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 \( K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \) be a function of continuous, symmetric and positive semi-definite, i.e., for any finite set of distinct points \( \{x_1, x_2, \ldots, x_T\} \subseteq \mathbb{R}^n \), the matrix \( (K(x_i, x_j))_{i,j=1}^T \) 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 \( \{K_x: x \in \mathbb{R}^n\} \) with the inner product \( \langle K_x, K_{t} \rangle_K = K(x, t) \). The reproducing property is given by

\[
\langle K_x, f \rangle_K = f(x), \forall x \in \mathbb{R}^n, f \in H_K.
\]

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

\[
\|f\| \leq k \|f\|_K, k = \sqrt{\sup_{x \in X} K(x, x)}, \forall f \in H_K.
\]

Let \( X \) be a compact subset of \( \mathbb{R}^n \) 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 \( \rho \) is known only through a set of samples \( Z = \{(x_i, y_i)\}_{i=1}^T \) independently drawn according to \( \rho \).

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

\[
\mathcal{R}(C) := \text{Prob}(C(x) \neq y) = \int_X P(y \neq C(x) \mid x) d\rho_x.
\]

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

\[
f^* = \begin{cases} 1, & P(y = 1 \mid x) \geq P(y = -1 \mid x), \\ -1, & \text{otherwise}. \end{cases}
\]

The performance of a classifier \( C \) can be measured by the excess misclassification error \( \mathcal{R}(C) - \mathcal{R}(f^*) \).

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

A loss function \( V: \mathbb{R} \rightarrow \mathbb{R} \) is often used for the real-valued function \( f \) to measure the local error suffered
from the use of \( \text{sgn}(f) \) as a model for the process producing \( y \) at \( x \in X \).

The batch learning algorithm for a classification associated with RKHS \( H_{\lambda} \), the sample and a classification loss function \( V(t) \) is

\[
f_{t,\lambda} := \arg\min_{f \in H_{\lambda}} \left\{ \frac{1}{T} \sum_{t=1}^{T} V(y, f(x_t)) + \frac{\lambda}{2} \|f\|_H^2 \right\}
\]

where we call \( V(t) \) a normal classification loss function if it is a convex function on \( \mathbb{R}^n, V'(0) < 0 \) 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 \( f_{t,\lambda} \) has behaviors similar to regularization function \( f_\lambda \) defined by

\[
f_\lambda := \arg\min_{f \in H_{\lambda}} \left\{ \mathcal{E}(f) + \frac{\lambda}{2} \|f\|_H^2 \right\},
\]

where

\[
\mathcal{E}(f) := \int V(yf(x))d\rho.
\]

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 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 descent online algorithm associating with RKHS \( KH \) and showed the convergence rates. 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 descent online algorithm associating with RKHS \( KH \) and showed the convergence rates. 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 descent online algorithm associating with RKHS \( KH \) and showed the convergence rates. 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 descent online algorithm associating with RKHS \( KH \) and showed the convergence rates.

**Definition 1.** The generalized gradient descent online algorithm is defined by \( f_t = 0 \) and

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

where \( G(t) \) 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 \( \text{sgn}(f_{T+1}) \) will converge to \( f \). The aim of theory analysis for the classification algorithm (2) is to bound the excess misclassification error

\[
\mathcal{R}(\text{sgn}(f_{T+1})) - \mathcal{R}(f).
\]

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

\[
\mathcal{R}(f) - \mathcal{R}(f_t) \leq C_r(\mathcal{E}(f) - \mathcal{E}(f_t))^{1/2},
\]

where

\[
f_t = \arg\min \{ \mathcal{E}(f) : f \text{ is measurable on } X \}.
\]

Moreover, there are

\[
\mathcal{E}(f) - \mathcal{E}(f_t) = \mathcal{E}(f) - \mathcal{E}(f_t) + \mathcal{E}(f_t) - \mathcal{E}(f_t) 
\leq \mathcal{E}(f) - \mathcal{E}(f_t) + D(\lambda),
\]

where

\[
D(\lambda) = \inf_{f, \lambda} \left\{ \mathcal{E}(f) - \mathcal{E}(f_t) + \frac{\lambda}{2} \|f\|_H^2 \right\}
\]

is called the regularization error which measures the approximation ability of the space \( H_{\lambda} \). If \( H_{\lambda} \) 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 \( D(\lambda) \leq A \lambda^{\beta} \) (see [14,19]).

Also, by [16] we know there exists a constant \( C_{r, \lambda} \) such that

\[
\mathcal{E}(f) - \mathcal{E}(f_t) \leq C_{r, \lambda} \|f_t - f\|_H.
\]

Then, to bound (3) we need to estimate \( \|f_{T+1} - f_t\|_H \). When \( V(t) \) is a normal convex classification loss, [16] shows that if we choose the step as \( \eta_t = \frac{1}{\mu(\lambda)^2} \) for some \( \theta \in (0,1] \) and \( \mu(\lambda) \geq M(\lambda)^k + \lambda \). Define \( f_t \) by (2) and \( \tilde{C}_{\lambda} = 4k^2 \|V_t\|_H^2 \|_{L^2[\mathbb{R}, \lambda]} \) for \( \lambda \in [\lambda_{\min}, \lambda_{\max}] \). Then

\[
E_{\mathbb{R}^T} \left( \|f_{T+1} - f_t\|_H^2 \right) \leq \begin{cases} \frac{2D(\lambda)}{\tilde{C}_{\lambda}} + \frac{9\tilde{C}_{\lambda}T^{-\theta}}{(1 - \theta)^2(\mu(\lambda))^{2/3}} & \quad \text{if } 0 < \theta < 1, \\ \frac{2D(\lambda)}{\tilde{C}_{\lambda}} + \frac{25\tilde{C}_{\lambda}}{(3\mu(\lambda) - \lambda)}T^{3/2} & \quad \text{if } \theta = 1. \end{cases}
\]

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 \( f_{T+1} \) as in (2) and \( \tilde{C}_{\lambda} = k\lambda^2 \left( \frac{\|V_t\|_{L^2[\mathbb{R}, \lambda]}^2}{\lambda} \right)^2 + 2kG(0) L \). Then, we have for \( \eta_t = \frac{1}{\mu(\lambda)^2}, 0 < \theta < 1 \) that

\[
E_{\mathbb{R}^T} \left( \|f_{T+1} - f_t\|_H^2 \right) \leq \begin{cases} \frac{2D(\lambda)}{\tilde{C}_{\lambda}} + \frac{9\tilde{C}_{\lambda}T^{-\theta}}{(1 - \theta)^2(\mu(\lambda))^{2/3}} & \quad \text{if } 0 < \theta < 1, \\ \frac{2D(\lambda)}{\tilde{C}_{\lambda}} + \frac{25\tilde{C}_{\lambda}}{(3\mu(\lambda) - \lambda)}T^{3/2} & \quad \text{if } \theta = 1. \end{cases}
\]
\[
E_{\text{seq}}(f_{T+1} - f_k^2) = \left\{ \begin{array}{l}
\frac{D(\lambda)}{\lambda} \exp \left\{ -\frac{3}{2(1-\theta)} [(T+1)^{1-\theta} - 2^{-\theta}] \right\} \\
\quad + \frac{64(\lambda)}{T^2} + \frac{9\theta T^{-\theta}}{(1-\theta)2^{-\theta}} \\
\quad \cdot \exp \left\{ -\frac{(1-2^{-\theta})c(T+1)^{1-\theta}}{(3-\theta)(T+1)} \right\},
\end{array} \right.
\]

(5)

\[ \iff K \iff H K \]

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 function with modulus \( c > 0 \), \( \{f_t\} \) be defined by (2). Then
\[
\|f_t\| \leq \frac{K}{\lambda} G(0), \quad \forall t \in \mathbb{N}.
\]

(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 \( G(yf_k^*(x)) \in \partial V(yf_k^*(x)) \) and any \( f \in H_k \), there holds
\[
\left\{ \int_x yG(yf_k^*(x))d\rho, f-f_k^* \right\}_k \leq 0.
\]

(7)

Proof. Since \( V(x) \) is a strongly convex function, we know \( \mathcal{E}(f_k^*) \) is also a strongly convex function on \( H_k \) as well. Therefore, we have
\[
0 \leq \frac{1}{\theta} \left\{ \mathcal{E}(f_k^* + \theta f) + \frac{\lambda}{2} \|f_k^* + \theta f\| \right\} - \left\{ \mathcal{E}(f_k^*) + \frac{\lambda}{2} \|f_k^*\| \right\}
\]
\[
\leq \frac{1}{\theta} \left\{ \int_x yG(yf_k^*(x) + \theta f(x))d\rho, f, f_k^* \right\}_k
\]
\[
+ \lambda \left\{ f_k^*, f \right\}_K + \frac{\lambda \theta}{2} \|f_k^*\| - \|f_k^*\|.
\]

(8)

Taking \( \theta \to 0 \), we have for any \( f \in H_k \) that
\[
0 \leq \left\{ \int_x yG(yf_k^*(x))d\rho, f, f_k^* \right\}_K + \lambda \left\{ f_k^*, f \right\}_K = 0,
\]
which together with the various of \( f \in H_k \) gives
\[
\left\{ \int_x yG(yf_k^*(x))d\rho, f, f_k^* \right\}_K + \lambda f_k^* = 0.
\]

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
\[
\|f_f^* - f_k^*\| \leq \mathcal{E}(f_f^*) + \frac{\lambda}{2} \|f_f^*\| - \mathcal{E}(f_k^*) + \frac{\lambda}{2} \|f_k^*\|.
\]

(9)

Proof. Define a univariate function \( H = \mathcal{E}(f) + \frac{\lambda}{2} \|f\| \) on \([0, 1] \) by
\[
H(\theta) = \mathcal{E}(f_f^* + \theta(f_f^* - f_k^*)) + \frac{\lambda}{2} \|f_f^* + \theta(f_f^* - f_k^*)\|,
\]
\[ \theta \in [0, 1], f \in H_k. \]

Then
\[
H(1) = \mathcal{E}(f) + \frac{\lambda}{2} \|f\|, \quad H(0) = \mathcal{E}(f_f^*) + \frac{\lambda}{2} \|f_f^*\|.
\]

Since \( V(x) \) is strongly convex loss, as a function of \( \theta \), \( H \) is also strongly convex. Take \( f_0 = f_f^* + \theta(f_f^* - f_k^*) \). Then \( H(\theta) \) can be rewritten as \( H(\theta) = \mathcal{E}(f_0) + \frac{\lambda}{2} \|f_0\| \) and
\[
H(\theta + \Delta \theta) - H(\theta) = \mathcal{E}(f_0) + \frac{\lambda}{2} \|f_0\|.
\]

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 \( \hat{G}(u) \in \partial H(u) \) such that
\[
H(\theta + \Delta \theta) - H(\theta) = \Delta \theta \hat{G}(u),
\]
and also there is \( g \in (f_0, f_0 + \Delta \theta) \) such that
\[
H(\theta + \Delta \theta) - H(\theta) = \Delta \theta \hat{G}(u).
\]

(10)

(10) can be rewritten as \( \hat{G}(u) \geq \theta(\lambda + c) \theta \|f - f_k^*\| \). It follows
\[
H(1) - H(0) = \int_0^1 (H(1) - H(0))d\theta
\]
\[
\geq \int_0^1 \hat{G}(\theta)(1 - \theta)d\theta
\]
\[
\geq \int_0^1 (\lambda + c) \theta(1 - \theta)d\theta
\]
\[
\geq \frac{\lambda + c}{6} \|f - f_k^*\| \geq \frac{\lambda}{6} \|f - f_k^*\|.
\]

(11)

(9) then holds.

We now give the quantitative description for the convergence of \( \|f_{T+1} - f_k^*\| \).
Lemma 4. Let the sequence \( \{f_i\}_{i=0}^\infty \) be defined as (2) and there is a constant \( C_f > 0 \) such that the sequence \( H_t(x) = G(y_i f_i(x)) y_i K_{h_i}(x) + \lambda f_i(x) \) satisfies
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| H_t \right\|_2^2 \right) \leq C_f.
\] (12)

Then
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| f_{t+1} - f_{t+1}^* \right\|_2^2 \right)
\leq \left( 1 - \frac{c \eta_t}{3} \right) E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| f_t - f_t^* \right\|_2^2 \right) + C_f \eta_t^2.
\] (13)

Proof. Rewrite the algorithm (2) by \( f_{t+1} = f_t - \eta_t H_t \). Then, simple computation gives
\[
\left\| f_{t+1} - f_{t+1}^* \right\|_2^2 = \left\| f_t - f_t^* - \eta_t H_t \right\|_2^2
= \left\| f_t - f_t^* \right\|_2^2 + \eta_t^2 \left\| H_t \right\|_2^2 + 2 \eta_t \left\langle f_t^* - f_t, H_t \right\rangle.
\] (14)

Since
\[
\left\langle H_t, f_t^* - f_t \right\rangle = G(y_i f_i(x)) \left( y_i f_i^*(x) - y_i f_i(x) \right) + \lambda \left( f_t^* - f_t \right)
\leq V \left( y_i f_i^*(x) \right) - V \left( y_i f_i(x) \right) + \frac{\lambda}{2} \left( \left\| f_t^* \right\|_2^2 - 1 \right),
\]
and \( f_t \) depends on \( \{z_1, \ldots, z_t\} \) but not on \( z_t \), we have
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| H_t \right\|_2^2 \right) \leq \left( 1 - \frac{c \eta_t}{3} \right) E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| f_t - f_t^* \right\|_2^2 \right) + C_f \eta_t^2.
\] (15)

Combining (15) with (9), we have
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| H_t \right\|_2^2 \right) \leq \frac{c}{6} E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| f_t - f_t^* \right\|_2^2 \right),
\] (16)
(14),(16) and (12) give (13).

Lemma 5. Let \( \{f_i\}_{i=0}^\infty \) satisfy the assumptions of Lemma 4.

Then, we have for \( t > T \) that
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| f_{t+1} - f_{t+1}^* \right\|_2^2 \right)
\leq \exp \left\{ \frac{c}{3} \sum_{j=1}^T \eta_j \right\} E_{_{\eta_1,\eta_2,\cdots,\eta_T}} \left( \left\| f_T - f_T^* \right\|_2^2 \right) + C_T \sum_{j=1}^T \eta_j.
\] (17)

Proof. Applying the relation iteratively for \( t = T, T-1, \ldots, t \), we have
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_t}} \left( \left\| f_{t+1} - f_{t+1}^* \right\|_2^2 \right)
\leq C_T \eta_T + \left( 1 - \frac{c}{3} \eta_T \right) \exp \left\{ \frac{c}{3} \sum_{j=1}^{T-1} \eta_j \right\} E_{_{\eta_1,\eta_2,\cdots,\eta_{T-1}}} \left( \left\| f_{T-1} - f_{T-1}^* \right\|_2^2 \right) + C_{T-1} \eta_{T-1}\]
\[
= C_T \eta_T + \left( 1 - \frac{c}{3} \eta_T \right) \exp \left\{ \frac{c}{3} \sum_{j=1}^{T-2} \eta_j \right\} E_{_{\eta_1,\eta_2,\cdots,\eta_{T-2}}} \left( \left\| f_{T-2} - f_{T-2}^* \right\|_2^2 \right) + C_{T-2} \eta_{T-2} + \cdots
\]
\[
\leq \cdots \leq \sum_{i=0}^T \left( 1 - \frac{c}{3} \eta_{i+1} \right) \exp \left\{ \frac{c}{3} \sum_{j=1}^{i} \eta_j \right\} E_{_{\eta_1,\eta_2,\cdots,\eta_{i+1}}} \left( \left\| f_{i+1} - f_{i+1}^* \right\|_2^2 \right) + C_i \sum_{j=1}^i \eta_j.
\]

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

Lemma 6. Let \( \{f_i\}_{i=0}^\infty \) satisfy the assumptions of Lemma 4.

Then, for \( \eta = T^\theta \geq 0 < \theta \leq 1 \) we have
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_T}} \left( \left\| f_{T+1} - f_{T+1}^* \right\|_2^2 \right)
\leq \frac{2D(\lambda)}{\lambda} \exp \left\{ - \frac{c}{3} \sum_{j=1}^T \eta^j \right\} + C_T \sum_{j=1}^T \exp \left\{ - \frac{c}{3} \sum_{j=1}^T j^\theta \right\}.
\] (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 \( 0 < \theta \leq 1 \), there holds
\[
\sum_{j=1}^t j^\theta \geq \left( \frac{1}{1 - \theta} \right) \left( (t + 1)^{1-\theta} - t^{1-\theta} \right), \quad \theta < 1,
\] (20)
\[
\log(T+1) - \log(t+1), \quad \theta = 1.
\]

Lemma 8. (see Lemma 5 of [14]) Let \( 0 < \nu \leq 1 \) and \( 0 < \theta \leq 1 \). Then
\[
\sum_{j=1}^T \frac{1}{t^{\nu}} \exp \left\{ - \nu \sum_{j=1}^T j^{1-\theta} \right\}
\]
is bounded by
\[
\frac{18}{\nu T^{\nu-1}} \exp \left\{ \frac{9T^{1-\theta}}{1-\theta} \left( (T+1)^{1-\theta} - T^{1-\theta} \right) \right\}, \quad \theta < 1,
\]
\[
\frac{8}{1-\nu} \left( T+1 \right)^{\nu-1}, \quad \theta = 1.
\]

Proof of Theorem 1. By (6) we know
\[
\left\| f_t \right\|_{L^2(x)} \leq k \left\| f_t \right\|_{k} \leq \frac{k^2 G(0)}{\lambda}.
\]

It follows \( \left\| y_i f_i(x) \right\| \leq \frac{k^2 G(0)}{\lambda} \) and for any \( G(t) \in \partial V(t) \) we have
\[
\left\| G(y_i f_i(x)) \right\| \leq L \left\| \frac{k^2 G(0)}{\lambda} \right\|.
\]

Therefore,
\[
E_{_{\eta_1,\eta_2,\cdots,\eta_T}} \left( \left\| H_t \right\|_2^2 \right) = E_{_{\eta_1,\eta_2,\cdots,\eta_T}} \left( \left\| G(y_i f_i(x)) \right\|^2 K(x_i, x) + 2 \lambda G(y_i f_i(x)) y_i f_i(x) + \lambda^2 f_i(x)^2 \right) \]
\[
= E_{_{\eta_1,\eta_2,\cdots,\eta_T}} \left( \left\| G(y_i f_i(x)) \right\|^2 K(x_i, x) + 2 \lambda G(y_i f_i(x)) y_i f_i(x) + \lambda^2 f_i(x)^2 \right) \leq C_A.
\] (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 \(f : X \rightarrow R\) is strongly convex on \(X\) if for any \(\lambda \in [0, 1)\) and \(x, x' \in X\) there holds

\[
f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda)f(x') - \frac{c\lambda(1-\lambda)}{2} \| x - x' \|^2,
\]

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

\[
\| f(x) - f(x') \| \leq L \| x - x' \|.
\]

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 \(l \in X\)

\[
\langle \xi, l \rangle \leq f^0(x + tl) - f(x).
\]

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

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

which equals that for all \(\xi \in \partial f(x)\) holds

\[
\langle \xi - \xi_i, x_i - x \rangle \geq c \| x_i - x \|^2.
\]

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

\[
f(x) - f(x') \leq \langle \xi, x - x' \rangle.
\]

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 \([x_i, x_j]\) for \(x_i, x_j \in X\), then, there exists \(u \in (x_i, x_j)\) such that

\[
f(x_i) - f(x_j) \leq \langle \partial f(u), x_i - x_j \rangle.
\]

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 \(\| \xi \| \leq L\).

ACKNOWLEDGMENT

The authors thank the referees and editors for valuable suggestions.

REFERENCES


© 2014 ACADEMY PUBLISHER
Baohuai Sheng attended Baoji Normal College, Baoji, Shaanxi, from 1981 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.