1}{\mu(\lambda) t^\theta}$ for some $\theta \in (0,1]$ and $\mu(\lambda) \ge M(\lambda) k^2 + \lambda$. Define $\{ f_t \}$ by (2) and $\overline{C}_\lambda = 4k^2 \| V^o \|_{C_\infty[-k^2|G(0)/\lambda|, k^2|G(0)/\lambda|]}$ for $t = 1, \ldots, T$. Then

$$\begin{aligned} & E_{x \in Z^T} \left(\| f_{T+1} - f_\lambda \|_K^2 \right) \\ & \le \begin{cases} \left(\frac{2D(\lambda)}{\lambda} + \frac{9\overline{C}_\lambda T^{1-\theta}}{(1-\theta) 2^{1-\theta} (\mu(\lambda))^2} \right) \\ \quad \times \exp\left\{ -\frac{(1-2^{\theta-1})\lambda}{3(1-\theta)\mu(\lambda)} T^{1-\theta} \right\}, & 0 < \theta < 1, \\ \left(\frac{2D(\lambda)}{\lambda} + \frac{25\overline{C}_\lambda}{\mu(\lambda)(3\mu(\lambda) - \lambda)} \right) \times T^{-\frac{\lambda}{3\mu(\lambda)}}, & \theta = 1. \end{cases} \end{aligned} \qquad (4)$$

On the other hand, we notice that, besides [17,18], there are other papers(see e.g.[10]) which borrow the strong convexity of the loss function to design online algorithm. Then, whether the convergence is influenced by the strong convexity is a topic needed to be investigated. This is the main motivation for writing the present paper. We give the following results.

Theorem 1. Let $V(t)$ be a strong convex loss with modulus $0 < c < 3$. Define $\{ f_t \}_{t=1}^T$ as in (2) and

$$\overline{C}_\lambda = kL^2_{V, \left[-\frac{k^2 G(0)}{\lambda}, \frac{k^2 G(0)}{\lambda} \right]} + 2kG(0) L_{V, \left[-\frac{k^2 G(0)}{\lambda}, \frac{k^2 G(0)}{\lambda} \right]} + k^2.$$

Then, we have for $\eta = t^{-\theta}, 0 < \theta < 1$ that

$$E_{x\in Z^T}\left(\|f_{T+1}-f_\lambda\|_K^2\right)$$
$$\le\begin{cases}\dfrac{D(\lambda)}{\lambda}\times\exp\left\{-\dfrac{3}{3(1-\theta)}\left[(T+1)^{1-\theta}-2^{1-\theta}\right]\right\}\\ \qquad+\dfrac{64\overline{C}_\lambda}{T^\theta}+\dfrac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}} & (5)\\ \qquad\times\exp\left\{-\dfrac{(1-2^{\theta-1})c(T+1)^{1-\theta}}{3(1-\theta)}\right\}, & 0<\theta<1,\\ \dfrac{D(\lambda)}{\lambda}\times\exp\left\{-\dfrac{c}{3}\log\dfrac{T+1}{2}\right\}+\dfrac{24\overline{C}_\lambda}{(3-c)(T+1)^{c/3}}, & \theta=1.\end{cases}$$

Comparing (5) with (4), we can see that the strong convexity actually increase the learning rates since the modulus c.

II. PROOFS

To prove (5), we need some lemmas.

Lemma 1. (see [16]) Let V be a strongly convex with modulus $c>0$, $\{f_t\}$ be defined by (2). Then

$$\|f_t\|_K\le\frac{k|G(0)|}{\lambda},\qquad\forall t\in\mathrm{N}. \tag{6}$$

(6) shows that the sequence $\{f_t\}_{t=1}^T$ is bounded in H_K.

Lemma 2. Assume $V(x)$ is a convex loss function, then for any $G(yf_\lambda^V(x))\in\partial V(yf_\lambda^V(x))$ and any $f\in H_K$ there holds

$$\left\langle\int_Z yG(yf_\lambda^V(x))d\rho, f-f_\lambda^V\right\rangle_K+\lambda\left\langle f_\lambda^V, f-f_\lambda^V\right\rangle_K=0. \tag{7}$$

Proof. Since $V(x)$ is a strong convex function, we know $\mathcal{E}(f_\lambda^V)$ is also a strong convex function on H_K as well. Therefore, we have

$$0\le\frac{1}{\theta}\left\{\left(\mathcal{E}(f_\lambda^V+\theta f)+\frac{\lambda}{2}\|f_\lambda^V+\theta f\|_K^2\right)-\left(\mathcal{E}(f_\lambda^V)+\frac{\lambda}{2}\|f_\lambda^V\|_K^2\right)\right\}$$
$$\le\frac{1}{\theta}\left\langle\int_Z yG\left(y(f_\lambda^V(x)+\theta f(x))\right)d\rho,\theta f\right\rangle_K$$
$$+\lambda\left\langle f_\lambda^V, f\right\rangle_K+\frac{\lambda\theta}{2}\|f\|_K^2. \tag{8}$$

Taking $\theta\to0$, we have for any $f\in H_K$ that

$$0\le\left\langle\int_Z yG\left(yf_\lambda^V(x)\right)d\rho, f\right\rangle_K+\lambda\left\langle f_\lambda^V, f\right\rangle_K=0,$$

which together with the variousness of $f\in H_K$ gives

$$\left\langle\int_Z yG\left(yf_\lambda^V(x)\right)d\rho, f\right\rangle_K+\lambda f_\lambda^V=0.$$

Thus (7) holds.

Lemma 3. Let V be a strongly convex loss function with the modulus $c>0$ and $\lambda>0$. Then, for any $f\in H_K$, there holds

$$\frac{c}{6}\|f-f_\lambda^V\|_K^2\le\left(\mathcal{E}(f)+\frac{\lambda}{2}\|f\|_K^2\right)-\left(\mathcal{E}(f_\lambda^V)+\frac{\lambda}{2}\|f_\lambda^V\|_K^2\right). \tag{9}$$

Proof. Define a univariate function $H=H(\theta)$ on [0,1] by

$$H(\theta)=\mathcal{E}\left(f_\lambda^V+\theta(f-f_\lambda^V)\right)+\frac{\lambda}{2}\|f_\lambda^V+\theta(f-f_\lambda^V)\|_K^2,$$

$\theta\in[0,1]$, $f\in H_K$.

Then

$$H(1)=\mathcal{E}(f)+\frac{\lambda}{2}\|f\|_K^2,\quad H(0)=\mathcal{E}(f_\lambda^V)+\frac{\lambda}{2}\|f_\lambda^V\|_K^2.$$

Since $V(x)$ is strongly convex loss, as a function of θ, H is also strongly convex. Take $f_\theta=f_\lambda^V+\theta(f-f_\lambda^V)$. Then $H(\theta)$ can be rewritten as $H(\theta)=\mathcal{E}(f_\theta)+\frac{\lambda}{2}\|f_\theta\|_K^2$ and

$$H(\theta+\Delta\theta)-H(\theta)$$
$$=\mathcal{E}(f_{\theta+\Delta\theta})-\mathcal{E}(f_\theta)+\frac{\lambda}{2}\left(\|f_{\theta+\Delta\theta}\|_K^2-\|f_\theta\|_K^2\right).$$

On one hand, by the mean value theorem for the generalized gradient (see the Appendix), we have a $u\in(\theta,\theta+\Delta\theta)$ and a $\hat{G}(u)\in\partial H(u)$ such that

$$H(\theta+\Delta\theta)-H(\theta)=\Delta\theta\hat{G}(u),$$

and also there is $g\in(f_\theta, f_{\theta+\Delta\theta})$ such that

$$H(\theta+\Delta\theta)-H(\theta)$$
$$=\langle\mathcal{E}(g), f_{\theta+\Delta\theta}-f_\theta\rangle_K+\lambda\Delta\theta\left\langle f_\theta, f-f_\lambda^V\right\rangle_K+\frac{\lambda(\Delta\theta)^2}{2}\|f-f_\lambda^V\|_K^2.$$

When $\Delta\theta\to0$, then there is a $G(yf_\lambda^V(x))\in\partial V(yf_\lambda^V(x))$ such that

$$\hat{G}(\theta)=\left\langle\int_Z G\left(yf_\theta(x)\right)yd\rho, f-f_\lambda^V\right\rangle_K+\lambda\left\langle f_\theta, f-f_\lambda^V\right\rangle_K.$$

The definition of f_θ gives

$$\hat{G}(\theta)=\left\langle\int_Z yG(yf_\theta(x))d\rho, f-f_\lambda^V\right\rangle_K+\lambda\left\langle f_\lambda^V, f-f_\lambda^V\right\rangle_K$$
$$+\lambda\theta\left\langle f-f_\lambda^V, f-f_\lambda^V\right\rangle_K.$$

On the other hand, by (7) we have

$$\lambda\left\langle f_\lambda^V, f-f_\lambda^V\right\rangle_K=-\left\langle\int_Z yG\left(yf_\lambda^V(x)\right)d\rho, f-f_\lambda^V\right\rangle_K.$$

Therefore,

$$\hat{G}(\theta)=\left\langle\int_Z y\left(G(yf_\theta(x))-G(yf_\lambda^V(x))\right)d\rho, f-f_\lambda^V\right\rangle_K$$
$$+\lambda\theta\|f-f_\lambda^V\|_K^2. \tag{10}$$

Since V is a strongly convex function with modulus $c>0$, we have

$$\left\langle\int_Z y\left(G\left(yf_\theta(x)\right)-G\left(yf_\lambda^V(x)\right)\right)d\rho, f-f_\lambda^V\right\rangle_K\ge c\theta\|f-f_\lambda^V\|_K^2. \tag{11}$$

(10) can be rewritten as $\hat{G}(\theta)\ge(\lambda+c)\theta\|f-f_\lambda^V\|_K^2$. It follows

$$H(1)-H(0)=\int_0^1(H(1)-H(0))d\theta$$
$$=\int_0^1(H(1)-H(\theta))d\theta+\int_0^1(H(\theta)-H(0))d\theta$$
$$\ge\int_0^1\hat{G}(\theta)\cdot(1-\theta)d\theta+0$$
$$\ge(\lambda+c)\|f-f_\lambda^V\|_K^2\int_0^1\theta(1-\theta)d\theta$$
$$=\left(\frac{\lambda+c}{6}\right)\|f-f_\lambda^V\|_K^2\ge\frac{c}{6}\|f-f_\lambda^V\|_K^2.$$

(9) then holds.

We now give the quantitative description for the convergence of $\|f_{T+1}-f_{\lambda_T}^V\|_K$.

Lemma 4. Let the sequence $\{f_t\}_{t=1}^T$ be defined as (2) and there is a constant $C_\lambda > 0$ such that the sequence $H_t(x) = G(y_t f_t(x_t)) y_t K_{x_t}(x) + \lambda f_t(x)$ satisfies

$$E_{z_1,z_2,\cdots,z_t}\left(\|H_t\|_K^2\right) \le C_\lambda . \tag{12}$$

Then

$$E_{z_1,z_2,\cdots,z_t}\left(\left\|f_{t+1} - f_\lambda^V\right\|_K^2\right)$$
$$\le \left(1 - \frac{c\eta_t}{3}\right) \times E_{z_1,\cdots,z_{t-1}}\left(\left\|f_t - f_\lambda^V\right\|_K^2\right) + C_\lambda \eta_t^2 . \tag{13}$$

Proof. Rewrite the algorithm (2) by $f_{t+1} = f_t - \eta_t H_t$. Then, simple computation gives

$$\left\|f_{t+1} - f_\lambda^V\right\|_K^2 = \left\|f_t - f_\lambda^V - \eta_t H_t\right\|_K^2$$
$$= \left\|f_t - f_\lambda^V\right\|_K^2 + \eta_t^2 \|H_t\|_K^2 + 2\eta_t \left\langle H_t, f_\lambda^V - f_t\right\rangle_K . \tag{14}$$

Since

$$\left\langle H_t, f_\lambda^V - f_t\right\rangle_K = G(y_t f_t(x_t)) \cdot \left(y_t f_\lambda^V(x_t) - y_t f_t(x_t)\right) + \lambda \left\langle f_t, f_\lambda^V - f_t\right\rangle_K$$
$$\le V\left(y_t f_\lambda^V(x_t)\right) - V\left(y_t f_t(x_t)\right) + \frac{\lambda}{2}\left(\left\|f_\lambda^V\right\|_K^2 - \|f_t\|_K^2\right)$$

and f_t depends on $\{z_1, \ldots, z_{t-1}\}$ but not on z_t, we have

$$E_{z_t}\left(\left\langle H_t, f_\lambda^V - f_t\right\rangle_K\right) \le \left(\mathcal{E}(f_\lambda^V) + \frac{\lambda}{2}\|f_\lambda^V\|_K^2\right) - \left(\mathcal{E}(f_t) + \frac{\lambda}{2}\|f_t\|_K^2\right) . \tag{15}$$

Combining (15) with (9), we have

$$E_{z_1,\cdots,z_t}\left(\left\langle H_t, f_\lambda^V - f_t\right\rangle_K\right) \le -\frac{c}{6} E_{z_1,\cdots,z_{t-1}}\left(\left\|f_t - f_\lambda^V\right\|_K^2\right) . \tag{16}$$

(14),(16) and (12) give (13).

Lemma 5. Let $\{f_t\}_{t=1}^T$ satisfy the assumptions of Lemma 4. Then, we have for $T > t_0$ that

$$E_{z_1,\ldots,z_T}\left(\left\|f_{T+1} - f_\lambda^V\right\|_K^2\right)$$
$$\le \exp\left\{-\frac{c}{3}\sum_{t=t_0}^T \eta_t\right\} \times \mathrm{E}_{z_1,z_2,\ldots,z_{t_0-1}}\left(\left\|f_{t_0} - f_\lambda^V\right\|_K^2\right)$$
$$+ C_\lambda \sum_{t=t_0}^T \eta_t^2 \times \exp\left\{-\frac{c}{3}\sum_{j=t+1}^T \eta_j\right\} . \tag{17}$$

Proof. Applying the relation iteratively for $t = T, T-1, \ldots, t_0$, we have

$$E_{z_1,\ldots,z_T}\left(\left\|f_{T+1} - f_\lambda^V\right\|_K^2\right)$$
$$\le C_\lambda \eta_T^2 + \left(1 - \frac{c}{3}\eta_T\right)\left[\left(1 - \frac{c}{3}\eta_{T-1}\right) \times E_{z_1,\ldots,z_{T-2}}\left(\left\|f_{T-1} - f_\lambda^V\right\|_K^2\right) + C_\lambda \eta_{T-1}^2\right]$$
$$= C_\lambda \eta_T^2 + \left(1 - \frac{c}{3}\eta_T\right)\left(1 - \frac{c}{3}\eta_{T-1}\right) \times E_{z_1,\ldots,z_{T-2}}\left(\left\|f_{T-1} - f_\lambda^V\right\|_K^2\right)$$
$$+ C_\lambda \left(1 - \frac{c}{3}\eta_T\right)\eta_{T-1}^2$$
$$\le \cdots$$
$$\le \prod_{t=t_0}^T \left(1 - \frac{c}{3}\eta_t\right) \times E_{z_1,\cdots,z_{t_0-1}}\left(\left\|f_{t_0} - f_\lambda^V\right\|_K^2\right) + C_\lambda \sum_{t=t_0}^T \eta_t^2$$
$$\times \prod_{j=t+1}^T \left(1 - \frac{c}{3}\eta_j\right), \qquad \eta_{T+1} = 0 . \tag{18}$$

Since $1 - u \le e^{-u}$ for $u > 0$, we obtain (17) by (18).

Lemma 6. Let $\{f_t\}_{t=1}^T$ satisfy the assumptions of Lemma 4. Then, for $\eta = t^{-\theta}, 0 < \theta \le 1$ we have

$$E_{x \in Z^T}\left(\left\|f_{T+1} - f_\lambda^V\right\|_K^2\right)$$
$$\le \frac{2D(\lambda)}{\lambda} \times \exp\left\{-\frac{c}{3}\sum_{t=1}^T t^{-\theta}\right\} + C_\lambda \sum_{t=1}^T t^{-2\theta} \times \exp\left\{-\frac{c}{3}\sum_{j=t+1}^T j^{-\theta}\right\} . \tag{19}$$

To make precise estimate for the right side of (19), we cite two lemmas.

Lemma 7. (see Lemma 4 of [14]) For any $t < T$ and $0 < \theta \le 1$, there holds

$$\sum_{j=t+1}^T j^{-\theta} \ge \begin{cases} \dfrac{1}{1-\theta}\left[(T+1)^{1-\theta} - (t+1)^{1-\theta}\right], & \theta < 1, \\ \log(T+1) - \log(t+1), & \theta = 1. \end{cases} \tag{20}$$

Lemma 8. (see Lemma 5 of [14]) Let $0 < \upsilon \le 1$ and $0 < \theta \le 1$. Then

$$\sum_{t=1}^{T_1} \frac{1}{t^{2\theta}} \exp\left\{-\upsilon \sum_{j=t+1}^T j^{-\theta}\right\}$$

is bounded by

$$\begin{cases} \dfrac{18}{\upsilon T^\theta} + \dfrac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}} \times \exp\left\{-\dfrac{\upsilon(1-2^{\theta-1})}{1-\theta}(T+1)^{1-\theta}\right\}, & \theta < 1, \\ \dfrac{8}{1-\upsilon}(T+1)^{-\upsilon}, & \theta = 1. \end{cases} \tag{21}$$

Proof of Theorem 1. By (6) we know

$$\|f_t\|_{C(X)} \le k\|f_t\|_K \le \frac{k^2 G(0)}{\lambda} .$$

It follows $|y_t f_t(x_t)| \le \dfrac{k^2 G(0)}{\lambda}$ and for any $G(t) \in \partial V(t)$ we have

$$\left|G\left(y_t f_t(x_t)\right)\right| \le L_{V,\left[-\frac{k^2G(0)}{\lambda}, \frac{k^2G(0)}{\lambda}\right]} .$$

Therefore,

$$E_{z_1,z_2,\cdots,z_t}\left(\|H_t\|_K^2\right)$$
$$= E_{z_1,z_2,\cdots,z_t}\left(\left|G(y_t f_t(x_t))^2 K(x_t,x_t) + 2\lambda G(y_t f_t(x_t)) y_t f_t(x_t) + \lambda^2 f_t(x)^2\right|\right)$$
$$= E_{z_1,z_2,\cdots,z_t}\left(\left|\int_Z G(y f_t(x))^2 K(x,x) d\rho + 2\lambda \int_Z G(y f_t(x)) y f_t(x) d\rho\right.\right.$$
$$\left.\left. + \lambda^2 f_t(x)^2\right|\right) \le \overline{C}_\lambda . \tag{22}$$

Combining (22),(21) and (20) with (19), we have (5).

APPENDIX : RESULTS ON CONVEX ANALYSIS

Let $(X,\langle\cdot,\cdot\rangle)$ a given Hilbert space with inner product $\langle\cdot,\cdot\rangle$ which induces the norm $\|\cdot\|=\sqrt{\langle\cdot,\cdot\rangle}$. Then, we say a function $f: X\to R$ is strongly convex on X if for any $\lambda\in[0,1]$ and $x,x'\in X$ there holds

$$f(\lambda x+(1-\lambda)x')\le\lambda f(x)+(1-\lambda)f(x')-\frac{c\lambda(1-\lambda)}{2}\|x-x'\|^2,$$

where $c>0$ is the modulus of f . It is known that if f is a strongly convex function on X , then it is Lipschitz on X , i.e., there is a constant $L>0$ such that for any $x,x'\in X$ holds

$$|f(x)-f(x')|\le L\|x-x'\|.$$

The generalized gradient of f at x , denoted by $\partial f(x)$, is a subset of X defined by $\xi\in\partial f(x)$ if and only if for any $x\in X$ there holds for all $l\in X$

$$\langle\xi,l\rangle\le f^0(x;l)=\limsup_{x'\to x,t\downarrow 0}\frac{f(x'+tl)-f(x')}{t}.$$

By Theorem 6.1.2 of [20] we know if $f(x)$ is strongly convex with the modulus $c>0$ on X , then

$$\partial f(x)=\left\{\xi\in X: f(x')-f(x)\ge\langle\xi,x'-x\rangle+\frac{c}{2}\|x'-x\|^2\right\}$$

which equals that for all $\xi_i\in\partial f(x_i)$ holds

$$\langle\xi_2-\xi_1,x_2-x_1\rangle\ge c\|x_2-x_1\|^2.$$

Also, for any $\xi\in\partial f(x)$ there holds

$$f(x)-f(x')\le\langle\xi,x-x'\rangle.$$

The following mean value theorem for the generalized gradient is very important (see Theorem 2.3.7 of [20]). When f is Lipschitz on an open set containing the line segment $[x_1,x_2]$ for $x_1,x_2\in X$, then, there exists $u\in(x_1,x_2)$ such that

$$f(x_1)-f(x_2)\le\langle\partial f(u),x_1-x_2\rangle.$$

Moreover, by Proposition 6.2.2 and Theorem 3.1.2 of [20] we know if $V(t)$ is a convex function on R with Lipschitz constant $L>0$, then, the image $\partial V(B)$ for a bounded set $B\subset R$ is a bounded set and for any $\xi\in\partial V(B)$ there holds $\|\xi\|\le L$.

ACKNOWLEDGMENT

The authors thank the referees and editors for valuable suggestions.

REFERENCES

[1] J. Kivinen, A. J. Smola and R.C. Williamson, "Online learning with kernels," IEEE Trans. Signal Proc., 2004, vol.52, pp. 2165-2176, 2004.
[2] S.Smale, Y. Yao, "Online learning algorithms," Found. Comput. Math., vol.6, pp. 145-170, 2006.
[3] X.M.Dong, D.R.Chen, "Learning rates of gradient descent algorithm for classification, " J. Comp. Appl. Math. , vol.224, pp.182-192, 2009.
[4] Y.M.Ying, "Convergence analysis of online algorithms," Adv. .Compu. Math.,vol. 27, pp.273-291., 2007.
[5] B. H. Sheng, P. X. Ye, "Learning rates of support vector machine classifiers with data dependent hypothesis spaces," J. Computer, vol.7(1), pp.252-257, 2012.
[6] H.Z.Tong, D.R.Chen and L.Z.Peng, "Learning rates for regularized classifiers using multivariate, polynomial kernel," J. Complexity, vol.24(5-6), pp. 619-631,2008.
[7] I. Steinwart, "Support vector machines are universally consistent, "J. Complexity, vol. 18, pp.768-791 ,2002.
[8] S.Smale and D.X.Zhou, "Shannon sampling and function reconstruction from point values, " Bull. Amer. Math. Soc., vol.41, pp.279-305.,2004.
[9] G.P.Li, "Batch-to-Batch iterative learning control for end-point qualities based on kernel principal component regression model," J. Computer, vol.12(8),pp.3184-3190,2013.
[10] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," J. Mach. Learn. Res., vol.11,pp.2543-2596, 2010.
[11] Y. Xie, B. Luo, R.B. Xu and S.B. Chen, "Smooth harmonic transductive learning," J. Computer, vol.8(12), pp.3079-3085,2013.
[12] X.X. Yang, Y.Z. Zhang and J.H. Shi, "Iterative learning control on carrying robot," J. Computer, vol.9(1),pp.196-201, 2014.
[13] S.F.Ding, J.Z.Yu, H.J. Huang and H. Zhao, "Twin support vector machines based on particle swarm optimization," J. Computer,vol.8(9), pp.2296-2303, 2013.
[14] Y.Ying, D.X.Zhou, "Online regularized classification algorithms, "IEEE. Trans. Inform. Theory, vol.52, pp.4775-4788, 2006.
[15] G.B.Ye, D.X.Zhou, "Fully online classification by regularization, " Appl. Comput. Harmon. Anal.,vol. 23, pp.198-214, 2007.
[16] L.L.Zhang, B.H.Sheng and J.L.Wang, "Online regularized generalized gradient classification algorithm," Anal. Theory Appl., vol.26(3),pp.278-300, 2010.
[17] B.H.Sheng, P.X.Ye, "Fully online regularized classification algorithm with strongly convex loss," High Performance Networking, Computing, and Communication Systems Communications in Computer and Information Science, vol.163, pp.223-228, 2011.
[18] M.D.Tian, B.H.Sheng, "Convergence of coefficient regularized fully online algorithm," 2011 International Conference on Multimedia Technology (ICMT), Hangzhou,26-28, July 2011,pp.2059-2065.
[19] D.R.Chen, Q.Wu, Y.Ying and D.X.Zhou, "Support vector machine soft margin classifiers: error analysis,"J.Mach. Learn. Res. , vol.5, pp.1143-1175, 2004.
[20] A. Chenciner, S.S.Chern, et.al., Fundamentals of Convex Analysis, Springer, New York, 2004.

Baohuai Sheng attended Baoji Normal College, Baoji, Shaanxi, from1981 to 1985. He earned his BS degree in mathematical teaching from the department of mathematics in 1985. He earned his MS degree in basic mathematics from the department of mathematics of Hangzhou University in 1988. From 1988-2001 he worked towards the Doctor of Sciences in the applied mathematics at Xi dian University, Xian, P. R. China, and earned his S. D. degree in March, 2001. From March, 2001, to September, 2003, he served as a Professor at the Ningbo University. Since September, 2003, he has been a faculty member of Shaoxing College of Arts and Sciences, Shaoxing, Zhejiang, P. R. China, where he serves currently as a Professor and the Chairman of Mathematical Department. His current research interests focus on the area of approximation theory, nonlinear optimization, learning theory.

Liqin Duan Shanghai, China. Birthdate: August, 1978. is Mathematics Ph.D., graduated from School of Mathematics Science, Beijing Normal University. And research interests on approximation theory and information-based complexity. She is a lecturer of Mathematics & Science College, Shanghai Normal University.

Peixin Ye Tianjin, China. Birthdate: December, 1972. He received the M.S. degree in mathematics from Xiamen University, Fujian, China, in 1998 and the Ph.D. degree in mathematics from Beijing Normal University, in 2001.He is a Full Professor at Nankai University. He has published more than 50 journal and conference papers. His current research interests include information complexity, approximation theory, machine learning and compressed sensing.