Hindawi Publishing Corporation
Journal of Applied Mathematics
Volume 2012, Article ID 189753, 13 pages
doi:10.1155/2012/189753

Research Article

Approximation Analysis of Gradient Descent Algorithm for Bipartite Ranking

Hong Chen,[1] Fangchao He,[2,3] and Zhibin Pan[1]

[1] *College of Science, Huazhong Agricultural University, Wuhan 430070, China*
[2] *School of Science, Hubei University of Technology, Wuhan 430068, China*
[3] *Faculty of Mathematics and Computer Science, Hubei University, Wuhan 430062, China*

Correspondence should be addressed to Hong Chen, chenhongml@163.com

Received 9 March 2012; Revised 12 May 2012; Accepted 26 May 2012

Academic Editor: Yuesheng Xu

We introduce a gradient descent algorithm for bipartite ranking with general convex losses. The implementation of this algorithm is simple, and its generalization performance is investigated. Explicit learning rates are presented in terms of the suitable choices of the regularization parameter and the step size. The result fills the theoretical gap in learning rates for ranking problem with general convex losses.

1. Introduction

In this paper we consider a gradient descent algorithm for bipartite ranking generated from Tikhonov regularization scheme with general convex losses and reproducing kernel Hilbert spaces (RKHS).

Let $\mathcal{X}$ be a compact metric space and $\mathcal{Y} = \{-1, 1\}$. In bipartite ranking problem, the learner is given positive samples $S^+ = \{x_i^+\}_{i=1}^m$ and negative samples $S^- = \{x_i^-\}_{i=1}^n$, which are randomly independent drawn from ρ^+ and ρ^-, respectively. Given training set $S := (S^+, S^-)$, the goal of bipartite ranking is to learn a real-valued ranking function $f : \mathcal{X} \to \mathbb{R}$ that ranks future positive samples higher than negative ones.

The expected loss incurred by a ranking function f on a pair of instances (x^+, x^-) is $I_{\{f(x^+)-f(x^-)\le 0\}}$, where $I_{\{t\}}$ is 1 if t is true and 0 otherwise. However, due to the nonconvexity of I, the empirical minimization method based on I is NP-hard. Thus, we consider replacing I by a convex upper loss function $\phi(f(x^+) - f(x^-))$. Typical choices of ϕ include the hinge loss, the least square loss, and the logistic loss.

The expected convex risk is

$$\mathcal{E}(f) = \iint_{\mathcal{X}} \phi(f(x^+) - f(x^-)) d\rho^+(x^+) d\rho^-(x^-). \tag{1.1}$$

The corresponding empirical risk is

$$\mathcal{E}_S(f) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi\left(f(x_i^+) - f\left(x_j^-\right)\right). \tag{1.2}$$

Let $\vartheta = \{f \in \mathfrak{F} : f = \arg\min_{f\in\mathfrak{F}} \mathcal{E}(f)\}$ be the target function set, where $\mathfrak{F}$ is the measurable function space. We can observe that the target function is not unique. In particular, for the least square loss, the regression function is one element in this set.

The ranking algorithm we investigate in this paper is based on a Tikhonov regularization scheme associated with a Mercer kernel. We usually call a symmetric and positive semidefinite continuous function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a Mercer kernel. The RKHS $\mathcal{H}_K$ associated with the kernel K is defined (see [1]) to be the closure of the linear span of the set of functions $\{K_x := K(x, \cdot) : x \in \mathcal{X}\}$ with the inner product $\langle \cdot \rangle_K$ given by $\langle K_x, K_{x'} \rangle_K = K(x, x')$. The reproducing property takes the form $f(x) = \langle f, K_x \rangle_K$, for all $x \in \mathcal{X}, f \in \mathcal{H}_K$. The reproducing property with the Schwartz inequality yields that $|f(x)| \le \sqrt{K(x,x)} \|f\|_{\mathcal{H}_K}$. Then, $\|f\|_\infty \le \kappa \|f\|_{\mathcal{H}_K}$, where $\kappa := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)}$.

The regularized ranking algorithm is implemented by an offline regularization scheme [2] in $\mathcal{H}_K$

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \mathcal{E}_S(f) + \lambda \|f\|_{\mathcal{H}_K}^2 \right\}, \tag{1.3}$$

where $\lambda > 0$ is the regularization parameter. A data free-limit of (1.3) is

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

Though the offline algorithm (1.3) has been well understood in [2], it might be practically challenging when the sample size m or n is large. The same difficulty for classification and regression algorithms is overcome by reducing the computational complexity through a stochastic gradient descent method. Such algorithms have been proposed for online regression in [3, 4], online classification in [5, 6], and gradient learning in [7, 8]. In this paper, we use the idea of gradient descent to propose an algorithm for learning a target function in ϑ.

Since ϕ is convex, we know that its left derivative ϕ'_- is well defined and nondecreasing on $\mathbb{R}$. By taking functional derivatives in (1.3), we introduce the following algorithm for ranking.

Definition 1.1. The stochastic gradient descent ranking algorithm is defined for the sample S by $f_1^S = 0$ and

$$f_{t+1}^S = (1 - \eta_t\lambda) f_t^S - \frac{\eta_t}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \phi'_- \left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right)\left(K_{x_i^+} - K_{x_j^-}\right), \tag{1.5}$$

where $t \in \mathbb{N}$ and $\{\eta_t\}$ is the sequence of step sizes.

In fact, Burges et al. [9] investigate gradient descent methods for learning ranking functions and introducing a neural network to model the underlying ranking function. From the idea of maximizing the generalized Wilcoxon-Mann-Whitney statistic, a ranking algorithm using gradient approximation has been proposed in [10]. However, these approaches are different from ours and their analysis focuses on computational complexity. Recently, for least square loss, numerical experiments by gradient descent algorithm have been presented in [11]. The aim of this paper is to provide generalization bounds for the gradient descent ranking algorithm (1.5) with general convex losses. To the best of our knowledge, there is no error analysis in this case; This is why we conduct our study in this paper.

We mainly analyze the errors $\|f_t^S - f_\lambda\|_{\mathscr{H}_K}$ and $\inf_{f\in\mathfrak{H}}\|f_t^S - f\|_{\mathscr{H}_K}$, which is different from previous error analysis for ranking algorithms based on uniform convergence (e.g., [12–16]) and stability analysis in [2, 17, 18]. Though the convergence rates of $\mathscr{H}_K$ norm for classification and regression algorithms have been elegantly investigated in [19, 20], there is no such analysis in the ranking setting. The main difference in the formulation of the ranking problem as compared to the problems of classification and regression is that the performance or loss in ranking is measured on pairs of examples, rather than on individual examples. This means in particular that, unlike the empirical error in classification or regression, the empirical error in ranking cannot be expressed as a sum of independent random variables [17]. This makes the convergence analysis of $\mathscr{H}_K$ norm difficult and previous techniques invalid. Fortunately, we observe that similar difficulty for gradient learning has been well overcome in [7, 21, 22] for gradient learning by introducing some novel techniques. In this paper, we will develop an elaborative analysis in terms of these analysis techniques.

2. Main Result

In this section we present our main results on learning rates of algorithm (1.5) for learning ranking functions. We assume that $\phi \in C^1(\mathbb{R})$ satisfies

$$\left|\phi(u)\right| \le C_0(1 + |u|)^q, \quad \left|\phi'(u)\right| \le C_0(1 + |u|)^{q-1}, \quad \forall u \in \mathbb{R} \tag{2.1}$$

for some $C_0 > 0$ and $q \ge 1$. Denote the constant

$$\Delta_* = 1 + 4\kappa^2\left(\sup_{|u|\le 1}\frac{\left|\phi'_-(u) - \phi'(0)\right|}{|u|} + C_0\phi'(0) + \left(8\kappa^2\phi'(0)\right)^{q-1}\right). \tag{2.2}$$

Table 1: The values of parameters for different convex losses.

Loss function	C_0	q	Δ_*	α
$\phi(t) = t^2$	1	2	$1 + 12\kappa^2$	$\theta \min\{(1/2)(\theta - 3\gamma), \gamma\beta\}$
$\phi(t) = \log(1 + e^{-t})$	2	1	$1 + 11\kappa^2$	$(\gamma + \theta) \min\{\theta - \gamma, \gamma - \beta\}$
$\phi(t) = (1 - t)_+$	1	1	$1 + 8\kappa^2$	$\theta \min\{(1/2)(\theta - 3\gamma), \gamma\beta\}$

Theorem 2.1. *Assume ϕ satisfies* (2.1)*, and choose the step size as*

$$\eta_t = \eta_* \lambda^{\max\{q-2,0\}} t^{-\theta} \text{ for some } 0 < \theta < 1,\ 0 < \eta_* \le \frac{1}{\Delta_*}. \tag{2.3}$$

For $0 < \gamma < (1 - \theta)/\min\{q - 1, 1\}, s > 0$, one takes $\lambda = t^{-\gamma}$ with $(mn/(m+n)^{3/2})^s \le t \le 2(mn/(m+n)^{3/2})^s$. Then, for any $0 < \delta < 1$, with confidence at least $1 - \delta$, one has

$$\left\| f_t^S - f_\lambda \right\|_{\mathscr{H}_K}^2 \le C \left(\frac{mn}{(m+n)^{3/2}} \right)^{-\alpha}, \tag{2.4}$$

where C is a constant independent of m, n, and

$$\alpha = \min\{s\theta - s\gamma \min\{q + 1, 2q - 1\}, 1 - s\gamma(1 + q)\}. \tag{2.5}$$

Theorem 2.1 will be proved in the next section where the constant C can be obtained explicitly. The explicit parameters in Theorem 2.1 are described in Table 1 for some special loss functions ϕ. Note that the iteration steps and iterative numbers depend on sample number m, n. When $m = O(n)$ and $m \to \infty$, we have $t \to \infty$ and $\eta_t \to 0$.

From the results in Theorem 2.1, we know that the balance of samples is crucial to reach fast learning rates. For $m = O(n)$ and the least square loss, the approximation order is $O(m^{(-1/2)\min\{s\theta - 3s\gamma, 1 - 3s\gamma\}})$. Moreover, when $s\theta \to 1$ and $s\gamma \to 0$, we have $\|f_t^S - f_\lambda\|_{\mathscr{H}_K}^2 \to 0$ with the order $O(m^{(-1/2)})$.

Now we present the estimates of $\inf_{f \in \vartheta} \|f_t^S - f\|_{\mathscr{H}_K}$ under some approximation conditions.

Corollary 2.2. *Assume that there is $f^* \in \vartheta$ such that $\|f_\lambda - f^*\|_{\mathscr{H}_K}^2 \le C_\beta \lambda^\beta$ for some $0 < \beta < 1$. Under the condition in Theorem 2.1, for any $0 < \delta < 1$, with confidence at least $1 - \delta$, one has*

$$\inf_{f \in \vartheta} \left\| f_t^S - f \right\|_{\mathscr{H}_K}^2 \le \widetilde{C} \left(\frac{mn}{(m+n)^{3/2}} \right)^{-\tilde{\alpha}}, \tag{2.6}$$

where $\widetilde{C}$ is a constant independent of m, n, and

$$\tilde{\alpha} = \min\{s\theta - s\gamma \min\{q + 1, 2q - 1\}, 1 - s\gamma(1 + q), s\gamma\beta\}. \tag{2.7}$$

For $m = O(n)$ and the least square loss, by setting $s = 1/\gamma(3+\beta)$, we can derive the learning rate $O(m^{(-1/\gamma(6+2\beta))\min\{\theta-3\gamma,\gamma\beta\}})$. Moreover, if $\beta < (\theta-3\gamma)/\gamma$, we get the approximation order $O(m^{-(\beta)/(6+2\beta)})$.

For the least square loss, the regression function is an optimal predictor in ϑ. Then, the bipartite ranking problem can be reduced as a regression problem. Based on the theoretical analysis in [19, 20], we know that the approximation condition in Corollary 2.2 can be achieved when the regression function lies in the $(\beta+1)/2$th power of the integral operator associated with the kernel K.

The highlight of our theoretical analysis results is to provide the estimate of the distance between f_t^S and the target function set ϑ in RKHS. This is different from the previous result on error analysis that focuses on establishing the estimate of $|\mathcal{E}(f) - \mathcal{E}_S(f)|$. Compared with the previous theoretical studies, the approximation analysis in $\mathcal{H}_K$-norm is new and fills the gap on learning rates for ranking problem with general convex losses.

We also note that the techniques of previous error estimate for ranking problem mainly include stability analysis in [2, 17], concentration estimation based on U-statistics in [14], and uniform convergence bounds based on covering numbers [15, 16]. Our analysis presents a novel capacity-independent procedure to investigate the generalization performance of ranking algorithms.

3. Proof of Main Result

We introduce a special property of $\mathcal{E}(f) + (\lambda/2)\|f\|_{\mathcal{H}_K}^2$. Since the proof is the same as that in [5], we will omit it here.

Lemma 3.1. *Let $\lambda > 0$. For any $f \in \mathcal{H}_K$, there holds*

$$\frac{\lambda}{2}\|f - f_\lambda\|_{\mathcal{H}_K} \le \left\{\mathcal{E}(f) + \frac{\lambda}{2}\|f\|_{\mathcal{H}_K}^2\right\} - \left\{\mathcal{E}(f_\lambda) + \frac{\lambda}{2}\|f_\lambda\|_{\mathcal{H}_K}^2\right\}. \tag{3.1}$$

Denote

$$\begin{aligned} f_t^\lambda &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi_-'\left(f(x_i^+) - f\left(x_j^-\right)\right)\left(K_{x_i^+} - K_{x_j^-}\right) + \lambda f_t^S, \\ f_{t+1}^S &= f_t^S - \eta_t f_t^\lambda. \end{aligned} \tag{3.2}$$

Now we give the one-step analysis.

Lemma 3.2. *For $t \ge 0$, one has*

$$\left\|f_{t+1}^S - f_\lambda\right\|_{\mathcal{H}_K}^2 \le (1-\eta_t\lambda)\left\|f_t^S - f_\lambda\right\|_{\mathcal{H}_K}^2 + \eta_t^2\left\|f_t^\lambda\right\|_{\mathcal{H}_K}^2 + 2\eta_t\varphi(S,t), \tag{3.3}$$

where $\varphi(S,t) = \mathcal{E}_S(f_\lambda) - \mathcal{E}(f_\lambda) + \mathcal{E}(f_t^S) - \mathcal{E}_S(f_t^S)$.

Proof. Observe that

$$\left\|f_{t+1}^S - f_\lambda\right\|_{\mathscr{H}_K}^2 = \left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2 + \eta_t^2\left\|f_t^\lambda\right\|_{\mathscr{H}_K}^2 + 2\eta_t\left\langle f_\lambda - f_t^S, f_t^\lambda\right\rangle_K. \tag{3.4}$$

Note that

$$\begin{aligned}\left\langle f_\lambda - f_t^S, f_t^\lambda\right\rangle_K &= \frac{1}{mn}\sum_{i=1}^m\sum_{j=1}^n \phi'_-\left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right)\left(f_\lambda(x_i^+) - f_\lambda\left(x_j^-\right)\right.\\ &\qquad \left. - f_t^S(x_i^+) + f_t^S\left(x_j^-\right)\right) + \lambda\left\langle f_\lambda - f_t^S, f_t^S\right\rangle_K\\ &\le \mathscr{E}_S(f_\lambda) - \mathscr{E}_S\left(f_t^S\right) - \lambda\left\|f_t^S\right\|_{\mathscr{H}_K}^2 + \lambda\left\langle f_\lambda, f_t^S\right\rangle_K\\ &\le \left\{\mathscr{E}_S(f_\lambda) + \frac{\lambda}{2}\|f_\lambda\|_{\mathscr{H}_K}^2\right\} - \left\{\mathscr{E}_S\left(f_t^S\right) + \frac{\lambda}{2}\left\|f_t^S\right\|_{\mathscr{H}_K}^2\right\},\end{aligned} \tag{3.5}$$

where the first and the second inequalities are derived by the convexity of ϕ and the Schwartz inequality, respectively.

By Lemma 3.1, we know that

$$\begin{aligned}&\left\{\mathscr{E}_S(f_\lambda) + \frac{\lambda}{2}\|f_\lambda\|_{\mathscr{H}_K}^2\right\} - \left\{\mathscr{E}_S\left(f_t^S\right) + \frac{\lambda}{2}\left\|f_t^S\right\|_{\mathscr{H}_K}^2\right\}\\ &\quad\le \left\{\mathscr{E}_S(f_\lambda) - \mathscr{E}(f_\lambda) + \mathscr{E}\left(f_t^S\right) - \mathscr{E}_S\left(f_t^S\right)\right\} - \frac{\lambda}{2}\left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2.\end{aligned} \tag{3.6}$$

Thus, the desired result follows by combining (3.5) and (3.6) with (3.4). □

To deal with the sample error iteratively by applying (3.3), we need to bound the quantity $\varphi(S,t)$ by the theory of uniform convergence. To this end, a bound for the norm of f_t^S is required.

Definition 3.3. One says that ϕ'_- is locally Lipschitz at the origin if the local Lipschitz constant

$$M(\lambda) = \sup\left\{\frac{\left|\phi'_-(u) - \phi'_-(0)\right|}{|u|} : |u| \le \frac{4\kappa^2\left|\phi'_-(0)\right|}{\lambda}\right\} \tag{3.7}$$

is finite for any $\lambda > 0$.

Now we estimate the bound of f_t^S from the ideas given in [5].

Lemma 3.4. *Assume that ϕ'_- is locally Lipschitz at the origin. If the step size η_t satisfies $\eta_t(4\kappa^2 M(\lambda)+\lambda) \le 1$ for each t, then $\|f_t^S\|_{\mathscr{H}_K} \le 2\kappa|\phi'_-(0)|/\lambda$.*

Proof. We prove by induction. It is trivial that $f_1^S = 0$ satisfies the bound.

Suppose that this bound holds true for f_t^S, $\|f_t^S\|_{\mathscr{H}_K} \le 2\kappa|\phi'_-(0)|/\lambda$. Consider

$$\begin{aligned} f_{t+1}^S &= (1-\eta_t\lambda)f_t^S - \frac{\eta_t}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi'_-\left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right)\left(K_{x_i^+} - K_{x_j^-}\right) \\ &= (1-\eta_t\lambda)f_t^S - \frac{\eta_t}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\phi'_-\left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right) - \phi'_-(0)}{f_t^S(x_i^+) - f_t^S\left(x_j^-\right)} \\ &\quad \cdot \left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right)\left(K_{x_i^+} - K_{x_j^-}\right) - \frac{\eta_t}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\phi'_-(0)\left(K_{x_i^+} - K_{x_j^-}\right). \end{aligned} \tag{3.8}$$

Let $L_{ij}f = \langle f, K_{x_i^+} - K_{x_j^-}\rangle_K (K_{x_i^+} - K_{x_j^-})$. Since

$$0 \le \langle L_{ij}f, f\rangle_K = \left|\left\langle f, K_{x_i^+} - K_{x_j^-}\right\rangle_K\right|^2 \le 4\kappa^2\|f\|_{\mathscr{H}_K}^2, \tag{3.9}$$

we have $\|L_{ij}\| \le 4\kappa^2$.

Meanwhile, $(\phi'_- f_t^S(x_i^+) - f_t^S(x_j^-) - \phi'_-(0)/f_t^S(x_i^+) - f_t^S(x_j^-) \le M(\lambda))$. Then,

$$\frac{\eta_t}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\phi'_-\left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right) - \phi'_-(0)}{f_t^S(x_i^+) - f_t^S\left(x_j^-\right)}L_{ij} \tag{3.10}$$

is a positive linear operator on $\mathscr{H}_K$ and its norm is bounded by $4\kappa^2 M(\lambda)$.

Since $\eta_t(4\kappa^2 M(\lambda) + \lambda) \le 1$, the operator

$$A := (1-\eta_t\lambda)I - \frac{\eta_t}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\phi'_-\left(f_t^S(x_i^+) - f_t^S\left(x_j^-\right)\right) - \phi'_-(0)}{f_t^S(x_i^+) - f_t^S\left(x_j^-\right)}L_{ij} \tag{3.11}$$

on $\mathscr{H}_K$ is positive and $A \le (1-\eta_t\lambda)I$.

Thus,

$$\begin{aligned} \left\|f_{t+1}^S\right\|_{\mathscr{H}_K} &\le (1-\eta_t\lambda)\left\|f_t^S\right\|_{\mathscr{H}_K} + \frac{\eta_t}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|\phi'_-(0)\right|\left\|K_{x_i^+} - K_{x_j^-}\right\|_{\mathscr{H}_K} \\ &= \frac{2\kappa\left|\phi'_-(0)\right|}{\lambda}. \end{aligned} \tag{3.12}$$

This proves the lemma. □

For $r > 0$, denote $\mathfrak{F}_r = \{f \in \mathscr{H}_K : \|f\|_{\mathscr{H}_K} \le r\}$. Meanwhile, denote $L_r = \max\{|\phi'_-(2\kappa r)|, |\phi'_-(-2\kappa r)|\}$ and $M_r = \max\{|\phi(2\kappa r)|, |\phi(-2\kappa r)|\}$.

Based on analysis techniques in [21, 23], we derive the capacity-independent bounds for $W(S,r) := \sup_{f\in\mathfrak{F}_r}|\mathcal{E}_S(f) - \mathcal{E}(f)|$.

Lemma 3.5. *For every $r > 0$ and $\varepsilon > 0$, one has*

$$\begin{aligned}\mathrm{Prob}_S\{|W(S,r) - EW(S,r)| > \varepsilon\} &\le \exp\left\{-\frac{2m^2n^2\varepsilon^2}{(m+n)^3M_r^2}\right\},\\ EW(S,r) &\le \left(4L_r\kappa r + 2\phi(0)\right)\frac{\sqrt{m}+\sqrt{n}}{\sqrt{mn}}.\end{aligned} \tag{3.13}$$

Proof. Because of the feature of S, four cases of samples change should be taken into account to use McDiarmid's inequality. Denote by S_k the sample coinciding with S except for x_k^+ (or x_k^-) replaced by $\widetilde{x}_k^+$ (or $\widetilde{x}_k^-$). It is easy to verify that

$$\begin{aligned}|W(S,r) - W(S_k,r)| &= \left|\sup_{f\in\mathfrak{F}_r}\left|\mathcal{E}_S(f) - \mathcal{E}(f)\right| - \sup_{f\in\mathfrak{F}_r}\left|\mathcal{E}_{S_k}(f) - \mathcal{E}(f)\right|\right|\\ &\le \sup_{f\in\mathfrak{F}_r}\left|\mathcal{E}_S(f) - \mathcal{E}_{S_k}(f)\right| \le \frac{m+n}{mn}M_r.\end{aligned} \tag{3.14}$$

Based on MicDiarmid's inequality in [24], we can derive the first result in Lemma 3.5. To derive the second result, we denote $\xi(x^+,u^-) = \phi(f(x^+)-f(u^-))$. Then, $\mathcal{E}(f) = E_{x^+}E_{x^-}\xi(x^+,x^-)$ and $\mathcal{E}_S(f) = 1/mn\sum_{i=1}^m\sum_{j=1}^n\xi(x_i^+,x_j^-)$. Observe that

$$\begin{aligned}W(S,r) &\le \sup_{f\in\mathfrak{F}_r}\left|\mathcal{E}(f) - \frac{1}{n}\sum_{j=1}^n E_{x^+}\xi\left(x^+,x_j^-\right)\right| + \sup_{f\in\mathfrak{F}_r}\left|\frac{1}{n}\sum_{j=1}^n E_{x^+}\xi\left(x^+,x_j^-\right) - \mathcal{E}_S(f)\right|\\ &\le E_{x^+}\sup_{f\in\mathfrak{F}_r}\left|E_{x^-}\xi(x^+,x^-) - \frac{1}{n}\sum_{j=1}^n\xi\left(x^+,x_j^-\right)\right|\\ &\quad + \frac{1}{n}\sum_{j=1}^n\sup_{f\in\mathfrak{F}_r}\sup_{x^-}\left|E_{x^+}\xi(x^+,x^-) - \frac{1}{m}\sum_{i=1}^m\xi(x_i^+,x^-)\right|\\ &= W_1 + W_2.\end{aligned} \tag{3.15}$$

Denote $G_{x^+} = \{h(x^-) = f(x^+) - f(x^-) : f\in\mathfrak{F}\}$. Then,

$$EW_1 = E_xE\sup_{h\in G_{x^+}}\left|E_{x^-}\phi(h(x^-)) - \frac{1}{n}\sum_{j=1}^n\phi\left(h\left(x_j^-\right)\right)\right| \le 2\sup_{x^+}E\sup_{h\in G_{x^+}}\left|\frac{1}{n}\sum_{j=1}^n\varepsilon_j\phi\left(h\left(x_j^-\right)\right)\right|. \tag{3.16}$$

Since $\phi(h(x_j^-)) - \phi(0) \le L_r(h(x_j^-) - 0)$, we have

$$\begin{aligned} EW_1 &\le 2L_r \sup_{x^+} E \sup_{h\in G_{x^+}} \left| \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j h\left(x_j^-\right) \right| + \frac{2\phi(0)}{n} E\left|\sum_{j=1}^{n} \varepsilon_j\right| \\ &\le \frac{4L_r\kappa r}{n} + \frac{2\phi(0)}{n} E\left|\sum_{j=1}^{n} \varepsilon_j\right| \\ &\le \frac{4L_r\kappa r + 2\phi(0)}{n}\left(E\left|\sum_{j=1}^{n} \varepsilon_j\right|^2\right)^{1/2} \le \frac{4L_r\kappa r + 2\phi(0)}{n}\left(E\sum_{j,j'=1}^{n} \varepsilon_j\varepsilon_{j'}\right)^{1/2} \\ &= \frac{4L_r\kappa r + 2\phi(0)}{\sqrt{n}}. \end{aligned} \tag{3.17}$$

With the same fashion, we can also derive

$$EW_2 \le \frac{4L_r\kappa r + 2\phi(0)}{\sqrt{m}}. \tag{3.18}$$

Thus, the second desired result follows by combining (3.17) and (3.18). □

Now we can derive the estimate of $\varphi(S,t)$.

Lemma 3.6. *If η_t satisfies (1) for each t and $r = 2\kappa|\phi'_-(0)|/\lambda + \sqrt{2\phi(0)/\lambda}$, then with confidence at least $1-\delta$ one has*

$$\varphi(S,t) \le B_\lambda := \left(4L_r\kappa r + 2\phi(0) + M_r\sqrt{2\log\left(\frac{2}{\delta}\right)}\right)\frac{(m+n)^{3/2}}{mn}. \tag{3.19}$$

Proof. By Lemma 3.5, we have, with confidence at least $1-\delta$,

$$W(S,r) \le \left(4L_r\kappa r + 2\phi(0)\right)\frac{\sqrt{m}+\sqrt{n}}{\sqrt{mn}} + \frac{(m+n)^{3/2}M_r}{2mn}\sqrt{2\log\left(\frac{2}{\delta}\right)}. \tag{3.20}$$

By taking $f = 0$ in the definition of f_λ, we see that

$$\frac{\lambda}{2}\|f_\lambda\|_{\mathcal{H}_K}^2 \le \mathcal{E}(0) + 0 \le \phi(0). \tag{3.21}$$

Then, for any $\lambda > 0$, we have $\|f_\lambda\|_{\mathcal{H}_K} \le \sqrt{2\phi(0)/\lambda}$. Thus, $f_t^S, f_\lambda \in \mathfrak{F}_r$ for $r = 2\kappa|\phi'_-(0)|/\lambda + \sqrt{2\phi(0)/\lambda}$. So, $\varphi(S,t) \le 2W(S,r)$ for each t. This completes the proof. □

We are in a position to give bounds for the sample error. We need the following elementary inequalities that can be found in [3, 5].

Lemma 3.7. *(1) For $\alpha \in (0,1]$ and $\theta \in [0,1]$,*

$$\sum_{i=1}^{t} \frac{1}{i^{\theta}} \prod_{j=i+1}^{t} \left(1 - \frac{\alpha}{j^{\theta}}\right) \leq \frac{3}{\alpha}. \tag{3.22}$$

(2) Let $v \in (0,1]$ and $\theta \in (0,1]$. Then

$$\sum_{t=1}^{T-1} \frac{1}{t^{2\theta}} \exp\left\{-v \sum_{j=t+1}^{T} j^{-\theta}\right\} \leq \begin{cases} \dfrac{18}{vT^{\theta}} + \dfrac{9T^{1-\theta}}{(1-\theta)2^{1-\theta}} \exp\left\{-\dfrac{v(1-2^{\theta-1})}{1-\theta}(T+1)^{1-\theta}\right\} & \text{if } \theta < 1, \\ \dfrac{8}{1-v}(T+1)^{-v} & \text{if } \theta = 1. \end{cases} \tag{3.23}$$

(3) For any $t < T$ and $\theta \in (0,1]$, there holds

$$\sum_{j=t+1}^{T} j^{-\theta} \leq \begin{cases} \dfrac{1}{1-\theta}\left[(T+1)^{1-\theta} - (t+1)^{1-\theta}\right] & \text{if } \theta < 1, \\ \log(T+1) - \log(t+1) & \text{if } \theta = 1. \end{cases} \tag{3.24}$$

Proposition 3.8. *Let $\eta_t = \eta_1 t^{-\theta}$ for some $\theta \in [0,1]$, and let η_1 satisfy $\eta_1(4M(\lambda) + \lambda) \leq 1$. Set r and B_λ as in Lemma 3.6. Denote $\widetilde{B}_\lambda = 2\kappa L_r + \kappa|\phi'_-(0)|$. Then, with confidence at least $1-\delta$, the following bound holds for $t \geq 1$: when $\theta < 1$,*

$$\begin{aligned} \left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2 \leq \left\|f_\lambda\right\|_{\mathscr{H}_K}^2 \exp\left\{-\frac{\eta_1\lambda}{1-\theta}\left(t^{1-\theta}-1\right)\right\} + \frac{18\widetilde{B}_\lambda^2\eta_1}{\lambda t^{\theta}} \\ + \frac{2\widetilde{B}_\lambda^2\eta_1^2 t^{1-\theta}}{(1-\theta)2^{1-\theta}} \exp\left\{-\frac{\eta_1\lambda(1-2^{\theta-1})}{1-\theta}(t+1)^{1-\theta}\right\} + \frac{6B_\lambda}{\lambda}, \end{aligned} \tag{3.25}$$

when $\theta = 1$,

$$\left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2 \leq \left\|f_\lambda\right\|_{\mathscr{H}_K}^2 t^{-\eta_1\lambda} + \frac{8\widetilde{B}_\lambda^2\eta_1^2}{1-\eta_1\lambda}(t+1)^{-\eta_1\lambda} + \frac{6B_\lambda}{\lambda}. \tag{3.26}$$

Proof. Since $\|f_t^S\|_{\mathscr{H}_K} \leq 2\kappa|\phi'_-(0)|/\lambda$, we have $|f_t^S(x_i^+) - f_t^S(x_j^-)| \leq 2\kappa\|f_t^S\|_{\mathscr{H}_K} \leq 2\kappa r$. From the definition of f_t^λ, we know that $\|f_t^\lambda\|_{\mathscr{H}_K} \leq 2\kappa L_r + \kappa|\phi'_-(0)| = \widetilde{B}_\lambda$. Thus, when $\varphi(S,t) \leq B_\lambda$, we have from Lemma 3.2

$$\left\|f_{t+1}^S - f_\lambda\right\|_{\mathscr{H}_K}^2 \leq (1-\eta_t\lambda)\left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2 + \eta_t^2\widetilde{B}_\lambda^2 + 2\eta_t B_\lambda. \tag{3.27}$$

Applying this relation iteratively, we have

$$\left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2 \leq \prod_{i=1}^{t-1}(1-\eta_i\lambda)\left\|f_\lambda\right\|_{\mathscr{H}_K}^2 + \sum_{i=1}^{t-1}\prod_{j=i+1}^{t-1}(1-\eta_i\lambda)\left(\eta_i^2\widetilde{B}_\lambda^2 + 2\eta_i B_\lambda\right). \tag{3.28}$$

Since $\eta_i = \eta_1 i^{-\theta}$, by Lemma 3.7(2), we have for $\theta < 1$

$$\begin{aligned}\sum_{i=1}^{t-1}\prod_{j=i+1}^{t-1}(1-\eta_i\lambda)\eta_i^2 &\le \eta_1^2\sum_{i=1}^{t-1}\frac{1}{2\theta}\exp\left\{-\eta_1\lambda\sum_{j=i+1}^{t-1}j^{-\theta}\right\}\\ &\le \frac{18\eta_1}{\lambda t}+\frac{9\eta_1^2 t^{1-\theta}}{(1-\theta)2^{1-\theta}}\exp\left\{-\frac{\eta_1\lambda(1-2^{\theta-1})}{1-\theta}(t+1)^{1-\theta}\right\}\end{aligned} \tag{3.29}$$

and for $\theta = 1$

$$\sum_{i=1}^{t-1}\prod_{j=i+1}^{t-1}(1-\eta_i\lambda)\eta_i^2 \le \frac{8\eta_1^2}{1-\eta_1\lambda}(t+1)^{-\eta_1\lambda}. \tag{3.30}$$

Lemma 3.7(1) yields

$$\sum_{i=1}^{t-1}\prod_{j=i+1}^{t-1}(1-\eta_i\lambda)\eta_i \le \eta_1\sum_{i=1}^{t-1}\frac{1}{i^\theta}\prod_{j=i+1}^{t-1}\left(1-\frac{\eta_1\lambda}{j^\theta}\right)\le\frac{3}{\lambda}. \tag{3.31}$$

By Lemma 3.7(3), we also have for $\theta < 1$

$$\sum_{i=1}^{t-1}(1-\eta_i\lambda)\le\exp\left\{-\sum_{i=1}^{t-1}\eta_i\lambda\right\}\le\exp\left\{\frac{\eta_1\lambda}{1-\theta}\left(1-t^{1-\theta}\right)\right\} \tag{3.32}$$

and for $\theta = 1$

$$\sum_{i=1}^{t-1}(1-\eta_i\lambda)\le t^{-\eta_1\lambda}. \tag{3.33}$$

Combining the above estimations with Lemma 3.6, we derive the desired results. □

Now we present the proof of Theorem 2.1.

Proof of Theorem 2.1. First we derive explicit expressions for the quantities in Proposition 3.8. Since $\lambda = t^{-\gamma}$, we have $r \le C_3 t^\gamma$, where $C_3 = 2\kappa|\phi'_-(0)| + \sqrt{2\phi(0)}$. By (2.1), we find that

$$\begin{aligned}L_r &\le C_0(1+2\kappa r)^{q-1}\le C_0(1+2\kappa)^{q-1}C_3^{q-1}t^{(q-1)\gamma},\\ M_r &\le C_0(1+2\kappa)^q C_3^q t^{q\gamma}.\end{aligned} \tag{3.34}$$

Then,

$$B_\lambda \le C_4\sqrt{\log\left(\frac{2}{\delta}\right)}t^{q\gamma}\frac{(m+n)^{3/2}}{mn}, \qquad \widetilde{B}_\lambda \le C_5 t^{(q-1)\gamma}, \tag{3.35}$$

where $C_4 = 4C_0\kappa(1+2\kappa)^{q-1}C_3^{q-1} + C_0(1+2\kappa)^q C_3^q + 2\phi(0)$ and $C_5 = 2\kappa C_0(1+2\kappa)^{q-1}C_3^{q-1} + 2\phi(0)$.

Next, we bound $M(\lambda)$. When $1 \le |u| \le 4\kappa^2|\phi'_-(0)|/\lambda$, we have

$$\frac{\left|\phi'_-(u) - \phi'_-(0)\right|}{|u|} \le \left|\phi'_-(0)\right| + 2^{q-1}C_0\left(4\kappa^2\phi'_-(0)\right)^{q-1}\lambda^{2-q}. \tag{3.36}$$

Hence,

$$M(\lambda) \le \left(\sup_{|u|\le 1}\frac{\left|\phi'_-(u) - \phi'_-(0)\right|}{|u|} + \left|\phi'_-(0)\right| + 2^{q-1}C_0\left(4\kappa^2\phi'_-(0)\right)^{q-1}\right)\lambda^{\min\{2-q,0\}}. \tag{3.37}$$

It follows that the condition $\eta_1(4\kappa^2 M(\lambda)+\lambda) \le 1$ in Proposition 3.8 holds true when $\eta_t = \eta_1 t^{-\theta}$ and $\eta_1 = \eta_* \lambda^{\max\{q-2,0\}}$. Based on Proposition 3.8 and $(mn/m+n^{3/2})^s \le t \le 2(mn/m+n^{3/2})^s$, we have, with confidence at least $1-\delta$,

$$\begin{aligned}\left\|f_t^S - f_\lambda\right\|_{\mathscr{H}_K}^2 \le{}& \left(\widetilde{c}_1\|f_\lambda\|_{\mathscr{H}_K}^2 \exp\left\{\frac{\eta_*}{1-\theta} + \frac{\widetilde{c}_2\eta_*^2}{(1-\theta)2^{1-\theta}}\left(\frac{mn}{(m+n)^{3/2}}\right)^{s(1-\theta-2(q-1)\gamma)}\right\}\right)\\ &\times \exp\left\{-\frac{\eta_*(1-2^{\theta-1})}{1-\theta}t^{1-\theta-\gamma\max\{q-1,1\}}\right\} + \widetilde{c}_3\left(\frac{mn}{(m+n)^{3/2}}\right)^{s\gamma(1+q)-1}\\ &+ \widetilde{c}_4\eta_*\left(\frac{mn}{(m+n)^{3/2}}\right)^{s\gamma\min\{q+1,2q-1\}-s\theta},\end{aligned} \tag{3.38}$$

where $\widetilde{c}_1, \widetilde{c}_2, \widetilde{c}_3$, and $\widetilde{c}_4$ are constants independent of m, n,and t.

Thus, when $1-\theta-\gamma\max\{q-1,1\} > 0$, we can derive the desired result in Theorem 2.1. □

Acknowledgments

This work was supported partially by the National Natural Science Foundation of China (NSFC) under Grant no. 11001092 and the Fundamental Research Funds for the Central Universities (Program no. 2011PY130, 2011QC022). The authors are indebted to the anonymous reviewers for their constructive comments.

References

[1] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[2] S. Agarwal and P. Niyogi, "Stability and generalization of bipartite ranking algorithms," In COLT, 2005.

[3] S. Smale and Y. Yao, "Online learning algorithms," *Foundations of Computational Mathematics*, vol. 6, no. 2, pp. 145–170, 2006.

[4] S. Smale and D. X. Zhou, "Online learning with Markov sampling," *Analysis and Applications*, vol. 7, no. 1, pp. 87–113, 2009.

[5] Y. Ying and D. X. Zhou, "Online regularized classification algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 11, pp. 4775–4788, 2006.

[6] X. M. Dong and D. R. Chen, "Learning rates of gradient descent algorithm for classification," *Journal of Computational and Applied Mathematics*, vol. 224, no. 1, pp. 182–192, 2009.

[7] J. Cai, H. Wang, and D. X. Zhou, "Gradient learning in a classification setting by gradient descent," *Journal of Approximation Theory*, vol. 161, no. 2, pp. 674–692, 2009.

[8] X. Dong and D. X. Zhou, "Learning gradients by a gradient descent algorithm," *Journal of Mathematical Analysis and Applications*, vol. 341, no. 2, pp. 1018–1027, 2008.

[9] C. Burges, T. Shaked, E. Renshaw et al., "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, 2005.

[10] V. C. Raykar, R. Duraiswami, and B. Krishnapuram, "A fast algorithm for learning a ranking function from large-scale data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1158–1170, 2008.

[11] H. Chen, Y. Tang, L. Q. Li, and X. Li, "Ranking by a gradient descent algorithm," manuscript.

[12] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, no. 6, pp. 933–969, 2004.

[13] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, "Generalization bounds for the area under the ROC curve," *Journal of Machine Learning Research*, vol. 6, pp. 393–425, 2005.

[14] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *The Annals of Statistics*, vol. 36, no. 2, pp. 844–874, 2008.

[15] C. Rudin and R. E. Schapire, "Margin-based ranking and an equivalence between AdaBoost and RankBoost," *Journal of Machine Learning Research*, vol. 10, pp. 2193–2232, 2009.

[16] C. Rudin, "The P-norm push: a simple convex ranking algorithm that concentrates at the top of the list," *Journal of Machine Learning Research*, vol. 10, pp. 2233–2271, 2009.

[17] S. Agarwal and P. Niyogi, "Generalization bounds for ranking algorithms via algorithmic stability," *Journal of Machine Learning Research*, vol. 10, pp. 441–474, 2009.

[18] D. Cossock and T. Zhang, "Statistical analysis of Bayes optimal subset ranking," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5140–5154, 2008.

[19] S. Smale and D. X. Zhou, "Shannon sampling. II. Connections to learning theory," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 285–302, 2005.

[20] S. Smale and D. X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, no. 2, pp. 153–172, 2007.

[21] S. Mukherjee and Q. Wu, "Estimation of gradients and coordinate covariation in classification," *Journal of Machine Learning Research*, vol. 7, pp. 2481–2514, 2006.

[22] S. Mukherjee and D. X. Zhou, "Learning coordinate covariances via gradients," *Journal of Machine Learning Research*, vol. 7, pp. 519–549, 2006.

[23] H. Chen and L. Q. Li, "Learning rates of multi-kernel regularized regression," *Journal of Statistical Planning and Inference*, vol. 140, no. 9, pp. 2562–2568, 2010.

[24] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics, 1989 (Norwich, 1989)*, vol. 141, pp. 148–188, Cambridge University Press, Cambridge, UK, 1989.