Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Unregularized online learning algorithms with general loss functions

Yiming Ying [a,*], Ding-Xuan Zhou [b]

[a] *Department of Mathematics and Statistics, State University of New York at Albany, Albany, NY, 12222, USA*
[b] *Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China*

ARTICLE INFO

Article history:
Received 13 November 2014
Received in revised form 12 May 2015
Accepted 15 August 2015
Available online xxxx
Communicated by Bin Han

Keywords:
Learning theory
Online learning
Reproducing kernel Hilbert space
Pairwise learning
Bipartite ranking

ABSTRACT

In this paper, we consider unregularized online learning algorithms in a Reproducing Kernel Hilbert Space (RKHS). Firstly, we derive explicit convergence rates of the unregularized online learning algorithms for classification associated with a general α-activating loss (see Definition 1 below). Our results extend and refine the results in [30] for the least square loss and the recent result [3] for the loss function with a Lipschitz-continuous gradient. Moreover, we establish a very general condition on the step sizes which guarantees the convergence of the last iterate of such algorithms. Secondly, we establish, for the first time, the convergence of the unregularized pairwise learning algorithm with a general loss function and derive explicit rates under the assumption of polynomially decaying step sizes. Concrete examples are used to illustrate our main results. The main techniques are tools from convex analysis, refined inequalities of Gaussian averages [5], and an induction approach.

1. Introduction

Let the input space $\mathcal{X}$ be a complete metric space and the output space $\mathcal{Y} = \{\pm 1\}$. In the standard framework of learning theory [10,11,23], one considers the problem of learning from a set of examples $\mathbf{z} = \{z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \ldots, T\}$ which are independently and identically distributed (i.i.d.) according to an unknown distribution ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

In the task of classification, a univariate loss function $\phi(yf(x))$ measures the error when $f(x)$ is used to predict the true label y. In this case, one aims to find a predictor in a hypothesis space to minimize the following true (generalization) error which is defined, for a function $g : \mathcal{X} \to \mathbb{R}$, by

$$\mathcal{E}(g) = \iint_{\mathcal{Z}} \phi(yg(x)) d\rho(x, y).$$

* Corresponding author.
E-mail address: yying@albany.edu (Y. Ying).

http://dx.doi.org/10.1016/j.acha.2015.08.007
1063-5203/

In contrast to the task of classification, pairwise learning problems involve a pairwise loss function $\phi((y-y')f(x,x'))$ for a hypothesis function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Notable examples of pairwise learning tasks include bipartite ranking [1,9,19], similarity and metric learning [6,26], AUC maximization [35], minimum error entropy principle [12–14], and gradient learning [16–18,31]. The aim of pairwise learning is to minimize the true error which is defined, for a pairwise function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, by

$$\widetilde{\mathcal{E}}(f) = \iint_{\mathcal{Z}\times\mathcal{Z}} \phi((y-y')f(x,x'))d\rho(x,y)d\rho(x',y').$$

In this paper, we consider online learning algorithms for both classification and pairwise learning tasks in a Reproducing Kernel Hilbert Space (RKHS). Specifically, let $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *Mercer kernel*, i.e. a continuous, symmetric and positive semi-definite kernel, see e.g. [10,23]. According to [2], the RKHS $\mathcal{H}_G$ associated with kernel G is defined to be the completion of the linear span of the set of functions $\{G_x(\cdot) := G(x,\cdot) : x \in \mathcal{X}\}$ with an inner product satisfying the reproducing property, i.e., for any $x', x \in \mathcal{X}$, $\langle G_x, G_{x'}\rangle_G = G(x,x')$. Similarly, for pairwise learning, we assume that the pairwise function $f : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is from an RKHS defined on the domain $\mathcal{X}^2 := \mathcal{X}\times\mathcal{X}$ with a (pairwise) kernel $K : \mathcal{X}^2\times\mathcal{X}^2 \to \mathbb{R}$. Throughout this paper, we consider a specific family of loss functions called α-activating loss defined as follows.

Definition 1. A function $\phi : \mathbb{R} \to \mathbb{R}^+$ is called an α-activating loss with some $\alpha \in (0,1]$ if it is convex and differentiable, $\phi'(0) < 0$, and $L := \sup_{\widetilde{s},s\in\mathbb{R}} |\phi'(\widetilde{s}) - \phi'(s)|/|\widetilde{s}-s|^\alpha < \infty$.

Our definition of α-activating loss follows [28] where the concept of the activating loss was first introduced. One can find in-depth discussions in [4,34] on loss functions for classification. Typical examples of α-activating losses includes q-norm loss [8,34] $\phi(s) = (1-s)_+^q = \max\{1-s,0\}^q$ for the support vector machine (SVM) classification with $1 < q \le 2$, the least square loss $\phi(s) = (1-s)^2$ and the logistic regression loss $\phi(s) = \log(1+e^{-s})$.

The first purpose of this paper is to study the unregularized online learning algorithm for classification associated with a general α-activating loss defined as follows.

Algorithm 1. Given the i.i.d. generated training data $\mathbf{z} = \{z_i = (x_i, y_i) : i = 1, 2, \ldots, T\}$, the unregularized online learning algorithm is given by $g_1 = 0$ and, for any $1 \le t \le T$,

$$g_{t+1} = g_t - \gamma_t \phi'(y_t g_t(x_t)) y_t G_{x_t}, \tag{1.1}$$

where $\{\gamma_t > 0 : t \in \mathbb{N}\}$ is usually referred to as the step size.

Online learning algorithms for classification or regression have drawn much attention [3,21,24,29,30,32]. Most of them focused on regularized online learning algorithms, i.e. $g_{t+1} = g_t - \gamma_t(\phi'(y_t g_t(x_t))y_t G_{x_t} + \lambda g_t)$. In particular, regularized online learning with a fixed $\lambda > 0$ was studied in [21] for the least square loss and in [32] for the general loss function, and in [24,29] for a time-varying regularization, i.e. $\lambda = \lambda(t) > 0$.

Instead, we focus on deriving explicit convergence rates of the unregularized online learning algorithms (i.e. $\lambda = 0$) with a general α-activating loss. Our results extend and refine those in [30] for the least square loss and the recent result [3, Theorem 4] for the loss function with a Lipschitz-continuous gradient. In contrast to the results [3,30] derived with the step sizes being chosen in a special form of $\mathcal{O}(t^{-\theta})$, we shall establish a very general condition on the step sizes which guarantees the convergence of the last iterate g_{T+1} of Algorithm 1. Moreover, in the contrast to the proof in [3], we will soon see below that our new proof here is much simpler and more powerful to handle general loss functions.

The second purpose of this paper is to study the convergence of the last iterate of the following online pairwise learning algorithm, which is associated with an α-activating loss function and the RKHS $\mathcal{H}_K$.

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Algorithm 2. Given the i.i.d. generated training data $\mathbf{z} = \{z_i = (x_i, y_i) : i = 1, 2, \ldots, T\}$, the unregularized online pairwise learning algorithm is given by $f_1 = f_2 = 0$ and, for any $2 \le t \le T$,

$$f_{t+1} = f_t - \frac{\gamma_t}{t-1} \sum_{j=1}^{t-1} \phi'((y_t - y_j) f_t(x_t, x_j))(y_t - y_j) K_{(x_t, x_j)}. \tag{1.2}$$

Online pairwise learning involves non-i.i.d.d. pairs of examples, which introduces more difficulty than the analysis in the univariate case. The research in this direction was recently conducted in [15,27,33]. In particular, in [15,27] the convergence of the average of the iterates (i.e. $\frac{1}{T}\sum_{t=2}^{T+1} f_t$) was established in the linear case by following online-to-batch conversion approach similar to those in the univariate case [7]. Recent work [33] focuses on Algorithm 2 with the least square loss. However, the analysis techniques there heavily depend on the nature of the least square loss (e.g. its derivative is a linear function) and do not apply to the general loss function.

In this paper, we establish, for the first time, the convergence of the last iterate of the unregularized pairwise learning algorithm (Algorithm 2) with a general loss function and derive explicit rates under the assumption of polynomially decaying step sizes. Concrete examples are used to illustrate our main results. The main techniques are tools from convex analysis and refined inequalities related to the Gaussian averages [5].

2. Main results

In this section, we present our main results related to Algorithms 1 and 2. The following theorem states a general convergence result for Algorithm 1.

Theorem 1. *Assume that ϕ is α-activating with some $0 < \alpha \le 1$ and let $\{g_t : t = 1, \ldots, T+1\}$ be given by Algorithm 1. If the step sizes satisfy that $\sum_{t=1}^{\infty} \gamma_t^{1+\alpha} < \infty$, then $\lim_{T\to\infty} \mathbb{E}\big[\mathcal{E}(g_{T+1})\big]$ exists. If, furthermore, $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ exists and $\sum_{t=1}^{\infty} \gamma_t = \infty$, then $\lim_{T\to\infty} \mathbb{E}\big[\mathcal{E}(g_{T+1})\big] = \inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$.*

By the above theorem, the step sizes can be chosen in the form of $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{1+\alpha}, 1)$, and $c > 0$. Indeed, we can further derive the explicit convergence rate for the last iterate of Algorithm 1.

Theorem 2. *Assume that ϕ is α-activating with some $0 < \alpha \le 1$ and $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ exists. Choose step sizes $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{1+\alpha}, 1)$ and $c > 0$. Then,*

$$\mathbb{E}\big[\mathcal{E}(g_{T+1}) - \mathcal{E}(g_{\mathcal{H}})\big] \le C_{\theta,\alpha,\mathcal{H}} T^{-\min(\frac{\alpha\theta}{2}, 1-\theta)},$$

where the constant $C_{\theta,\alpha,\mathcal{H}}$ depends on θ, α, c and $\|g_{\mathcal{H}}\|_G$ (see its explicit form in the proof).

From the above theorem, the maximal rate for α-activating losses is of the form $\mathcal{O}(T^{-\frac{\alpha}{\alpha+2}})$ which is achieved by choosing $\gamma_t = c\, t^{-\frac{2}{\alpha+2}}$. When $\alpha = 1$, the rate is of $\mathcal{O}(T^{-\frac{1}{3}})$ which is consistent with that in [3]. We can directly get the following examples from the above theorems, since $\phi(t) = (1-t)_+^q$ with $q \in (1, 2]$ is a $(q-1)$-activating loss and $\phi(t) = \log(1 + e^{-t})$ is a 1-activating loss.

Example 1. Let $\phi(t) = (1-t)_+^q$ with $1 < q \le 2$ and assume that $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ exists. Let $\{g_t : t = 1, \ldots, T+1\}$ be given by Algorithm 1 with step sizes $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{q}, 1)$, and $c > 0$. Then,

$$\mathbb{E}\big[\mathcal{E}(g_{T+1}) - \mathcal{E}(g_{\mathcal{H}})\big] = \mathcal{O}\big(T^{-\min(\frac{(q-1)\theta}{2}, 1-\theta)}\big).$$

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Example 2. Let $\phi(t) = \log(1+e^{-t})$ and assume that $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ exists. Let $\{g_t : t = 1, \ldots, T+1\}$ be given by Algorithm 1 with step sizes $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{2}, 1)$, and $c > 0$. Then,

$$\mathbb{E}\big[\mathcal{E}(g_{T+1}) - \mathcal{E}(g_{\mathcal{H}})\big] = \mathcal{O}\big(T^{-\min(\frac{\theta}{2}, 1-\theta)}\big).$$

Now we turn our attention to the convergence rates of Algorithm 2.

Theorem 3. *Assume ϕ is 1-activating, and $f_{\mathcal{H}} = \arg\inf_{f\in\mathcal{H}_K} \widetilde{\mathcal{E}}(f)$ exists. Let $\{f_t : t = 1, \ldots, T+1\}$ be given by Algorithm 2 with step sizes $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{2}, 1)$ and $0 < c \le \frac{1}{4\kappa^2 L}$. Then, for any $\delta \in (0, \min(\theta - \frac{1}{2}, 1-\theta))$, there holds*

$$\mathbb{E}\big[\widetilde{\mathcal{E}}(f_{T+1}) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})\big] \le \tilde{C}_{\theta,\delta,\mathcal{H}}\, T^{-\min(\frac{\theta}{2} - \frac{1}{4} - \frac{\delta}{2}, 1-\theta-\delta)},$$

where the constant $\tilde{C}_{\theta,\delta,\mathcal{H}}$ depends on θ, δ and $\|f_{\mathcal{H}}\|_G$ (see its explicit form in the proof).

If, moreover, the gradient of ϕ is uniformly bounded then the rate in the above theorem can further be improved.

Theorem 4. *Under the same assumptions of Theorem 3 and further assuming $|\phi'(s)| \le B < \infty$ for any $s \in \mathbb{R}$, then, for any $\delta \in (0, \min(\frac{\theta}{4}, 1-\theta))$, we have*

$$\mathbb{E}\big[\widetilde{\mathcal{E}}(f_{T+1}) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})\big] \le \bar{C}_{\theta,\delta,\mathcal{H}} T^{-\min(\frac{\theta}{4} - \frac{\delta}{2}, 1-\theta-\delta)},$$

where the constant $\bar{C}_{\theta,\delta,\mathcal{H}}$ depends on θ, δ and $\|f_{\mathcal{H}}\|_G$ (see its explicit form in the proof).

From the above theorem, we see that the maximal rate for Algorithm 2 associated with an α-activating loss is arbitrarily close to $\mathcal{O}(T^{-\frac{1}{6}})$. If, moreover, the gradient of the loss function ϕ is uniformly bounded then the maximal rate is improved to $\mathcal{O}(T^{-\frac{1}{5}})$. In particular, from the above theorem, we can immediately get the following examples since $\phi(t) = (1-t)_+^2$ and $\phi(t) = \log(1+e^{-t})$ are both 1-activating loss functions, and the gradient of $\phi(t) = \log(1+e^{-t})$ is uniformly bounded by one.

Example 3. Let $\phi(t) = (1-t)_+^2$ with $1 < q \le 2$ and assume that $f_{\mathcal{H}} = \arg\inf_{f\in\mathcal{H}_K} \widetilde{\mathcal{E}}(f)$ exists. Let $\{g_t : t = 1, \ldots, T+1\}$ be given by Algorithm 2 with step sizes $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{2}, 1)$ and $c > 0$. Then, for any $\delta \in (0, \min(\theta - \frac{1}{2}, 1-\theta))$, there holds

$$\mathbb{E}\big[\widetilde{\mathcal{E}}(f_{T+1}) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})\big] = \mathcal{O}\Big(T^{-\min(\frac{\theta}{2} - \frac{1}{4} - \frac{\delta}{2}, 1-\theta-\delta)}\Big).$$

Example 4. Let $\phi(t) = \log(1+e^{-t})$ and assume that $f_{\mathcal{H}} = \arg\inf_{f\in\mathcal{H}_K} \widetilde{\mathcal{E}}(f)$ exists. Let $\{g_t : t = 1, \ldots, T+1\}$ be given by Algorithm 2 with step sizes $\gamma_t = c\, t^{-\theta}$ with some $\theta \in (\frac{1}{2}, 1)$, and $c > 0$. Then, for any $\delta \in (0, \min(\frac{\theta}{4}, 1-\theta))$,

$$\mathbb{E}\big[\widetilde{\mathcal{E}}(f_{T+1}) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})\big] = \mathcal{O}\Big(T^{-\min(\frac{\theta}{4} - \frac{\delta}{2}, 1-\theta-\delta)}\Big).$$

3. Proofs of main results

We derive some useful properties of the α-activating loss function ϕ, which play critical roles in proving main theorems. Some of them may be of interest in their own rights.

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Proposition 1. *Assume that $\phi : \mathbb{R} \to \mathbb{R}$ is convex and its gradient is α-Hölder continuous, i.e. $L := \sup_{\widetilde{s},s\in\mathbb{R}} |\phi'(\widetilde{s}) - \phi'(s)|/|\widetilde{s} - s|^\alpha < \infty$. Then, for any $s, \widetilde{s} \in \mathbb{R}$, the following properties hold true.*

(a) $\phi(s) - \phi(\widetilde{s}) - \phi'(\widetilde{s})(s - \widetilde{s}) \le \frac{L}{1+\alpha}|s - \widetilde{s}|^{1+\alpha}$.

(b) $\phi(\widetilde{s}) \ge \phi(s) + \phi'(s)(\widetilde{s} - s) + \frac{\alpha L^{-\frac{1}{\alpha}}}{1+\alpha}|\phi'(s) - \phi'(\widetilde{s})|^{\frac{1+\alpha}{\alpha}}$.

(c) $(\phi'(s) - \phi'(\widetilde{s}))(s - \widetilde{s}) \ge \frac{2\alpha L^{-\frac{1}{\alpha}}}{1+\alpha}|\phi'(s) - \phi'(\widetilde{s})|^{\frac{1+\alpha}{\alpha}}$.

(d) *If, moreover, $\phi(s) \ge 0$ for any $s \in \mathbb{R}$, then $|\phi'(s)|^{\frac{1+\alpha}{\alpha}} \le \frac{(1+\alpha)^{1+\frac{1}{\alpha}}}{\alpha} L^{\frac{1}{\alpha}}\, \phi(s)$.*

Proof. Part (a) directly follows from the fact that the assumption that $|\phi'(s) - \phi'(\widetilde{s})| \le L|s - \widetilde{s}|^\alpha$ and the fact

$$\phi(s) - \phi(\widetilde{s}) - \phi'(\widetilde{s})(s - \widetilde{s}) = \int_0^1 (\phi'(\theta s + (1-\theta)\widetilde{s}) - \phi'(\widetilde{s}))(s - \widetilde{s})d\theta.$$

For part (b), let $\psi_s(\widetilde{s}) = \phi(\widetilde{s}) - \phi'(s)\widetilde{s}$. Notice that $\psi_s(\cdot)$ is convex, differentiable and its gradient $\psi_s'(\widetilde{s}) = \phi'(\widetilde{s}) - \phi'(s)$ is α-Hölder continuous. In addition, $\psi_s(\cdot)$ achieves the minimum at s since $\psi_s'(s) = 0$. Hence, for $\delta = L^{\frac{1}{\alpha}}$,

$$\begin{aligned}
\psi_s(s) &\le \psi_s\big(\widetilde{s} - \tfrac{1}{\delta}(\phi'(\widetilde{s}) - \phi'(s))|\phi'(\widetilde{s}) - \phi'(s)|^{\frac{1-\alpha}{\alpha}}\big) \\
&\le \psi_s(\widetilde{s}) + \psi_s'(\widetilde{s})\big(-\tfrac{1}{\delta}(\phi'(\widetilde{s}) - \phi'(s))|\phi'(\widetilde{s}) - \phi'(s)|^{\frac{1-\alpha}{\alpha}}\big) \\
&\quad + \tfrac{L}{1+\alpha}\big|\tfrac{1}{\delta}(\phi'(\widetilde{s}) - \phi'(s))|\phi'(\widetilde{s}) - \phi'(s)|^{\frac{1-\alpha}{\alpha}}\big|^{1+\alpha} \\
&= \psi_s(\widetilde{s}) - \tfrac{\alpha L^{-\frac{1}{\alpha}}}{1+\alpha}|\phi'(s) - \phi'(\widetilde{s})|^{\frac{1+\alpha}{\alpha}},
\end{aligned}$$

where the second to last inequality used the fact that $\psi_s(\cdot)$ satisfies part (a). By the definition of $\psi_s(\cdot)$, re-arranging the terms in the above estimation yields the desired result of part (b).

For part (c), switching the roles of $\widetilde{s}, s$ in part (b) yields that

$$\phi(s) \ge \phi(\widetilde{s}) + \phi'(\widetilde{s})(s - \widetilde{s}) + \frac{\alpha L^{-\frac{1}{\alpha}}}{1+\alpha}|\phi'(s) - \phi'(\widetilde{s})|^{\frac{1+\alpha}{\alpha}}.$$

Adding part (b) and the above inequality implies part (c).

For part (d), the case for $\alpha = 1$ was proved in [22]. We generalize their proof to the general case $0 < \alpha \le 1$. Indeed, we only need to prove the case $\phi'(s) \ne 0$. For any $s \in \mathbb{R}$, let $r = s - ((1+\alpha)L)^{-\frac{1}{\alpha}}|\phi'(s)|^{\frac{1}{\alpha}}\frac{\phi'(s)}{|\phi'(s)|}$. By the mean-value theorem, there exists ξ in the range (s, r) (if $\phi'(s) < 0$) or (r, s) (if $\phi'(s) > 0$) such that $\phi(r) = \phi(s) + \phi'(\xi)(r - s)$. Hence,

$$\begin{aligned}
0 \le \phi(r) &= \phi(s) + \phi'(s)(r - s) + (\phi'(\xi) - \phi'(s))(r - s) \\
&\le \phi(s) + \phi'(s)(r - s) + L|r - s||\xi - s|^\alpha \\
&\le \phi(s) + \phi'(s)(r - s) + L|r - s|^{1+\alpha} = \phi(s) - \tfrac{\alpha}{(1+\alpha)^{1+\frac{1}{\alpha}}} L^{-\frac{1}{\alpha}}|\phi'(s)|^{\frac{1+\alpha}{\alpha}},
\end{aligned}$$

which completes the proof of part (d). □

3.1. Proofs for the convergence of Algorithm 1

The main idea for proving the convergence of Algorithm 1 is to derive a recursive inequality for the sequence $\{R_t := \mathbb{E}[\mathcal{E}(g_t) - \mathcal{E}(g_{\mathcal{H}})] : 1 \le t \le T+1\}$ (i.e. the relationship between R_{t+1} and R_t), and then

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

apply induction based on this inequality. To this end, we need to establish the boundedness of the learning sequence $\{g_t : t = 1, 2, \ldots, T+1\}$ generated by Algorithm 1. Throughout the paper, we use the conventional notion that $\sum_{j=k}^{t} \gamma_j^{1+\alpha} = 0$ whenever $t < k$. Denote $\kappa = \sup_{x\in\mathcal{X}} \sqrt{G(x,x)}$.

Lemma 1. *Let $\{g_t : t = 1, \ldots, T+1\}$ be generated by Algorithm 1. Then,*

$$\mathbb{E}\big[\mathcal{E}(g_{t+1})\big] \le (1 + \mathcal{E}(g_1)) \exp\big(A_\alpha \sum_{j=1}^{t} \gamma_j^{1+\alpha}\big),$$

where $A_\alpha = L^2(1 + \frac{1}{\alpha})^\alpha \kappa^{2(1+\alpha)}$.

Proof. Since ϕ is convex and ϕ' is of α-Hölder continuous, by part (a) and part (d) of Proposition 1 we have

$$\begin{aligned}
\phi(yg_{t+1}(x)) &\le \phi(yg_t(x)) + \phi'(yg_t(x))y(g_{t+1}(x) - g_t(x)) + \tfrac{L}{1+\alpha}|g_{t+1}(x) - g_t(x)|^{1+\alpha} \\
&= \phi(yg_t(x)) - \gamma_t \langle \phi'(yg_t(x))yG_x, \phi'(y_t g_t(x_t))y_t G_{x_t}\rangle_G + \tfrac{L}{1+\alpha}|g_{t+1}(x) - g_t(x)|^{1+\alpha} \\
&\le \phi(yg_t(x)) - \gamma_t \langle \phi'(yg_t(x))yG_x, \phi'(y_t g_t(x_t))y_t G_{x_t}\rangle_G + \tfrac{L\kappa^{2(1+\alpha)}\gamma_t^{1+\alpha}}{1+\alpha}|\phi'(y_t g_t(x_t))|^{1+\alpha} \\
&\le \phi(yg_t(x)) - \gamma_t \langle \phi'(yg_t(x))yG_x, \phi'(y_t g_t(x_t))y_t G_{x_t}\rangle_G + A_\alpha \gamma_t^{1+\alpha}|\phi(y_t g_t(x_t))|^\alpha.
\end{aligned}$$

Taking expectation of both sides of the above inequality with respect to $z = (x, y)$ and samples $\{z_1, \ldots, z_t\}$, and noting that g_t only depends on $\{z_1, \ldots, z_{t-1}\}$, we have

$$\begin{aligned}
\mathbb{E}\big[\mathcal{E}(g_{t+1})\big] &\le \mathbb{E}\big[\mathcal{E}(g_t)\big] - \gamma_t \mathbb{E}\Big[\| \textstyle\int_{\mathcal{Z}} \phi'(yg_t(x))yG_x d\rho(x,y)\|_G^2\Big] \\
&\quad + A_\alpha \gamma_t^{1+\alpha} \mathbb{E}\Big[\textstyle\int_{\mathcal{Z}} |\phi(yg_t(x))|^\alpha d\rho(x,y)\Big] \\
&\le \mathbb{E}\big[\mathcal{E}(g_t)\big] - \gamma_t \mathbb{E}\Big[\| \textstyle\int_{\mathcal{Z}} \phi'(yg_t(x))yG_x d\rho(x,y)\|_G^2\Big] \\
&\quad + A_\alpha \gamma_t^{1+\alpha} \Big(\mathbb{E}[\textstyle\int_{\mathcal{Z}} \phi(yg_t(x)) d\rho(x,y)]\Big)^\alpha \\
&= \mathbb{E}\big[\mathcal{E}(g_t)\big] - \gamma_t \mathbb{E}\Big[\| \textstyle\int_{\mathcal{Z}} \phi'(yg_t(x))yG_x d\rho(x,y)\|_G^2\Big] + A_\alpha \gamma_t^{1+\alpha} \big(\mathbb{E}[\mathcal{E}(g_t)]\big)^\alpha \\
&\le (1 + A_\alpha \gamma_t^{1+\alpha}) \mathbb{E}\big[\mathcal{E}(g_t)\big] - \gamma_t \mathbb{E}\Big[\| \textstyle\int_{\mathcal{Z}} \phi'(yg_t(x))yG_x d\rho(x,y)\|_G^2\Big] + A_\alpha \gamma_t^{1+\alpha}. \qquad (3.1)
\end{aligned}$$

Consequently,

$$\mathbb{E}\big[\mathcal{E}(g_{t+1})\big] \le (1 + A_\alpha \gamma_t^{1+\alpha}) \mathbb{E}\big[\mathcal{E}(g_t)\big] + A_\alpha \gamma_t^{1+\alpha}.$$

The above inequality implies that

$$\begin{aligned}
\mathbb{E}\big[\mathcal{E}(g_{t+1})\big] &\le \textstyle\prod_{j=1}^{t}(1 + A_\alpha \gamma_j^{1+\alpha}) \mathcal{E}(g_1) + A_\alpha \sum_{j=1}^{t} \prod_{k=j+1}^{t}(1 + A_\alpha \gamma_k^{1+\alpha}) \gamma_j^{1+\alpha} \\
&\le \textstyle\prod_{j=1}^{t}(1 + A_\alpha \gamma_j^{1+\alpha}) \mathcal{E}(g_1) + \sum_{j=1}^{t} \big[\prod_{k=j}^{t}(1 + A_\alpha \gamma_k^{1+\alpha}) - \prod_{k=j+1}^{t}(1 + A_\alpha \gamma_k^{1+\alpha})\big] \\
&= \textstyle\prod_{j=1}^{t}(1 + A_\alpha \gamma_j^{1+\alpha}) \mathcal{E}(g_1) + \big[\prod_{k=1}^{t}(1 + A_\alpha \gamma_k^{1+\alpha}) - 1\big] \\
&\le (1 + \mathcal{E}(g_1)) \exp\big(A_\alpha \textstyle\sum_{j=1}^{t} \gamma_j^{1+\alpha}\big).
\end{aligned}$$

This completes the proof of the lemma. □

From the above lemma, we know that if $\sum_{j=1}^{\infty} \gamma_j^{1+\alpha} < \infty$ then, for any $t \in \mathbb{N}$, there holds

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(g_{t+1})] &\le (1 + \mathcal{E}(g_1)) \exp\Big(A_\alpha \textstyle\sum_{j=1}^{t} \gamma_j^{1+\alpha}\Big) \\
&\le D_\infty := (1 + \mathcal{E}(g_1)) \exp\Big(A_\alpha \textstyle\sum_{j=1}^{\infty} \gamma_j^{1+\alpha}\Big) < \infty. \qquad (3.2)
\end{aligned}$$

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

One typical example of step sizes is of the form $\gamma_t = \frac{c}{t^\theta}$ with some $\theta \in (\frac{1}{1+\alpha}, 1)$. In this case, notice that

$$\begin{aligned}\sum_{j=2}^t \gamma_j^{1+\alpha} &= c^{1+\alpha} \sum_{j=1}^t j^{-\theta(1+\alpha)} = c^{1+\alpha}(1 + \sum_{j=2}^t j^{-\theta(1+\alpha)}) \\ &\le c^{1+\alpha}(1 + \int_1^t s^{-\theta(1+\alpha)} ds \le \tfrac{c^{1+\alpha}\theta(1+\alpha)}{\theta(1+\alpha)-1} \le \tfrac{2c^{1+\alpha}}{\theta(1+\alpha)-1}. \end{aligned} \tag{3.3}$$

Hence, for any $t \in \mathbb{N}$,

$$\mathbb{E}\big[\mathcal{E}(g_t)\big] \le D_\infty \le \big(1 + \mathcal{E}(g_1)\big) \exp\Big(\frac{2A_\alpha c^{1+\alpha}}{\theta(1+\alpha)-1}\Big). \tag{3.4}$$

We now turn our attention to estimating the boundedness of $\mathbb{E}\big[\|g_t - g_\mathcal{H}\|_G^2\big]$.

Lemma 2. *Assume that $g_\mathcal{H} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ exists and let the learning sequence $\{g_t : t = 1, \dots, T+1\}$ be generated by Algorithm 1. If $\sum_{j=1}^\infty \gamma_j^{1+\alpha} < \infty$, then*

$$\mathbb{E}\big[\|g_{t+1} - g_\mathcal{H}\|_G^2\big] \le \|g_\mathcal{H}\|_G^2 + B_\alpha D_\infty^{\frac{2\alpha}{1+\alpha}} \sum_{j=1}^t \gamma_j^2,$$

where $B_\alpha := \kappa^2(1+\alpha)^2 L^{\frac{2}{1+\alpha}} \alpha^{-\frac{2\alpha}{1+\alpha}}$.

Proof. Notice that, since $g_\mathcal{H} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$,

$$\int \phi'(yg_\mathcal{H}(x)) y G_x d\rho(x,y) = 0.$$

By the definition of g_{t+1} in Algorithm 1, $\mathbb{E}[\|g_{t+1} - g_\mathcal{H}\|_G^2]$ is therefore bounded by

$$\begin{aligned}&\mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] - 2\gamma_t \mathbb{E}[\langle \phi'(y_t g_t(x_t)) y_t G_{x_t}, g_t - g_\mathcal{H}\rangle_G] + \gamma_t^2 \mathbb{E}\big[\|\phi'(y_t g_t(x_t)) G_{x_t}\|_G^2\big] \\ &\le \mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] - 2\gamma_t \mathbb{E}[\langle \phi'(y_t g_t(x_t)) y_t G_{x_t}, g_t - g_\mathcal{H}\rangle_G] + \gamma_t^2 \kappa^2 \mathbb{E}\big[|\phi'(y_t g_t(x_t))|^2\big] \\ &= \mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] - 2\gamma_t \mathbb{E}[\langle \textstyle\int [\phi'(yg_t(x)) y G_x - \phi'(yg_\mathcal{H}(x)) y G_x] d\rho(x,y), g_t - g_\mathcal{H}\rangle_G] \\ &\quad + \gamma_t^2 \kappa^2 \mathbb{E}\big[|\phi'(y_t g_t(x_t))|^2\big] \\ &\le \mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] + \gamma_t^2 \kappa^2 \mathbb{E}\big[|\phi'(y_t g_t(x_t))|^2\big] \\ &\le \mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] + \gamma_t^2 \kappa^2 \big(\mathbb{E}[|\phi'(y_t g_t(x_t))|^{\frac{1+\alpha}{\alpha}}]\big)^{\frac{2\alpha}{1+\alpha}}, \end{aligned} \tag{3.5}$$

where the second to last inequality used the fact, by part (c) of Proposition 1,

$$\begin{aligned}&\langle \textstyle\int [\phi'(yg_t(x)) y G_x - \phi'(yg_\mathcal{H}(x)) y G_x] d\rho(x,y), g_t - g_\mathcal{H}\rangle_G \\ &= \textstyle\int [\phi'(yg_t(x)) - \phi'(yg_\mathcal{H}(x))] y (g_t(x) - g_\mathcal{H}(x)) d\rho(x,y) \ge 0. \end{aligned}$$

Also, by part (d) of Proposition 1, we have $|\phi'(y_t g_t(x_t))|^{\frac{1+\alpha}{\alpha}} \le \frac{(1+\alpha)^{1+\frac{1}{\alpha}}}{\alpha} L^{\frac{1}{\alpha}} \phi(y_t g_t(x_t))$. Putting this back into (3.5), we know from (3.2) that

$$\begin{aligned}\mathbb{E}[\|g_{t+1} - g_\mathcal{H}\|_G^2] &\le \mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] + \gamma_t^2 \kappa^2 \tfrac{(1+\alpha)^2 L^{\frac{2}{1+\alpha}}}{\alpha^{\frac{2\alpha}{1+\alpha}}} \big[\mathbb{E}(\mathcal{E}(g_t))\big]^{\frac{2\alpha}{1+\alpha}} \\ &\le \mathbb{E}[\|g_t - g_\mathcal{H}\|_G^2] + \gamma_t^2 \kappa^2 \tfrac{(1+\alpha)^2 L^{\frac{2}{1+\alpha}}}{\alpha^{\frac{2\alpha}{1+\alpha}}} \big(D_\infty\big)^{\frac{2\alpha}{1+\alpha}}, \end{aligned}$$

which directly yields the desired result. This completes the proof of the lemma. □

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Denote $\bar{D}_\infty = \|g_{\mathcal{H}}\|_G^2 + B_\alpha D_\infty^{\frac{2\alpha}{1+\alpha}} \sum_{j=1}^\infty \gamma_j^2$. Then, if the step sizes are in the form of $\gamma_t = \frac{c}{t^\theta}$ with $\theta \in (\frac{1}{1+\alpha}, 1)$, then, by (3.4),

$$\begin{aligned}\mathbb{E}\big[\|g_t - g_{\mathcal{H}}\|_G^2\big] &\le \|g_{\mathcal{H}}\|_G^2 + c^2 B_\alpha D_\infty^{\frac{2\alpha}{1+\alpha}} \sum_{j=1}^{t-1} j^{-2\theta} \\ &\le \bar{D}_\infty \le \|g_{\mathcal{H}}\|_G^2 + \tfrac{2\theta c^2 B_\alpha D_\infty^{\frac{2\alpha}{1+\alpha}}}{2\theta - 1}. \end{aligned} \tag{3.6}$$

We are now in a position to prove the main theorems for Algorithm 1.

Proof of Theorem 1. By (3.1) and (3.2), we have

$$\mathbb{E}\big[\mathcal{E}(g_{t+1})\big] \le \mathbb{E}\big[\mathcal{E}(g_t)\big] - \gamma_t \mathbb{E}\Big[\big\| \int_{\mathcal{Z}} \phi'(y g_t(x)) y G_x d\rho(x,y)\big\|_G^2\Big] + A_\alpha (1 + D_\infty)\gamma_t^{1+\alpha}. \tag{3.7}$$

The above inequality implies that

$$\mathbb{E}\big[\mathcal{E}(g_{t+1})\big] \le \mathbb{E}\big[\mathcal{E}(g_t)\big] + A_\alpha (1 + D_\infty)\gamma_t^{1+\alpha}.$$

Consequently, for any fixed $t \le T$,

$$\mathbb{E}\big[\mathcal{E}(g_{T+1})\big] \le \mathbb{E}\big[\mathcal{E}(g_t)\big] + A_\alpha (1 + D_\infty) \sum_{j=t}^\infty \gamma_t^{1+\alpha}.$$

This means that $\overline{\lim}_{T\to\infty} \mathbb{E}\big[\mathcal{E}(g_{T+1})\big] \le \mathbb{E}\big[\mathcal{E}(g_t)\big] + A_\alpha(1 + D_\infty) \sum_{j=t}^\infty \gamma_t^{1+\alpha}$, which also implies, since $\sum_{j=1}^\infty \gamma_t^{1+\alpha} < \infty$, that

$$\overline{\lim}_{T\to\infty} \mathbb{E}\big[\mathcal{E}(g_{T+1})\big] \le \underline{\lim}_{t\to\infty} \Big\{\mathbb{E}\big[\mathcal{E}(g_t)\big] + A_\alpha(1 + D_\infty) \sum_{j=t}^\infty \gamma_t^{1+\alpha}\Big\} = \underline{\lim}_{t\to\infty} \mathbb{E}\big[\mathcal{E}(g_t)\big].$$

Hence, $\varepsilon := \lim_{t\to\infty} \mathbb{E}\big[\mathcal{E}(g_t)\big]$ exists and, apparently, $\inf_{g\in\mathcal{H}_G} \mathcal{E}(g) \le \varepsilon \le D_\infty < \infty$ where the last inequality follows from equation (3.2). This completes the proof for the first part of the theorem.

Now it remains to prove, if we further assume that $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ exists and $\sum_{j=1}^\infty \gamma_j = \infty$, that $\varepsilon = \inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$. Let us assume, on the contrary, that $\varepsilon_1 = \varepsilon - \inf_{g\in\mathcal{H}_G} \mathcal{E}(g) > 0$. Let $R_t := \mathbb{E}\big[\mathcal{E}(g_t)\big] - \inf_{g\in\mathcal{H}_G} \mathcal{E}(g)$ for any $t \in \mathbb{N}$. In this case, there exists t_1 such that, for any $t \ge t_1$, $R_t \ge \frac{\varepsilon_1}{2}$. However, from (3.7), we know that

$$R_{t+1} \le R_t - \gamma_t \mathbb{E}\Big[\big\| \int_{\mathcal{Z}} \phi'(y g_t(x)) y G_x d\rho(x,y)\big\|_G^2\Big] + A_\alpha (1 + D_\infty)\gamma_t^{1+\alpha}. \tag{3.8}$$

By the convexity of ϕ,

$$\begin{aligned}\mathcal{E}(g_t) - \mathcal{E}(g_{\mathcal{H}}) &\le \textstyle\int_{\mathcal{Z}} \phi'(y g_t(x)) y (g_t(x) - g_{\mathcal{H}}(x)) d\rho(x,y) \\ &= \big\langle \textstyle\int_{\mathcal{Z}} \phi'(y g_t(x)) y G_x d\rho(x,y), g_t - g_{\mathcal{H}}\big\rangle_G \\ &\le \big[\| \textstyle\int_{\mathcal{Z}} \phi'(y g_t(x)) y G_x d\rho(x,y)\|^2\big]^{\frac{1}{2}} \|g_t - g_{\mathcal{H}}\|_G. \end{aligned}$$

Also, observe that $\bar{D}_\infty = \|g_{\mathcal{H}}\|_G^2 + B_\alpha D_\infty^{\frac{2\alpha}{1+\alpha}} \sum_{j=1}^\infty \gamma_j^2 < \infty$, since $\sum_{j=1}^\infty \gamma_j^{1+\alpha} < \infty$ and $\alpha \le 1$. This implies that

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\mathbb{E}\Big[\Big\|\int_{\mathcal{Z}}\phi'(yg_t(x))yG_x d\rho(x,y)\Big\|_G^2\Big] \geq \frac{R_t^2}{\mathbb{E}\big[\|g_t-g_{\mathcal{H}}\|_G^2\big]} \geq \frac{R_t^2}{\bar{D}_\infty}.$$

Putting this back into (3.8) yields that

$$R_{t+1} \leq R_t - \gamma_t R_t^2/\bar{D}_\infty + A_\alpha(1+D_\infty)\gamma_t^{1+\alpha}. \tag{3.9}$$

This means that

$$\begin{aligned}\overline{\lim}_{T\to\infty}\textstyle\sum_{t=1}^T \gamma_t R_t^2/\bar{D}_\infty &\leq R_1 + A_\alpha(1+D_\infty)\textstyle\sum_{t=1}^T \gamma_t^{1+\alpha}\\ &\leq R_1 + A_\alpha(1+D_\infty)\textstyle\sum_{t=1}^\infty \gamma_t^{1+\alpha} < \infty.\end{aligned}$$

However, $\sum_{t=1}^T \gamma_t R_t^2/\bar{D}_\infty \geq \frac{\varepsilon_1^2}{4\bar{D}_\infty}\sum_{t=t_1}^T \gamma_t$, which implies, by the assumption that $\sum_{t=1}^\infty \gamma_t = \infty$, that

$$\overline{\lim}_{T\to\infty}\sum_{t=1}^T \gamma_t R_t^2/\bar{D}_\infty \geq \frac{\varepsilon_1^2}{4\bar{D}_\infty}\sum_{t=t_1}^\infty \gamma_t = \infty.$$

This leads to a contradiction. Hence, $\varepsilon_1 = \lim_{t\to\infty} R_t = 0$. This completes the proof the theorem. □

We now turn our attention to proving Theorem 2 by an induction based on the recursive inequality (3.9).

Proof of Theorem 2. We prove the theorem from the recursive inequality (3.9). Since $\gamma_t = \frac{c}{t^\theta}$ with some $\theta \in (\frac{1}{1+\alpha}, 1)$, inequalities (3.4) and (3.6) hold true. Let $\beta = \min(\frac{\alpha\theta}{2}, 1-\theta)$, and choose

$$D = \max\Big\{D_\infty, \big(\frac{2c}{\bar{D}_\infty}\big)^{\min(\frac{\alpha}{2},\frac{1-\theta}{\theta})}\big(2^\beta D_\infty\big)^{\min(1+\frac{\alpha}{2},\frac{1}{\theta})}, \frac{\bar{D}_\infty}{c} + \sqrt{A_\alpha(1+D_\infty)c^\alpha\bar{D}_\infty}\Big\}.$$

Denote

$$t_0 = \left\lfloor 2\big(\frac{2cD}{\bar{D}_\infty}\big)^{\frac{1}{\theta+\beta}}\right\rfloor.$$

By the definition of D and β, we know that $D \geq \frac{\bar{D}_\infty}{c}$ and $0 < \theta+\beta \leq 1$ which further implies that $t_0 \geq 4$. Since

$$D \geq \max\Big\{D_\infty, \big(\frac{2c}{\bar{D}_\infty}\big)^{\min(\frac{\alpha}{2},\frac{1-\theta}{\theta})}\big(2^\beta D_\infty\big)^{\min(1+\frac{\alpha}{2},\frac{1}{\theta})}\Big\},$$

we have

$$\mathbb{E}[\mathcal{E}(g_t)-\mathcal{E}(g_{\mathcal{H}})] \leq D_\infty \leq \frac{D}{t_0^\beta} \leq \frac{D}{t^\beta}, \quad \forall t \leq t_0.$$

Now we assume that $R_t \leq \frac{D}{t^\beta}$ for some $t \in \mathbb{N}$ and $t \geq t_0$ and we are going to prove that $R_{t+1} \leq \frac{D}{(t+1)^\beta}$ by induction.

To this end, let $F(x) := x - \gamma_t x^2/\bar{D}_\infty$ and notice that F is increasing when $x \in (0, \frac{\bar{D}_\infty t^\theta}{2c}]$. Observe that $t \geq t_0 \geq \big(\frac{2cD}{\bar{D}_\infty}\big)^{\frac{1}{\theta+\beta}}$ which implies that $\frac{D}{t^\beta} \in (0, \frac{\bar{D}_\infty t^\theta}{2c})$. Combining this with (3.9) and the induction assumption $R_t \leq \frac{D}{t^\beta}$ (i.e. $R_t \in (0, \frac{\bar{D}_\infty t^\theta}{2c})$), we have

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\begin{aligned} R_{t+1} &\le F(R_t) + A_\alpha(1+D_\infty)\gamma_t^{1+\alpha} \le F\big(\tfrac{D}{t^\beta}\big) + A_\alpha(1+D_\infty)\gamma_t^{1+\alpha} \\ &\le \tfrac{D}{t^\beta}\Big[1 - \big(\tfrac{cD}{D_\infty} - \tfrac{A_\alpha(1+D_\infty)c^{1+\alpha}}{D} t^{2\beta-\theta\alpha}\big)t^{-\theta-\beta}\Big] \\ &\le \tfrac{D}{t^\beta}\Big[1 - \big(\tfrac{cD}{D_\infty} - \tfrac{A_\alpha(1+D_\infty)c^{1+\alpha}}{D}\big)t^{-\theta-\beta}\Big], \end{aligned} \tag{3.10}$$

where the last inequality used that fact $2\beta-\theta\alpha \le 0$. By the definition of D, $D \ge \frac{\bar{D}_\infty}{c} + \sqrt{A_\alpha(1+D_\infty)c^\alpha \bar{D}_\infty}$ which implies that $\frac{cD}{D_\infty} - \frac{A_\alpha(1+D_\infty)c^{1+\alpha}}{D} \ge 1$. Putting this back into (3.10) yields that

$$\begin{aligned} R_{t+1} &\le \frac{D}{t^\beta}\Big[1 - t^{-\theta-\beta}\Big] \le \frac{D}{t^\beta}\Big[1 - t^{-1}\Big] \\ &= \frac{D}{t^\beta}\big(\frac{t-1}{t}\big) \le \frac{D}{t^\beta}\big(\frac{t}{t+1}\big)^\beta = \frac{D}{(t+1)^\beta}, \end{aligned}$$

where the second inequality used the fact that $\theta + \beta \le 1$. This completes the proof of the theorem. □

3.2. Proofs for the convergence of Algorithm 2

In this subsection, we prove the main theorems related to Algorithm 2. The main idea is to derive a recursive inequality on the sequence $\{R_t := \mathbb{E}[\widetilde{\mathcal{E}}(f_t) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})] : 1 \le t \le T+1\}$ (i.e. the relationship between R_{t+1} and R_t), and then conduct a smart induction based on this inequality. To do this, let us establish some useful lemmas. Denote $\widetilde{\kappa} = \sup_{x,\bar{x}\in\mathcal{X}\times\mathcal{X}} \sqrt{K((x,\bar{x}),(x,\bar{x}))}$.

Lemma 3. *Assume ϕ is 1-activating and $f_{\mathcal{H}} = \arg\inf_{f\in\mathcal{H}_K} \widetilde{\mathcal{E}}(f)$ exists. Let $\{f_t : t = 1,\dots,T+1\}$ be generated by Algorithm 2. Then*

$$\mathbb{E}\big[\|f_{t+1} - f_{\mathcal{H}}\|_K^2\big] \le \Big[\|f_{\mathcal{H}}\|_K^2 + 2\sigma_{\mathcal{H}}^2(3+\ln t)\Big]\exp\big((2+32\widetilde{\kappa}^4L^2)\sum_{j=2}^{t}\gamma_j^2\big),$$

where $\sigma_{\mathcal{H}}^2 = \int_{\mathcal{Z}}\int_{\mathcal{Z}} \big\|\phi'((y-\widetilde{y})f_{\mathcal{H}}(x,\widetilde{x}))K_{(x,\widetilde{x})}\big\|_K^2 d\rho(x,y)d\rho(\widetilde{x},\widetilde{y})$.

Proof. $\mathbb{E}\big[\|f_{t+1} - f_{\mathcal{H}}\|_K^2\big]$ is bounded by

$$\begin{aligned} &\mathbb{E}[\|f_t - f_{\mathcal{H}}\|_K^2] + \tfrac{\gamma_t^2}{(t-1)^2}\mathbb{E}\big[\big\|\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_t(x_t,x_j))(y_t-y_j)K_{(x_t,x_j)}\big\|_K^2\big] \\ &- \tfrac{2\gamma_t}{t-1}\mathbb{E}\big[\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_t(x_t,x_j))(y_t-y_j)(f_t(x_t,x_j) - f_{\mathcal{H}}(x_t,x_j))\big]. \end{aligned} \tag{3.11}$$

Noting that $\int_{\mathcal{Z}}\int_{\mathcal{Z}} \phi'((y-\widetilde{y})f_{\mathcal{H}}(x,\widetilde{x}))K_{(x,\widetilde{x})}d\rho(x,y)d\rho(\widetilde{x},\widetilde{y}) = 0$, we have

$$
\begin{aligned}
&-\mathbb{E}[\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_t(x_t,x_j))(y_t-y_j)(f_t(x_t,x_j)-f_{\mathcal{H}}(x_t,x_j))]\\
&=-\mathbb{E}[\sum_{j=1}^{t-1}[\phi'((y_t-y_j)f_t(x_t,x_j))-\phi'((y_t-y_j)f_{\mathcal{H}}(x_t,x_j))](y_t-y_j)(f_t(x_t,x_j)-f_{\mathcal{H}}(x_t,x_j))]\\
&\quad-\mathbb{E}[\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_{\mathcal{H}}(x_t,x_j))(y_t-y_j)(f_t(x_t,x_j)-f_{\mathcal{H}}(x_t,x_j))]\\
&\le\mathbb{E}[\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_{\mathcal{H}}(x_t,x_j))(y_t-y_j)(f_{\mathcal{H}}(x_t,x_j)-f_t(x_t,x_j))]\\
&=\mathbb{E}[\langle\sum_{j=1}^{t-1}\phi'((y-y_j)f_{\mathcal{H}}(x,x_j))(y-y_j)K_{(x,x_j)},f_{\mathcal{H}}-f_t\rangle_K]\\
&\le 2\sqrt{t-1}(\mathbb{E}[\|f_t-f_{\mathcal{H}}\|_K^2])^{\frac12}\sigma_{\mathcal{H}}\le(t-1)\gamma_t\mathbb{E}[\|f_t-f_{\mathcal{H}}\|_K^2]+\tfrac{\sigma_{\mathcal{H}}^2}{\gamma_t}.
\end{aligned}
$$

Also, $\mathbb{E}\big[\big\|\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_t(x_t,x_j))(y_t-y_j)K_{(x_t,x_j)}\big\|_K^2\big]$ can be bounded by

$$
\begin{aligned}
&2\mathbb{E}\big[\big\|\sum_{j=1}^{t-1}(\phi'((y_t-y_j)f_t(x_t,x_j))-\phi'((y_t-y_j)f_{\mathcal{H}}(x_t,x_j)))(y_t-y_j)K_{(x_t,x_j)}\big\|_K^2\big]\\
&\quad+2\mathbb{E}\big[\big\|\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_{\mathcal{H}}(x_t,x_j))(y_t-y_j)K_{(x_t,x_j)}\big\|_K^2\big]\\
&\le 32\widetilde{\kappa}^4L^2(t-1)^2\|f_t-f_{\mathcal{H}}\|_K^2+8(t-1)\sigma_{\mathcal{H}}^2.
\end{aligned}
$$

Putting these two estimates into (3.11), we have

$$
\mathbb{E}\big[\|f_{t+1}-f_{\mathcal{H}}\|_K^2\big]\le(1+(32\widetilde{\kappa}^4L^2+2)\gamma_t^2)\mathbb{E}\big[\|f_t-f_{\mathcal{H}}\|_K^2\big]+\frac{(8\gamma_t^2+2)\sigma_{\mathcal{H}}^2}{t-1}.
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\big[\|f_{t+1}-f_{\mathcal{H}}\|_K^2\big]&\le\textstyle\prod_{j=2}^t(1+(32\widetilde{\kappa}^4L^2+2)\gamma_j^2)\|f_{\mathcal{H}}\|_K^2\\
&\quad+\sigma_{\mathcal{H}}^2\textstyle\sum_{j=2}^t\prod_{k=j+1}^t(1+(32\widetilde{\kappa}^4L^2+2)\gamma_k^2)\big[8\gamma_j^2+\frac{2}{j-1}\big]\\
&\le\exp\big((32\widetilde{\kappa}^4L^2+2)\textstyle\sum_{j=2}^t\gamma_j^2\big)\|f_{\mathcal{H}}\|_K^2\\
&\quad+\frac{8\sigma_{\mathcal{H}}^2}{32\widetilde{\kappa}^4L^2+2}\textstyle\sum_{j=2}^t\Big[\prod_{k=j}^t(1+(32\widetilde{\kappa}^4L^2+2)\gamma_k^2)-\prod_{k=j+1}^t(1+(32\widetilde{\kappa}^4L^2+2)\gamma_k^2)\Big]\\
&\quad+\sigma_{\mathcal{H}}^2\textstyle\sum_{j=2}^t\prod_{k=j+1}^t(1+(32\widetilde{\kappa}^4L^2+2)\gamma_k^2)\frac{2}{j-1}\\
&\le\exp\big((2+32\widetilde{\kappa}^4L^2)\textstyle\sum_{j=2}^t\gamma_j^2\big)\big[\|f_{\mathcal{H}}\|_K^2+2\sigma_{\mathcal{H}}^2(3+\ln t)\big].
\end{aligned}
$$

This completes the proof of the lemma. □

From the above lemma, we know if $\gamma_t=\frac{c}{t^\theta}$ with some $\theta\in(\frac12,1)$. Then,

$$
\mathbb{E}\big[\|f_t-f_{\mathcal{H}}\|_K^2\big]\le E_t:=\exp\big(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)[\|f_{\mathcal{H}}\|_K^2+2\sigma_{\mathcal{H}}^2(3+\ln t)].\tag{3.12}
$$

The next lemma estimates the boundedness of the learning sequence under the RKHS norm.

Lemma 4. *Let ϕ be 1-activating and $\{f_t:t=1,\dots,T+1\}$ be given by Algorithm 2. If $\gamma_t\widetilde{\kappa}^2\le\frac{1}{4L}$ for any $t\in\mathbb{N}$ then*

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\|f_{t+1}\|_K \le \widetilde{D}_t := C_\phi \sqrt{\sum_{j=2}^{t} \gamma_j},$$

where $C_\phi = \sqrt{L} s_0$ *if there exists* $s_0 \in \mathbb{R}$ *such that* $\phi'(s_0) = 0$, *and* $C_\phi = \sqrt{2\phi(0) + \frac{2(\phi'(0))^2}{L}}$ *otherwise.*

Proof. Write

$$\begin{aligned}
\|f_{t+1}\|_K^2 &= \|f_t\|_K^2 + \tfrac{\gamma_t^2}{(t-1)^2} \| \textstyle\sum_{j=1}^{t-1} \phi'((y_t - y_j) f_t(x_t, x_j))(y_t - y_j) K_{(x_t, x_j)} \|_K^2 \\
&\quad - \tfrac{2\gamma_t}{t-1} \textstyle\sum_{j=1}^{t-1} \phi'((y_t - y_j) f_t(x_t, x_j))(y_t - y_j) f_t(x_t, x_j) \\
&\le \|f_t\|_K^2 + \tfrac{\gamma_t}{t-1} \textstyle\sum_{j=1}^{t-1} \Big[4\widetilde{\kappa}^2 \gamma_t |\phi'((y_t - y_j) f_t(x_t, x_j))|^2 \\
&\quad - 2\phi'((y_t - y_j) f_t(x_t, x_j))(y_t - y_j) f_t(x_t, x_j) \Big] \\
&\le \|f_t\|_K^2 + \gamma_t \sup_{s \in \mathbb{R}} \big[4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - 2\phi'(s) s \big].
\end{aligned}$$

Therefore, the desired result follows directly from the following claim:

$$\sup_{s \in \mathbb{R}} \big[4(\phi'(s))^2 \gamma_t \kappa^2 - 2\phi'(s) s \big] \le C_\phi^2, \qquad \text{if } \gamma_t \widetilde{\kappa}^2 \le \frac{1}{4L}. \tag{3.13}$$

To prove (3.13), we discuss the following two cases.

Case 1: $\phi'(s) \le 0$ *for any* $s \in \mathbb{R}$. Firstly, consider $s \ge 0$. By the convexity of ϕ, $-s\phi'(s) \le \phi(0) - \phi(s) \le \phi(0)$. In addition, $\phi'(0) \le \phi'(s) \le 0$. Hence, for $s \ge 0$, there holds

$$4(\phi'(s))^2 \gamma_t \kappa^2 - 2\phi'(s) s \le 4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 + 2\phi(0) \le \frac{(\phi'(0))^2}{L} + 2\phi(0). \tag{3.14}$$

Secondly, consider $s < 0$ which implies $s\phi'(0) > 0$. Since $\phi'(\cdot)$ is Lipschitz continuous, part (c) of Proposition 1 implies that $(\phi'(s) - \phi'(0)) s \ge \frac{(\phi'(s) - \phi'(0))^2}{L} = \frac{(|\phi'(s)| - |\phi'(0)|)^2}{L}$. Therefore, for $s < 0$, we have

$$\begin{aligned}
4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - 2\phi'(s) s &\le 4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - 2(\phi'(s) - \phi'(0)) s \\
&\le 4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - \tfrac{2(|\phi'(s)| - |\phi'(0)|)^2}{L} \\
&\le \tfrac{(\phi'(s))^2}{L} - \tfrac{2(|\phi'(s)| - |\phi'(0)|)^2}{L} \\
&= -\tfrac{1}{L}(|\phi'(s)| - 2|\phi'(0)|)^2 + \tfrac{2(\phi'(0))^2}{L} \le \tfrac{2(\phi'(0))^2}{L}.
\end{aligned} \tag{3.15}$$

Combining the above estimates (3.14) and (3.15) yields that

$$\sup_{s \in \mathbb{R}} \big[4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - 2\phi'(s) s \big] \le 2\phi(0) + \frac{2(\phi'(0))^2}{L}.$$

Case 2: $\phi'(s_1) > 0$ *for some* $s_1 \in \mathbb{R}$. Since ϕ' is increasing and $\phi'(0) < 0$ by assumption, therefore s_1 must be positive and there exists $s_0 > 0$ such that $\phi'(s_0) = 0$. Hence, by part (b) of Proposition 1, we have

$$\begin{aligned}
4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - 2\phi'(s) s &= 4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - 2(\phi'(s) - \phi'(s_0))(s - s_0) - 2s_0 \phi'(s) \\
&\le 4(\phi'(s))^2 \gamma_t \widetilde{\kappa}^2 - \tfrac{2}{L}(\phi'(s) - \phi'(s_0))^2 - 2s_0 \phi'(s) \\
&= (4\gamma_t \widetilde{\kappa}^2 - \tfrac{2}{L})(\phi'(s))^2 - 2s_0 \phi'(s) \\
&\le -\tfrac{1}{L}(\phi'(s))^2 - 2s_0 \phi'(s) = -\tfrac{1}{L}(\phi'(s) + L s_0)^2 + L(s_0)^2,
\end{aligned}$$

which implies that

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\sup_{s\in\mathbb{R}}\left[4(\phi'(s))^2\gamma_t\widetilde{\kappa}^2 - 2\phi'(s)s\right] \le L(s_0)^2.$$

Combining the estimates in the above two cases yields (3.13). This completes the proof of the lemma. □

From the above lemma, we know that if $\gamma_t = \frac{c}{t^\theta}$ with $\theta \in (0,1)$ then

$$\|f_t\|_K \le C_\phi \sqrt{\sum_{j=2}^{t-1}\gamma_j} \le \frac{\sqrt{c}C_\phi}{\sqrt{1-\theta}}\, t^{\frac{1-\theta}{2}}. \tag{3.16}$$

The analysis for Algorithm 2 also needs the concept of Rademacher averages [5]. Let $\mathcal{F}$ be a class of uniformly bounded functions. The (empirical) Rademacher average $R_n(\mathcal{F})$ over $\mathcal{F}$ is defined by

$$R_n(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{f\in F}\frac{1}{n}\sum_{j=1}^{n}\sigma_i f(z_j)\right],$$

where $\{z_j : j = 1,2,\dots,n\}$ are independent random variables distributed according to some probability measure and $\{\sigma_j : j = 1,2,\dots,n\}$ are independent Rademacher random variables, i.e. $P(\sigma_j = 1) = P(\sigma_j = -1) = \frac{1}{2}$. Another useful complexity to describe the capacity of $\mathcal{F}$ is the Gaussian average which is defined by

$$G_n(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{f\in F}\frac{1}{n}\sum_{j=1}^{n}g_j f(z_j)\right],$$

where $\{g_j : j = 1,2,\dots,n\}$ are independent Gaussian $\mathcal{N}(0,1)$ random variables. The following inequality (e.g. [20, Remark 2.26]) describes the relationship between the above complexity averages:

$$\frac{\rho G_n(F)}{\ln n} \le R_n(F) \le \mu\, G_n(F). \tag{3.17}$$

Here, $\mu > 0$ and $\rho > 0$ are absolute constants independent of F and n.

We begin with stating the well-known comparison principles for Gaussian process (e.g. [25]) which will be used to prove a useful property of Gaussian averages.

Lemma 5. *Let $\{X_\theta : \theta \in \Theta\}$ and $\{Y_\theta : \theta \in \Theta\}$ be two zero-mean Gaussian processes indexed by the same countable set Θ and suppose that*

$$\mathbb{E}_g[(Y_\theta - Y_{\bar\theta})^2] \le \mathbb{E}_g[(X_\theta - X_{\bar\theta})^2], \quad \forall\theta,\bar\theta \in \Theta.$$

Then,

$$\mathbb{E}_g[\sup_\theta Y_\theta] \le \mathbb{E}_g[\sup_\theta X_\theta].$$

We now can derive the following property related to the Gaussian average.

Lemma 6. *Let $F_j(\theta)$ be a set of functions indexed by parameters $\theta = (\theta_1,\theta_2) \in \Theta_1 \times \Theta_2$, $H_j(\theta_1)$, and $J_j(\theta_2)$ be a set of functions indexed, respectively, by parameter $\theta_1 \in \Theta_1$, and $\theta_2 \in \Theta_2$. Assume, for any $\theta = (\theta_1,\theta_2), \bar\theta = (\bar\theta_1,\bar\theta_2) \in \Theta_1 \times \Theta_2$, that*

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$|F_j(\theta) - F_j(\bar{\theta})|^2 \le |H_j(\theta_1) - H_j(\bar{\theta}_1)|^2 + |J_j(\theta_2) - J_j(\bar{\theta}_2)|^2.$$

Then,

$$\mathbb{E}_g[\sup_{(\theta_1,\theta_2)\in\Theta_1\times\Theta_2}\sum_{i=1}^n g_i F_i(\theta)] \le \mathbb{E}_g[\sup_{\theta_1\in\Theta_1}\sum_{j=1}^n g_j H_j(\theta_1)] + \mathbb{E}_g[\sup_{\theta_2\in\Theta_2}\sum_{j=1}^n g_j J_j(\theta_2)].$$

Proof. Let $g_1,\dots,g_{2n}$ be $2n$ independent $\mathcal{N}(0,1)$ Gaussian variables. Introduce two Gaussian processes:

$$X_\theta = \sum_{j=1}^n g_j F_j(\theta) \quad \text{and} \quad Y_\theta = \sum_{j=1}^n [g_j H_j(\theta_1) + g_{n+j} J_j(\theta_2)].$$

Then, $\mathbb{E}_g[(X_\theta - X_{\bar{\theta}})^2] = \sum_{j=1}^n [F_j(\theta) - F_j(\bar{\theta})]^2$, and $\mathbb{E}_g[(Y_\theta - Y_{\bar{\theta}})^2] = \sum_{j=1}^n \left[(H_j(\theta_1) - H_j(\bar{\theta}_1))^2 + (J_j(\theta_2) - J_j(\bar{\theta}_2))^2\right]$. According to Lemma 5, we have

$$\begin{aligned}
\mathbb{E}_g[\sup_{\theta\in\Theta}\sum_{j=1}^n g_j F_j(\theta)] &\le \mathbb{E}_g[\sup_{\theta\in\Theta}(\sum_{j=1}^n g_j H_j(\theta_1) + \sum_{j=1}^n g_{n+j} J_j(\theta_2))] \\
&\le \mathbb{E}_g[\sup_{\theta_1\in\Theta_1}\sum_{j=1}^n g_j H_j(\theta_1)] + \mathbb{E}_g[\sup_{\theta_2\in\Theta_2}\sum_{j=1}^n g_{n+j} J_j(\theta_2)] \\
&= \mathbb{E}_g[\sup_{\theta_1\in\Theta_1}\sum_{j=1}^n g_j H_j(\theta_1)] + \mathbb{E}_g[\sup_{\theta_2\in\Theta_2}\sum_{j=1}^n g_j J_j(\theta_2)].
\end{aligned}$$

This completes the proof of the lemma. □

Denote

$$M_t^\phi = \sup_{|u|\le 2\widetilde{\kappa}\widetilde{D}_t} |\phi'(u)|. \tag{3.18}$$

We also need to bound the following term defined by

$$\Delta_t := \nabla\widetilde{\mathcal{E}}(f_t) - \frac{1}{t-1}\sum_{j=1}^{t-1}\int_{\mathcal{Z}} \phi'((\widetilde{y}-y_j)f_t(\widetilde{x},x_j))(\widetilde{y}-y_j)K_{(\widetilde{x},x_j)}d\rho(\widetilde{x},\widetilde{y}),$$

where $\nabla\widetilde{\mathcal{E}}(f_t)$ denotes the functional derivative of $\widetilde{\mathcal{E}}(\cdot)$ at f_t given by

$$\nabla\widetilde{\mathcal{E}}(f_t) = \iint_{\mathcal{Z}\times\mathcal{Z}} \phi'((\widetilde{y}-y)f_t(\widetilde{x},x))(\widetilde{y}-y)K_{(\widetilde{x},x)}d\rho(z)d\rho(\widetilde{z}).$$

Using Lemma 6, we can prove the following estimation.

Lemma 7. *Let ϕ be 1-activating, and $\{f_t : t = 1,\dots,T+1\}$ be given by Algorithm 2. If $\gamma_t\widetilde{\kappa}^2 \le \frac{1}{4L}$ then, for any $t \ge 2$,*

$$\mathbb{E}[\|\Delta_t\|_K] \le \frac{8\sqrt{2}\mu(L\widetilde{\kappa}\widetilde{D}_t + M_t^\phi)\widetilde{\kappa}}{\sqrt{t-1}}.$$

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Proof. For any fixed $\widetilde{z}=(\widetilde{x},\widetilde{y})$ and $z=(x,y)$, denote $\xi_{f,h}(\widetilde{z},z_j)=\phi'((\widetilde{y}-y_j)f(\widetilde{x},x_j))(\widetilde{y}-y_j)h(\widetilde{x},x_j)$. Since $\gamma_t\widetilde{\kappa}^2\le\frac{1}{4L}$, by Lemma 4, $\|f_t\|_K\le\widetilde{D}_t$. Notice

$$\begin{aligned}\|\Delta_t\|_K &= \sup_{\|h\|_K\le 1}\Big[\iint\phi'((\widetilde{y}-y)f_t(\widetilde{x},x))(\widetilde{y}-y)h(\widetilde{x},x)d\rho(z)d\rho(\widetilde{z})\\ &\qquad -\tfrac{1}{t-1}\textstyle\sum_{j=1}^{t-1}\int_{\mathcal{Z}}\phi'((\widetilde{y}-y_j)f_t(\widetilde{x},x_j))(\widetilde{y}-y_j)h(\widetilde{x},x_j)d\rho(\widetilde{x},\widetilde{y})\Big]\\ &\le \sup_{\substack{\|f\|_K\le\widetilde{D}_t\\ \|h\|_K\le 1}}\Big[\iint\phi'((\widetilde{y}-y)f(\widetilde{x},x))(\widetilde{y}-y)h(\widetilde{x},x)d\rho(z)d\rho(\widetilde{z})\\ &\qquad -\tfrac{1}{t-1}\textstyle\sum_{j=1}^{t-1}\int_{\mathcal{Z}}\phi'((\widetilde{y}-y_j)f(\widetilde{x},x_j))(\widetilde{y}-y_j)h(\widetilde{x},x_j)d\rho(\widetilde{x},\widetilde{y})\Big]\\ &= \int_{\mathcal{Z}}\sup_{\substack{\|f\|_K\le\widetilde{D}_t\\ \|h\|_K\le 1}}\Big[\mathbb{E}_z\xi_{f,h}(\widetilde{z},z)-\frac{1}{t-1}\textstyle\sum_{j=1}^{t-1}\xi_{f,h}(\widetilde{z},z_j)\Big]d\rho(\widetilde{z}).\end{aligned}\tag{3.19}$$

For any fixed $\widetilde{z}=(\widetilde{x},\widetilde{y})$, by the standard symmetrization technique [4], from the above inequality we have

$$\begin{aligned}&\sup_{\substack{\|f\|_K\le\widetilde{D}_t\\ \|h\|_K\le 1}}\Big[\mathbb{E}_z\xi_{f,h}(\widetilde{z},z)-\frac{1}{t-1}\sum_{j=1}^{t-1}\xi_{f,h}(\widetilde{z},z_j)\Big]\\ &\le 2\mathbb{E}_z\mathbb{E}_\sigma\sup_{\substack{\|f\|\le\widetilde{D}_t\\ \|h\|_K\le 1}}\frac{1}{t-1}\sum_{j=1}^{t-1}\sigma_j\xi_{f,h}(\widetilde{z},z_j)\\ &\le 2\mu\mathbb{E}_z\mathbb{E}_g\sup_{\substack{\|f\|\le\widetilde{D}_t\\ \|h\|_K\le 1}}\frac{1}{t-1}\sum_{j=1}^{t-1}g_j\xi_{f,h}(\widetilde{z},z_j),\end{aligned}\tag{3.20}$$

where the last inequality follows from (3.17). Let $\Theta_1=\{f\in\mathcal{H}_K:\|f\|_K\le\widetilde{D}_t\}$ and $\Theta_2=\{h\in\mathcal{H}_K:\|h\|_K\le 1\}$. Then, for any $f,\bar{f}\in\Theta_1$ and $h,\bar{h}\in\Theta_2$, there holds

$$|\xi_{f,h}(\widetilde{z},z)-\xi_{\bar{f},\bar{h}}(\widetilde{z},z)|^2\le(4\sqrt{2}L\widetilde{\kappa}|f(x,x_j)-\bar{f}(x,x_j)|)^2+(2\sqrt{2}M_t^\phi|h(x,x_j)-\bar{h}(x,x_j)|)^2.$$

Applying Lemma 6 with $F_i(\theta)=\xi_{f,h}(\widetilde{z},z)$ with $\theta_1=f$, $\theta_2=h$, $H_j(\theta_1)=4\sqrt{2}L\widetilde{\kappa}f(x,x_j)$, and $J_j(\theta_2)=2\sqrt{2}M_t^\phi h(x,x_j)$ yields that

$$\begin{aligned}&\mathbb{E}_g\Big[\sup_{\substack{\|f\|\le\widetilde{D}_t\\ \|h\|_K\le 1}}\frac{1}{t-1}\sum_{j=1}^{t-1}g_j\xi_{f,h}(\widetilde{z},z_j)\Big]\\ &\le 4\sqrt{2}L\widetilde{\kappa}\mathbb{E}_g[\sup_{\|f\|\le\widetilde{D}_t}\frac{1}{t-1}\sum_{j=1}^{t-1}g_jf(x,x_j)]+2\sqrt{2}M_t^\phi\mathbb{E}_g[\sup_{\|h\|_K\le 1}\frac{1}{t-1}\sum_{j=1}^{t-1}g_jh(x,x_j)]\\ &= 4\sqrt{2}L\widetilde{\kappa}\,\mathbb{E}_g\sup_{\|f\|\le\widetilde{D}_t}\langle\frac{1}{t-1}\sum_{j=1}^{t-1}g_jK_{(x,x_j)},f\rangle_K+2\sqrt{2}M_t^\phi\,\mathbb{E}_g\sup_{\|h\|_K\le 1}\langle\frac{1}{t-1}\sum_{j=1}^{t-1}g_jK_{(x,x_j)},h\rangle_K\\ &\le 4\sqrt{2}L\widetilde{\kappa}\widetilde{D}_t\mathbb{E}_g\|\tfrac{1}{t-1}\textstyle\sum_{j=1}^{t-1}g_jK_{(x,x_j)}\|_K+2\sqrt{2}M_t^\phi\,\mathbb{E}_g\|\tfrac{1}{t-1}\textstyle\sum_{j=1}^{t-1}g_jK_{(x,x_j)}\|_K\\ &\le 4\sqrt{2}L\widetilde{\kappa}\widetilde{D}_t\big(\mathbb{E}_g\|\tfrac{1}{t-1}\textstyle\sum_{j=1}^{t-1}g_jK_{(x,x_j)}\|_K^2\big)^{1/2}+2\sqrt{2}M_t^\phi\,\big(\mathbb{E}_g\|\tfrac{1}{t-1}\textstyle\sum_{j=1}^{t-1}g_jK_{(x,x_j)}\|_K^2\big)^{1/2}\\ &\le\tfrac{4\sqrt{2}(L\widetilde{\kappa}\widetilde{D}_t+M_t^\phi)\widetilde{\kappa}}{\sqrt{t-1}}.\end{aligned}$$

Putting the above estimation, (3.19), and (3.20) together yields the desired result. □

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Denote, for any $t \in \mathbb{N}$, $R_t = \mathbb{E}[\widetilde{\mathcal{E}}(f_t) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})]$. We derive the following recursive inequality for R_t which is critical for proving Theorem 3.

Lemma 8. *Let ϕ be a 1-activating loss, $\{f_t : t = 1, \ldots, T+1\}$ be given by Algorithm 2. Then, for any $t \geq 2$,*

$$R_{t+1} \leq R_t - \gamma_t \frac{R_t^2}{E_t} + \frac{16\sqrt{2}\mu\widetilde{\kappa}^2 M_t^{\phi}(L\widetilde{\kappa}\widetilde{D}_t + M_t^{\phi})\gamma_t}{\sqrt{t-1}} + 8L\widetilde{\kappa}^4\gamma_t^2(M_t^{\phi})^2. \tag{3.21}$$

Proof. By part (a) of Proposition 1, we have

$$\begin{aligned}\phi((y-\widetilde{y})f_{t+1}(x,\widetilde{x})) \leq\ & \phi((y-\widetilde{y})f_t(x,\widetilde{x})) + \langle \phi'((y-\widetilde{y})f_t(x,\widetilde{x}))(y-\widetilde{y})K_{(x,\widetilde{x})}, f_{t+1}-f_t\rangle_K \\ & + 2L\big|f_{t+1}(x,\widetilde{x}) - f_t(x,\widetilde{x})\big|^2.\end{aligned}$$

Therefore, letting $\Delta_t = \nabla\widetilde{\mathcal{E}}(f_t) - \frac{1}{t-1}\sum_{j=1}^{t-1}\int_{\mathcal{Z}} \phi'((y-y_j)f_t(x,x_j))(y-y_j)K_{(x,x_j)}d\rho(x,y)$, we know that $\mathbb{E}[\widetilde{\mathcal{E}}(f_{t+1})]$ is bounded by

$$\begin{aligned}&\mathbb{E}[\widetilde{\mathcal{E}}(f_t)] - \mathbb{E}\langle\nabla\widetilde{\mathcal{E}}(f_t), \tfrac{\gamma_t}{t-1}\textstyle\sum_{j=1}^{t-1}\int_{\mathcal{Z}}\phi'((y-y_j)f_t(x,x_j))(y-y_j)K_{(x,x_j)}d\rho(x,y)\rangle_K \\ &\quad + \tfrac{2L\widetilde{\kappa}^4\gamma_t^2}{(t-1)^2}\mathbb{E}\big(\textstyle\sum_{j=1}^{t-1}|\phi'((y_t-y_j)f_t(x_t,x_j))(y_t-y_j)|\big)^2 \\ &\leq \mathbb{E}[\widetilde{\mathcal{E}}(f_t)] - \gamma_t\mathbb{E}[\|\nabla\widetilde{\mathcal{E}}(f_t)\|_K^2] + \gamma_t\mathbb{E}\langle\nabla\widetilde{\mathcal{E}}(f_t),\Delta_t\rangle_K \\ &\quad + \tfrac{8L\widetilde{\kappa}^4\gamma_t^2}{t-1}\mathbb{E}\big[\textstyle\sum_{j=1}^{t-1}|\phi'((y_t-y_j)f_t(x_t,x_j))|^2\big] \\ &\leq \mathbb{E}[\widetilde{\mathcal{E}}(f_t)] - \gamma_t\mathbb{E}[\|\nabla\widetilde{\mathcal{E}}(f_t)\|_K^2] + \gamma_t\mathbb{E}\big[\|\nabla\widetilde{\mathcal{E}}(f_t)\|_K\|\Delta_t\|_K\big] \\ &\quad + \tfrac{8L\widetilde{\kappa}^4\gamma_t^2}{t-1}\mathbb{E}\big[\textstyle\sum_{j=1}^{t-1}|\phi'((y_t-y_j)f_t(x_t,x_j))|^2\big] \\ &\quad + \tfrac{8L\widetilde{\kappa}^4\gamma_t^2}{t-1}\mathbb{E}\big[\textstyle\sum_{j=1}^{t-1}|\phi'((y_t-y_j)f_t(x_t,x_j))|^2\big] \\ &\leq \mathbb{E}[\widetilde{\mathcal{E}}(f_t)] - \gamma_t\mathbb{E}[\|\nabla\widetilde{\mathcal{E}}(f_t)\|_K^2] + 2\widetilde{\kappa}\gamma_t M_t^{\phi}\mathbb{E}\big[\|\Delta_t\|_K\big] \\ &\quad + \tfrac{8L\widetilde{\kappa}^4\gamma_t^2}{t-1}\mathbb{E}\big[\textstyle\sum_{j=1}^{t-1}|\phi'((y_t-y_j)f_t(x_t,x_j))|^2\big].\end{aligned} \tag{3.22}$$

Notice

$$\mathbb{E}\big[\frac{1}{t-1}\sum_{j=1}^{t-1}|\phi'((y_t-y_j)f_t(x_t,x_j))|^2\big] \leq (M_t^{\phi})^2. \tag{3.23}$$

By the convexity of ϕ, $\widetilde{\mathcal{E}}(f_t) - \widetilde{\mathcal{E}}(f_{\mathcal{H}}) \leq \langle\nabla\widetilde{\mathcal{E}}(f_t), f_t - f_{\mathcal{H}}\rangle$ which, combined with Lemma 3, implies that

$$\mathbb{E}[\|\nabla\widetilde{\mathcal{E}}(f_t)\|_K^2] \geq \frac{\mathbb{E}\big[(\widetilde{\mathcal{E}}(f_t) - \widetilde{\mathcal{E}}(f_{\mathcal{H}}))^2\big]}{\mathbb{E}[\|f_t - f_{\mathcal{H}}\|_K^2]} \geq \frac{R_t^2}{E_t}.$$

Combining the above inequality, (3.22) and (3.23) together, by noting $R_t = \mathbb{E}[\widetilde{\mathcal{E}}(f_t) - \widetilde{\mathcal{E}}(f_{\mathcal{H}})]$, we have

$$R_{t+1} \leq R_t - \gamma_t\frac{R_t^2}{E_t} + \frac{16\sqrt{2}\mu\widetilde{\kappa}^2 M_t^{\phi}(L\widetilde{\kappa}\widetilde{D}_t + M_t^{\phi})\gamma_t}{\sqrt{t-1}} + 8L\widetilde{\kappa}^4\gamma_t^2(M_t^{\phi})^2.$$

This completes the proof of the lemma. □

From (3.21), in analogy to the proof used in Theorem 1, one can easily see that a sufficient condition to guarantee the convergence of $\mathbb{E}[\widetilde{\mathcal{E}}(f_t)]$ to $\widetilde{\mathcal{E}}(f_{\mathcal{H}})$ can be stated as follows:

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\sum_{t=2}^{\infty}\frac{\gamma_t}{\ln t}=\infty,\quad \sum_{t=2}^{\infty}\Big[(\frac{M_t^{\phi}\widetilde{D}_t+(M_t^{\phi})^2)\gamma_t}{\sqrt{t-1}}+\gamma_t^2(M_t^{\phi})^2\Big]<\infty. \tag{3.24}$$

This sufficient condition is not as neat as its counterpart to guarantee the convergence of Algorithm 1 as given by Theorem 1. Observe that the randomized gradient $\frac{1}{t-1}\sum_{j=1}^{t-1}\phi'((y_t-y_j)f_t(x_t,x_j))(y_t-y_j)K_{(x_t,x_j)}$ in Algorithm 2 is not an unbiased estimator of the true gradient $\iint_{\mathcal{Z}\times\mathcal{Z}}\phi'((y-y')f_t(x,x'))(y-y')K_{(x,x)}d\rho(x,y)d\rho(x',y')$, even conditioned on $\{z_1,z_2,\dots,z_{t-1}\}$. This fact may partly explain why our techniques can not be used to derive a similar sufficient condition as the one for Algorithm 1 which is stated in Theorem 1.

Lemma 9. *For any $x,\nu,a>0$, there holds*

$$a\ln x\le \nu x+a\ln\big(\frac{a}{\nu e}\big).$$

Proof. The lemma directly follows from the inequality in [21], i.e. $e^{-\nu x}\le\big(\frac{a}{\nu e}\big)^a x^{-a}$. □

We are now in a position to prove Theorem 3 by induction.

Proof of Theorem 3. Denote $a_{\mathcal{H}}=\|f_{\mathcal{H}}\|_K^2+6\sigma_{\mathcal{H}}^2$, and for any $\delta\in(0,\min(\theta-\frac12,1-\theta))$, let

$$\beta=\min(\frac{\theta-\delta}{2}-\frac14,1-\theta-\delta)>0.$$

Now let

$$D:=\widetilde{C}_{\theta,\delta,\mathcal{H}}=\max\big\{D_1,D_2,D_3\big\}, \tag{3.25}$$

where $D_1=\frac{1}{2c}\exp(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1})a_{\mathcal{H}}$,

$$\begin{aligned}D_2=\;&2^{\frac{(\beta+1)(\theta+\beta)}{\theta}}\big(\tfrac{c}{a_{\mathcal{H}}}\big)^{\frac{\beta}{\theta}}\Big\{2L\widetilde{\kappa}^2\exp\big(\tfrac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)a_{\mathcal{H}}\\&+\big(\tfrac{4\sigma_{\mathcal{H}}^2L\widetilde{\kappa}^2\exp(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1})}{\theta}\big)\big[\tfrac{(2+32\widetilde{\kappa}^4L^2)c^2\theta}{(2\theta-1)(\theta+\beta)}+\ln\big(2^{3+\beta}c^{\frac{\beta}{\theta+\beta}}\sigma_{\mathcal{H}}^2L\widetilde{\kappa}^2\theta^{-1}a_{\mathcal{H}}^{-\frac{\beta}{\theta+\beta}}\big)\big]\\&\times 2\sigma_{\mathcal{H}}^2L\widetilde{\kappa}^2\exp\big(\tfrac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)\big[\ln 2+\tfrac{1}{\theta+\beta}\ln\big(\tfrac{c}{a_{\mathcal{H}}}\big)\big]\Big\}^{\frac{\theta+\beta}{\theta}},\end{aligned}$$

and

$$D_3=\frac{2}{c}\exp\big(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)\big[a_{\mathcal{H}}+\frac{4\sigma_{\mathcal{H}}^2}{\delta}\ln\frac{1}{\delta}\big]+8\widetilde{\kappa}^2(Lc\widetilde{\kappa}^2+4\sqrt{2}\mu)\big(\frac{3L\widetilde{\kappa}\sqrt{c}C_{\phi}}{\sqrt{1-\theta}}+|\phi'(0)|\big)^2.$$

Let $t_0=\big\lfloor 2\big(\frac{2cD}{\exp(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1})a_{\mathcal{H}}}\big)^{\frac{1}{\theta+\beta}}\big\rfloor$. Since $D\ge\frac{1}{2c}\exp(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1})a_{\mathcal{H}}$ and $\theta+\beta<1$, we have $t_0\ge2$. Notice

$$\begin{aligned}R_{t_0}&=\mathbb{E}[\widetilde{\mathcal{E}}(f_{t_0})-\widetilde{\mathcal{E}}(f_{\mathcal{H}})]\le 2L\widetilde{\kappa}^2\mathbb{E}(\|f_{t_0}-f_{\mathcal{H}}\|_K^2)\le 2L\widetilde{\kappa}^2E_{t_0}\\&\le 2L\widetilde{\kappa}^2\exp\big(\tfrac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)[a_{\mathcal{H}}+2\sigma_{\mathcal{H}}^2\ln t_0)]\\&\le 2L\widetilde{\kappa}^2\exp\big(\tfrac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)a_{\mathcal{H}}+\frac{2\sigma_{\mathcal{H}}^2L\widetilde{\kappa}^2\exp(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1})}{\theta}\ln D^{\frac{\theta}{\theta+\beta}}\\&\quad+4\sigma_{\mathcal{H}}^2L\widetilde{\kappa}^2\exp\big(\tfrac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)\big[\ln 2+\tfrac{1}{\theta+\beta}\ln\big(\tfrac{c}{a_{\mathcal{H}}}\big)\big].\end{aligned} \tag{3.26}$$

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

Applying Lemma 9 with $a = \frac{4\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2 \exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)}{\theta}$, $\nu = 2^{-1-\beta}\left(\frac{2\exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)a_{\mathcal{H}}}{c}\right)^{\frac{\beta}{\theta+\beta}}$ and $x = D^{\frac{\theta}{\theta+\beta}}$ implies that

$$\begin{aligned}\frac{4\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2 \exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)}{\theta}\ln D^{\frac{\theta}{\theta+\beta}} \le\ & 2^{-1-\beta}\left(\frac{2\exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)a_{\mathcal{H}}}{c}\right)^{\frac{\beta}{\theta+\beta}} D^{\frac{\theta}{\theta+\beta}} \\ &+ \left(\frac{4\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2 \exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)}{\theta}\right)\Big[\frac{(2+32\widetilde{\kappa}^4L^2)c^2\theta}{(2\theta-1)(\theta+\beta)} \\ &+ \ln\left(2^{3+\beta}c^{\frac{\beta}{\theta+\beta}}\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2\theta^{-1}a_{\mathcal{H}}^{-\frac{\beta}{\theta+\beta}}\right)\Big].\end{aligned}$$

Putting this estimation back into (3.26), we have, for any $t \le t_0$,

$$\begin{aligned}R_t \le\ & 2L\widetilde{\kappa}^2 E_{t_0} \le 2L\widetilde{\kappa}^2 \exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)a_{\mathcal{H}} \\ &+ 4\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2 \exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)\left[\ln 2 + \frac{1}{\theta+\beta}\ln\left(\frac{c}{a_{\mathcal{H}}}\right)\right] \\ &+ 2^{-1-\beta}\left(\frac{2\exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)a_{\mathcal{H}}}{c}\right)^{\frac{\beta}{\theta+\beta}} D^{\frac{\theta}{\theta+\beta}} \\ &+ \left(\frac{4\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2 \exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)}{\theta}\right)\Big[\frac{(2+32\widetilde{\kappa}^4L^2)c^2\theta}{(2\theta-1)(\theta+\beta)} + \ln\left(2^{3+\beta}c^{\frac{\beta}{\theta+\beta}}\sigma_{\mathcal{H}}^2 L\widetilde{\kappa}^2\theta^{-1}a_{\mathcal{H}}^{-\frac{\beta}{\theta+\beta}}\right)\Big] \\ \le\ & 2^{-\beta}\left(\frac{2\exp\left(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\right)a_{\mathcal{H}}}{c}\right)^{\frac{\beta}{\theta+\beta}} D^{\frac{\theta}{\theta+\beta}} \le \frac{D}{t_0^\beta} \le \frac{D}{t^\beta},\end{aligned} \tag{3.27}$$

where, in the last to third inequality, we have used the fact that $D \ge D_2$.

We can now prove the theorem by induction. Due to (3.27), $R_t \le \frac{D}{t^\beta}$ certainly holds true for $t \le t_0$. Now assume $R_t \le \frac{D}{t^\beta}$ for some $t \ge t_0$.

To estimate R_{t+1}, note, by the assumption on ϕ, that $M_t^\phi = \sup_{|u|\le 2\widetilde{\kappa}\widetilde{D}_t}|\phi'(u)| \le 2L\widetilde{\kappa}\widetilde{D}_t + |\phi'(0)|$, and $\gamma_t \le \frac{c}{\sqrt{t}}$ since $\theta > 1/2$, the recursive inequality (3.21) becomes

$$\begin{aligned}R_{t+1} &\le R_t - \gamma_t\frac{R_t^2}{E_t} + \frac{32\sqrt{2}\mu\widetilde{\kappa}^2 M_t^\phi(L\widetilde{\kappa}\widetilde{D}_t + M_t^\phi)\gamma_t}{\sqrt{t}} + \frac{8Lc\widetilde{\kappa}^4(M_t^\phi)^2\gamma_t}{\sqrt{t}} \\ &\le R_t - \gamma_t\frac{R_t^2}{E_t} + \frac{8\widetilde{\kappa}^2(Lc\widetilde{\kappa}^2+4\sqrt{2}\mu)(3L\widetilde{\kappa}\widetilde{D}_t+|\phi'(0)|)^2\gamma_t}{\sqrt{t}}.\end{aligned} \tag{3.28}$$

Consider the function $F(x) = x - \gamma_t\frac{x^2}{E_t}$ which is increasing if $x \in [0, \frac{E_t}{2\gamma_t}]$. By the definition of t_0, it is also easy to verify, for any $t \ge t_0$, that

$$\frac{D}{t^\beta} \le \frac{t^\theta E_t}{2c} = \frac{E_t}{2\gamma_t}.$$

Therefore, by recalling (3.16), i.e. $\widetilde{D}_t \le \frac{\sqrt{c}C_\phi}{\sqrt{1-\theta}}t^{\frac{1-\theta}{2}}$, we have

$$\begin{aligned}R_{t+1} &\le F(R_t) + \frac{8\widetilde{\kappa}^2(Lc\widetilde{\kappa}^2+4\sqrt{2}\mu)(3L\widetilde{\kappa}^2\widetilde{D}_t+|\phi'(0)|)^2\gamma_t}{\sqrt{t}} \\ &\le F(\tfrac{D}{t^\beta}) + \frac{8\widetilde{\kappa}^2(Lc\widetilde{\kappa}^2+4\sqrt{2}\mu)(3L\widetilde{\kappa}^2\widetilde{D}_t+|\phi'(0)|)^2\gamma_t}{\sqrt{t}} \\ &\le Dt^{-\beta} - \gamma_t\frac{D^2t^{-2\beta}}{E_t} + d_\theta t^{\frac{1}{2}-2\theta},\end{aligned} \tag{3.29}$$

where

$$d_\theta = 8\widetilde{\kappa}^2(Lc\widetilde{\kappa}^2 + 4\sqrt{2}\mu)\Big(\frac{3L\widetilde{\kappa}^2\sqrt{c}C_\phi}{\sqrt{1-\theta}} + |\phi'(0)|\Big)^2.$$

In addition, since $0 < \delta < \min(\theta - \frac{1}{2}, 1-\theta)$, applying Lemma 9 with $x = t^\delta, a = 1$, and $\nu = \delta$ implies that

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\ln t \le t^\delta + \frac{1}{\delta}\ln\frac{1}{\delta} \le \big[\frac{2}{\delta}\ln\frac{1}{\delta}\big]t^\delta.$$

Combining the above inequality with (3.12) yields that

$$E_t \le \exp\big(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)[a_{\mathcal{H}} + 2\sigma_{\mathcal{H}}^2\ln t)] \le \exp\big(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)[a_{\mathcal{H}} + \frac{4\sigma_{\mathcal{H}}^2}{\delta}\ln\frac{1}{\delta}]t^\delta := b_{\theta,\delta}\,t^\delta.$$

From the above inequality and (3.29), and noticing $\frac{1}{2}-\theta+2\beta+\delta\le 0$, $\theta+\beta+\delta\le 1$, we have

$$\begin{aligned} R_{t+1} &\le \tfrac{D}{t^\beta}\big[1-\tfrac{cD}{b_{\theta,\delta}}t^{-\theta-\beta-\delta}+\tfrac{d_\theta}{D}t^{\frac{1}{2}-2\theta+\beta}\big] \\ &= \tfrac{D}{t^\beta}\big[1-\big(\tfrac{cD}{b_{\theta,\delta}}-\tfrac{d_\theta}{D}t^{\frac{1}{2}-\theta+2\beta+\delta}\big)t^{-\theta-\beta-\delta}\big] \\ &\le \tfrac{D}{t^\beta}\big[1-\big(\tfrac{cD}{b_{\theta,\delta}}-\tfrac{d_\theta}{D}\big)t^{-\theta-\beta-\delta}\big] \\ &\le \tfrac{D}{t^\beta}\big[1-t^{-\theta-\beta-\delta}\big] \le \tfrac{D}{t^\beta}\big[1-t^{-1}\big] \\ &\le \tfrac{D}{t^\beta}\big[1-(t+1)^{-1}\big] \le \tfrac{D}{(t+1)^\beta}, \end{aligned} \tag{3.30}$$

where the last to fourth inequality used the fact that $\frac{cD}{b_{\theta,\delta}}-\frac{d_\theta}{D}\ge 1$ since $D\ge D_3=\frac{2b_{\theta,\delta}}{c}+d_\theta\ge\frac{1}{2}\big(\frac{b_{\theta,\delta}}{c}+\sqrt{\frac{b_{\theta,\delta}^2}{c^2}+\frac{4b_{\theta,\delta}d_\theta}{c}}\big)$. This completes the proof of the theorem. □

We turn our attention to the proof of Theorem 4.

Proof of Theorem 4. For any $\delta\in(0,\min(\frac{\theta}{4},1-\theta))$, and let

$$\beta=\min(\frac{\theta}{4}-\frac{\delta}{2},1-\theta-\delta)>0.$$

Let D_1, D_2 and t_0 be the same as those introduced in the proof for Theorem 3. Choose $D:=\bar{C}_{\theta,\delta,\mathcal{H}}=\max\{D_1,D_2,\widetilde{D}_3\}$, where

$$\widetilde{D}_3=\frac{2}{c}\exp\big(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1}\big)[a_{\mathcal{H}}+\frac{4\sigma_{\mathcal{H}}^2}{\delta}\ln\frac{1}{\delta}]+8\widetilde{\kappa}^2B(4\sqrt{2}\mu L\widetilde{\kappa}\frac{\sqrt{c}C_\phi}{\sqrt{1-\theta}}+(4\sqrt{2}\mu+Lc\widetilde{\kappa}^2)B).$$

Since $|\phi'(s)|\le B$ for any $s\in\mathbb{R}$, $M_t^\phi\le B$ holds true uniformly. Hence, for any $t\le t_0=\big\lfloor 2\big(\frac{2cD}{\exp(\frac{(2+32\widetilde{\kappa}^4L^2)c^2}{2\theta-1})a_{\mathcal{H}}}\big)^{\frac{1}{\theta+\beta}}\big\rfloor$, there holds $R_t\le\frac{D}{t_0^\beta}\le\frac{D}{t^\beta}$. Assume that, for some $t\ge t_0$, $R_t\le\frac{D}{t^\beta}$. We will prove that $R_{t+1}\le\frac{D}{(t+1)^\beta}$ by induction. To this end, observing that $M_t^\phi\le B$ holds true uniformly, we know from the recursive inequality (3.28) that

$$\begin{aligned} R_{t+1} &\le R_t-\gamma_t\frac{R_t^2}{E_t}+\frac{32\sqrt{2}\mu\widetilde{\kappa}^2M_t^\phi(L\widetilde{\kappa}\widetilde{D}_t+M_t^\phi)\gamma_t}{\sqrt{t}}+\frac{8Lc\widetilde{\kappa}^4(M_t^\phi)^2\gamma_t}{\sqrt{t}} \\ &\le R_t-\gamma_t\frac{R_t^2}{E_t}+\frac{8\widetilde{\kappa}^2B\big[4\sqrt{2}\mu L\widetilde{\kappa}\widetilde{D}_t+(4\sqrt{2}\mu+Lc\widetilde{\kappa}^2)B\big]\gamma_t}{\sqrt{t}}. \end{aligned}$$

Recalling (3.16) again, i.e. $\widetilde{D}_t\le\frac{\sqrt{c}C_\phi}{\sqrt{1-\theta}}t^{\frac{1-\theta}{2}}$, we have

$$\begin{aligned} R_{t+1} &\le F(R_t)+\frac{8\widetilde{\kappa}^2B\big[4\sqrt{2}\mu L\widetilde{\kappa}\widetilde{D}_t+(4\sqrt{2}\mu+Lc\widetilde{\kappa}^2)B\big]\gamma_t}{\sqrt{t}} \\ &\le Dt^{-\beta}-\gamma_t\frac{D^2t^{-2\beta}}{E_t}+\widetilde{d}_\theta t^{-\frac{3\theta}{2}}, \end{aligned} \tag{3.31}$$

where

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

$$\widetilde{d}_\theta = 8\widetilde{\kappa}^2 B(4\sqrt{2}\mu L\widetilde{\kappa}\frac{\sqrt{c}C_\phi}{\sqrt{1-\theta}} + (4\sqrt{2}\mu + Lc\widetilde{\kappa}^2)B).$$

In analogy to the argument in the proof of Theorem 3, from the above inequality and (3.31), and noticing $-\frac{\theta}{2} + 2\beta + \delta \le 0$, $\theta + \beta + \delta \le 1$, we have

$$\begin{aligned} R_{t+1} &\le \tfrac{D}{t^\beta}\big[1 - \tfrac{cD}{b_{\theta,\delta}}t^{-\theta-\beta-\delta} + \tfrac{\widetilde{d}_\theta}{D}t^{-\frac{3\theta}{2}+\beta}\big] \\ &= \tfrac{D}{t^\beta}\big[1 - \big(\tfrac{cD}{b_{\theta,\delta}} - \tfrac{\widetilde{d}_\theta}{D}t^{-\frac{\theta}{2}+2\beta+\delta}\big)t^{-\theta-\beta-\delta}\big] \\ &\le \tfrac{D}{t^\beta}\big[1 - \big(\tfrac{cD}{b_{\theta,\delta}} - \tfrac{\widetilde{d}_\theta}{D}\big)t^{-\theta-\beta-\delta}\big] \\ &\le \tfrac{D}{t^\beta}\big[1 - t^{-\theta-\beta-\delta}\big] \le \tfrac{D}{t^\beta}\big[1 - t^{-1}\big] \\ &\le \tfrac{D}{t^\beta}\big[1 - (t+1)^{-1}\big] \le \tfrac{D}{(t+1)^\beta}, \end{aligned} \tag{3.32}$$

where the last to fourth inequality used the fact that $D \ge \widetilde{D}_3 = \frac{2b_{\theta,\delta}}{c} + \widetilde{d}_\theta$, which means that $\frac{cD}{b_{\theta,\delta}} - \frac{\widetilde{d}_\theta}{D} \ge 1$. This completes the proof of the theorem. □

4. Conclusion

In this paper, we considered the unregularized online learning algorithms in the RKHSs for both classification and pairwise learning problems associated with general loss functions. We established their convergence and derived explicit convergence rates with polynomially decaying step sizes. This is in contrast to most of studies which mainly focused on regularized online learning [21,24,29,32]. Our novel results are obtained by using tools from convex analysis, refined properties of Gaussian averages and a simple induction approach. Below, we discuss some directions for future work.

Firstly, the rates for Algorithm 1 and Algorithm 2 are suboptimal. For instance, in the special case of the least-square loss, it was proved in [30] that Algorithm 1 can achieve $\mathcal{O}(T^{-\frac{1}{2}}\ln T)$ if $f_\rho \in \mathcal{H}_G$. However, by Theorem 2, the rate is only of $\mathcal{O}(T^{-\frac{1}{3}})$. It remains an open and challenging question on how to improve the rates for unregularized online learning algorithms with general loss functions. Secondly, our main theorems assume that $g_{\mathcal{H}} = \arg\inf_{g\in\mathcal{H}_G}\mathcal{E}(g)$ and $f_{\mathcal{H}} = \arg\inf_{f\in\mathcal{H}_K}\widetilde{\mathcal{E}}(f)$ exist. However, we know from [30,33] that this assumption can be removed for the least-square loss. It is a clearly important future work to discuss when this assumption can also be removed for general loss functions. Thirdly, the techniques in this paper rely on some smoothness assumptions of the loss function. Consequently, they can not directly be applied to the popular hinge loss. It remains an open question to us how to establish the convergence of unregularized online learning algorithms associated with the hinge loss. Lastly, our results are established in the form of expectation. It would be interesting to prove the almost surely convergence of the last iterate of Algorithms 1 and 2.

Acknowledgments

We would like to thank the referees for their invaluable comments and suggestions. We are also grateful to Dr. Yunwen Lei for improving Lemma 5 in an early version of the paper and providing Lemma 6 to us. The work by D.X. Zhou described in this paper is supported by a grant from the Research Grants Council of Hong Kong [Project No. CityU 104012] and by the National Natural Science Foundation of China under Grant 11461161006.

References

[1] S. Agarwal, P. Niyogi, Generalization bounds for ranking algorithms via algorithmic stability, J. Mach. Learn. Res. 10 (2009) 441–474.

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007

[2] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (1950) 337–404.
[3] F. Bach, E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in: Advances in Neural Information Processing Systems, NIPS, 2011.
[4] P.L. Bartlett, M.I. Jordan, J.D. McAuliffe, Convexity, classification, and risk bounds, J. Amer. Statist. Assoc. 101 (2006) 138–156.
[5] P.L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.
[6] Q. Cao, Z.C. Guo, Y. Ying, Generalization bounds for metric and similarity learning, Mach. Learn. J. (2015), http://dx.doi.org/10.1007/s10994-015-5499-7, in press.
[7] N. Cesa-Bianchi, A. Conconi, C. Gentile, On the generalization ability of on-line learning algorithms, IEEE Trans. Inform. Theory 50 (2004) 2050–2057.
[8] D.R. Chen, Q. Wu, Y. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, J. Mach. Learn. Res. 5 (2004) 1143–1175.
[9] S. Clémencon, G. Lugosi, N. Vayatis, Ranking and empirical minimization of U-statistics, Ann. Statist. 36 (2008) 844–874.
[10] F. Cucker, D.X. Zhou, Learning Theory: An Approximation Theory Viewpoint, Cambridge University Press, 2007.
[11] E. De Vito, A. Caponnetto, L. Rosasco, Model selection for regularized least-squares algorithm in learning theory, Found. Comput. Math. 5 (2005) 59–85.
[12] J. Fan, T. Hu, Q. Wu, D.X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2014.12.005, in press.
[13] T. Hu, J. Fan, Q. Wu, D.-X. Zhou, Learning theory approach to a minimum error entropy criterion, J. Mach. Learn. Res. 14 (2013) 377–397.
[14] T. Hu, J. Fan, Q. Wu, D.X. Zhou, Regularization schemes for minimum error entropy principle, Anal. Appl. (Singap.) 13 (2015) 437–455.
[15] P. Kar, B.K. Sriperumbudur, P. Jain, H.C. Karnick, On the generalization ability of online learning algorithms for pairwise loss functions, in: Proceedings of the 30th International Conference on Machine Learning, ICML, Atlanta, Georgia, USA, 2013.
[16] S. Mukherjee, Q. Wu, Estimation of gradients and coordinate covariation in classification, J. Mach. Learn. Res. 7 (2006) 2481–2514.
[17] S. Mukherjee, D.X. Zhou, Learning coordinate covariances via gradients, J. Mach. Learn. Res. 7 (2006) 519–549.
[18] R. Meir, T. Zhang, Generalization error bounds for Bayesian mixture algorithms, J. Mach. Learn. Res. 4 (2003) 839–860.
[19] W. Rejchel, On ranking and generalization bounds, J. Mach. Learn. Res. 13 (2012) 1373–1392.
[20] S. Mendelson, A few notes on statistical learning theory, in: S. Mendelson, A.J. Smola (Eds.), Advanced Lectures in Machine Learning, in: Lecture Notes in Computer Science, vol. 2600, Springer, 2003, pp. 1–40.
[21] S. Smale, Y. Yao, Online learning algorithms, Found. Comput. Math. 6 (2006) 145–170.
[22] N. Srebro, K. Sridharan, A. Tewari, Smoothness, low-noise, and fast rates, in: Advances in Neural Information Processing Systems, NIPS, 2010.
[23] I. Steinwart, A. Christmann, Support Vector Machines, Springer-Verlag, New York, 2008.
[24] P. Tarres, Y. Yao, Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence, IEEE Trans. Inform. Theory 60 (2014) 5716–5735.
[25] Richard A. Vitale, Some comparisons for Gaussian processes, Proc. Amer. Math. Soc. (2000) 3043–3046.
[26] K.Q. Weinberger, L. Saul, Distance metric learning for large margin nearest neighbour classification, J. Mach. Learn. Res. 10 (2009) 207–244.
[27] Y. Wang, R. Khardon, D. Pechyony, R. Jones, Generalization bounds for online learning algorithms with pairwise loss functions, in: Proceedings of the 25th Annual Conference on Learning Theory, COLT, 2012.
[28] Q. Wu, Y. Ying, D.X. Zhou, Multi-kernel regularized classifiers, J. Complexity 23 (2007) 108–134.
[29] G.B. Ye, D.X. Zhou, Fully online classification by regularization, Appl. Comput. Harmon. Anal. 23 (2007) 198–214.
[30] Y. Ying, M. Pontil, Online gradient descent algorithms, Found. Comput. Math. 5 (2008) 561–596.
[31] Y. Ying, Q. Wu, C. Campbell, Learning the coordinate gradients, Adv. Comput. Math. 37 (2012) 355–378.
[32] Y. Ying, D.X. Zhou, Online regularized classification algorithms, IEEE Trans. Inform. Theory 11 (2006) 4775–4788.
[33] Y. Ying, D.X. Zhou, Online pairwise learning algorithms with kernels, arXiv preprint, available on http://arxiv.org/abs/1502.07229, 2015.
[34] T. Zhang, Statistical behavior and consistency of classification methods based on convex risk minimization, Ann. Statist. 32 (2004) 56–85.
[35] P. Zhao, S.C.H. Hoi, R. Jin, T. Yang, Online AUC maximization, in: Proceedings of the 28th International Conference on Machine Learning, ICML, Bellevue, Washington, USA, 2011.

Please cite this article in press as: Y. Ying, D.-X. Zhou, Unregularized online learning algorithms with general loss functions, Appl. Comput. Harmon. Anal. (2015), http://dx.doi.org/10.1016/j.acha.2015.08.007