Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Uniform Convergence, Stability and Learnability for Ranking Problems *

Wei Gao and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{gaow, zhouzh}@lamda.nju.edu.cn

Abstract

Most studies were devoted to the design of efficient algorithms and the evaluation and application on diverse ranking problems, whereas few work has been paid to the theoretical studies on ranking learnability. In this paper, we study the relation between uniform convergence, stability and learnability of ranking. In contrast to supervised learning where the learnability is equivalent to uniform convergence, we show that the ranking uniform convergence is sufficient but not necessary for ranking learnability with AERM, and we further present a sufficient condition for ranking uniform convergence with respect to bipartite ranking loss. Considering the ranking uniform convergence being unnecessary for ranking learnability, we prove that the ranking average stability is a necessary and sufficient condition for ranking learnability.

1 Introduction

The most central issue in learning theory focuses on the problem of characterizing learnability. The fundamental work [Blumer *et al.*, 1989; Alon *et al.*, 1997], at least for supervised classification and regression, showed that learnability is equivalent to the uniform convergence of the empirical risk to the expected risk, and that if a problem is learnable, it is learnable with ERM (Empirical Risk Minimizer). Uniform convergence, however, does not hold in the general learning setting [Alon *et al.*, 1997; Vapnik, 1998], and Shalev-Shwartz et al. [2010] established stability as a sufficient and necessary condition for learnability.

Ranking problems try to learn a real-valued function so as to assign scores to instances, and then a relative ranking of instances is derived based on these scores. What is important is the relative rankings rather than the scores themselves. For example, Alice would like to search for documents that are relevant to a given topic, and what she wants is a ranked list in which the relevant documents are placed higher than irrelevant ones, no matter which score is assigned to which document. During the past decade, ranking problems have attracted much attention and have been found useful in various applications. Considering that ranking is quite different from traditional learning settings such as classification and regression, it is interesting to study what conditions can be used to characterize the learnability of ranking.

In this paper, we show that the ranking uniform convergence is sufficient but not necessary for ranking learnability with AERM (Asymptotic Empirical Risk Minimizer), and further present a sufficient condition for ranking uniform convergence with respect to the bipartite ranking loss, which is used commonly in literature. Then, we explore ranking stability as an alternative tool to characterize the ranking learnability in place of ranking uniform convergence. We show that the previous uniform stability is not necessary for ranking learnability, and thus introduce the average stability and prove that it is sufficient and necessary for learnability of ranking. This characterization holds even for ranking problems without ranking uniform convergence, and thus we provide a more convenient tool than ranking uniform convergence to characterize the learnability of ranking problems.

The rest of this paper is organized as follows. Section 2 introduces the related work and Section 3 provides background. Section 4 discusses the relationships between the ranking uniform convergence and ranking learnability. Section 5 introduces the ranking average stability and proves its equivalence with learnability. Section 6 concludes with future work.

2 Related Work

The study on ranking can be traced back to 1970's in social choice theory [Arrow, 1970] and some statistics research [Lehmann, 1975], while early studies on learning to rank in the machine learning community could be dated back to Cohen et al. [1999]. Many studies have been devoted to designing efficient algorithms for ranking problems under different settings, e.g., the online ranking algorithm [Crammer and Singer, 2002; Weng and Lin, 2011], neural network algorithm [Burges *et al.*, 2005], boosting-style algorithms [Cohen *et al.*, 1999; Freund *et al.*, 2003; Rudin and Schapire, 2009], univariate loss algorithms [Kotlowski *et al.*, 2011], regression algorithms [Herbrich *et al.*, 2000], listwise approaches [Cao *et al.*, 2007], etc. Considering that most approaches work with convex surrogate loss functions, many studies have been devoted to the asymptotic consistency of ranking approaches

*This research was supported by 973 Program (2010CB327903), NSFC (61073097, 61021062) and Baidu Fund (181315P00651).

based on surrogate loss functions [Clémençon *et al.*, 2008; Duchi *et al.*, 2010; Gao and Zhou, 2012].

Various generalization bounds have been developed to estimate the ranking performance of algorithms beyond training data based on different measures such as VC-dimension and cover number [Freund *et al.*, 2003; Agarwal *et al.*, 2005; Rudin and Schapire, 2009; Wang *et al.*, 2012]. In place of function space, Agarwal and Niyogi [2009] presented a new generalization bound based on the ranking uniform stability, whereas such stability is proven to be far from necessary for ranking learnability, as shown by Example 1 (Section 5.1).

Learnability has always been one of the most central issues in machine learning, whereas the traditional work [Blumer *et al.*, 1989; Alon *et al.*, 1997; Shalev-Shwartz *et al.*, 2010] can not be applied directly to ranking problems since the ranking loss is based on pairwise instances from different classes rather than single instance. Therefore, Agarwal and Roth [2005] presented the first work on ranking learnability by providing a sufficient condition and a necessary condition, separately, for ranking learnability based on two different measures of function space. To the best of our knowledge, our work presents the first sufficient and necessary condition for ranking learnability, and such characterization does not rely on any space complexity measure, but rather on the way the algorithm searches.

3 Background

The bipartite ranking problem focuses on two classes, i.e., a positive class and a negative class, where the objective is to learn a real-valued function to rank future positive instances higher than negative ones. Formally, let $\mathcal{X}$ be an instance space and assume that positive instances are chosen randomly and independently (i.i.d.) according to some (unknown) underlying distribution $\mathcal{D}_+$ over $\mathcal{X}$, and negative instances are chosen similarly according to some (unknown) underlying distribution $\mathcal{D}_-$ over $\mathcal{X}$. A ranking problem is specified by a real-valued function class $\mathcal{H}$, an instance space $\mathcal{X}$ with respect to distributions $\mathcal{D}_+$ and $\mathcal{D}_-$, and a cost (e.g., "loss" and "objective") function $c\colon \mathcal{H} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, where it aims to minimize the expected risk

$$R(h) = E_{x^+ \sim \mathcal{D}_+, x^- \sim \mathcal{D}_-}[c(h, x^+, x^-)]$$

over the class $\mathcal{H}$. Distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ are unknown, and what we have is a training sample (S_+, S_-) with

$$S_+ = (x_1^+, x_2^+, \cdots, x_m^+) \text{ and } S_- = (x_1^-, x_2^-, \cdots, x_n^-)$$

being sampled i.i.d. according to distributions $\mathcal{D}_+$ and $\mathcal{D}_-$, respectively. Based on this sample, we can define the empirical risk $\hat{R}(h, S_+, S_-)$, with respect to $h \in \mathcal{H}$, to be

$$\hat{R}(h, S_+, S_-) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} c(h, x_i^+, x_j^-).$$

Formally, a (deterministic) ranking algorithm is a mapping $\mathscr{A}\colon \cup_{m=1}^{\infty} \mathcal{X}^m \times \cup_{n=1}^{\infty} \mathcal{X}^n \to \mathcal{H}$ from training sample (S_+, S_-) to a real-valued function $\mathscr{A}_{S_+,S_-} : \mathcal{X} \to \mathbb{R}$, where $\mathscr{A}_{S_+,S_-} \in \mathcal{H}$.

Throughout this paper, we consider symmetric ranking algorithms, i.e., algorithms depending upon the given training sample but not on the order of instances in the training sample, and we assume that the cost function is bounded by some constant B, i.e.,

$$|c(h, x^+, x^-)| \leq B \text{ for all } h \in \mathcal{H} \text{ and } x^+, x^- \in \mathcal{X}.$$

The *bipartite ranking loss*, which is commonly studied in literature for bipartite ranking problems, is:

$$c(h, x^+, x^-) = I[h(x^+) < h(x^-)] + I[h(x^+) = h(x^-)]/2 \quad (1)$$

where $I[\cdot]$ denotes the indicator function which returns one if its argument is true and zero otherwise. Based on this cost function, the empirical risk $\hat{R}(h, S_+, S_-)$ becomes an evaluation criterion, Area Under the ROC Curve (AUC), with respect to function h on the training sample (S_+, S_-), which has been studied in detail [Cortes and Mohri, 2003; Agarwal *et al.*, 2005].

A ranking algorithm would like, in essence, to pick a real-valued function $h \in \mathcal{H}$ such that its expected risk approaches the minimal expected risk, i.e.,

$$R(h^*) = \inf_{h \in \mathcal{H}} R(h)$$

where we denote by $h^* = \arg\inf_{h \in \mathcal{H}} R(h)$. Since distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ are unknown but demonstrated by training sample (S_+, S_-), a natural method is to minimize the empirical $\hat{R}(h, S_+, S_-)$. A ranking algorithm $\mathscr{A}$ is said to be an ERM (Empirical Risk Minimizer) if it minimizes the empirical risk, i.e.,

$$\begin{aligned} \hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-) &= \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) \\ &= \inf_{h \in \mathcal{H}} \hat{R}(h, S_+, S_-), \end{aligned}$$

where we denote by $\hat{h}_{S_+,S_-} = \arg\inf_{h \in \mathcal{H}} \hat{R}(h, S_+, S_-)$.

A ranking algorithm $\mathscr{A}$ is said to be an AERM (Asymptotic Empirical Risk Minimizer) with rate $\epsilon_{\text{erm}}(m, n)$ under distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ if

$$E_{S_+ \sim \mathcal{D}_+^m, S_- \sim \mathcal{D}_-^n}[\hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-) - \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-)] \leq \epsilon_{\text{erm}}(m, n).$$

A ranking algorithm is universally an AERM with rate $\epsilon_{\text{erm}}(m, n)$ if it is an AERM with rate $\epsilon_{\text{erm}}(m, n)$ under all distributions $\mathcal{D}^+$ and $\mathcal{D}^-$ over $\mathcal{X}$. Here and wherever mentioning about $\epsilon(m, n)$, it requires $\epsilon(m, n) \to 0$ as $m, n \to \infty$. Note that ERM is a special case of AERM and we will mainly focus on AERM in this paper.

A ranking algorithm $\mathscr{A}$ is said to be consistent with rate $\epsilon_{\text{con}}(m, n)$ under distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ if

$$E_{S_+ \sim \mathcal{D}_+^m, S_- \sim \mathcal{D}_-^n}[R(\mathscr{A}_{S_+,S_-}) - R(h^*)] \leq \epsilon_{\text{con}}(m, n).$$

A ranking algorithm $\mathscr{A}$ is universally consistent with rate $\epsilon_{\text{con}}(m, n)$ if it is consistent with rate $\epsilon_{\text{con}}(m, n)$ under all distributions $\mathcal{D}^+$ and $\mathcal{D}^-$ over $\mathcal{X}$. Based on this definition, we give the following notion of ranking learnability:

Definition 1 *A bipartite ranking problem with real-valued class $\mathcal{H}$ is learnable if there exists a universally consistent algorithm for such problem, i.e., there exists a ranking algorithm $\mathscr{A}$ and rate $\epsilon_{con}(m, n)$ such that*

$$\sup_{\mathcal{D}_+, \mathcal{D}_-} \left\{ E_{S_+,S_-}[R(\mathscr{A}_{S_+,S_-}) - R(h^*)] \right\} \leq \epsilon_{con}(m, n)$$

where $\epsilon_{con}(m, n) \to 0$ as $m, n \to \infty$.

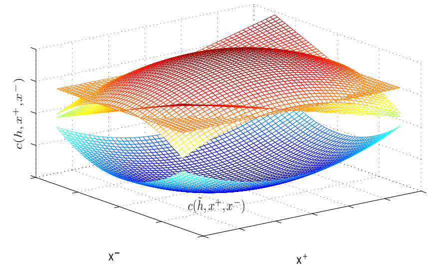

Figure 1: Each surface represents some $h \in \mathcal{H}$ and shows the value of $c(h, x^+, x^-)$. The function $\tilde{h}$ satisfies $c(\tilde{h}, x^+, x^-) < c(h, x^+, x^-)$ for all $h \in \mathcal{H}$. This problem is learnable with ERM without ranking uniform convergence.

4 Uniform Convergence and Learnability

In this section, we study the relationship between ranking uniform convergence and ranking learnability. Our result shows that ranking uniform convergence is sufficient but not necessary for ranking learnability. For bipartite ranking loss, we present a sufficient condition for uniform convergence.

We introduce the ranking uniform convergence as follows:

Definition 2 *For a ranking problem with real-valued function class $\mathcal{H}$, the ranking uniform convergence holds with rate $\epsilon_{\textit{uni-con}}(m, n)$ if*

$$\sup_{\mathcal{D}_+,\mathcal{D}_-} E_{S_+,S_-}[\sup_{h\in\mathcal{H}} |R(h) - \hat{R}(h, S_+, S_-)|] \leq \epsilon_{\textit{uni-con}}(m, n)$$

where $\epsilon_{\textit{uni-con}}(m, n) \to 0$ as $m, n \to \infty$.

By using the facts

$$\begin{aligned} R(\mathscr{A}_{S_+,S_-}) - R(h^*) &= \hat{R}(h^*, S_+, S_-) - R(h^*) \\ &+R(\mathscr{A}_{S_+,S_-}) - \hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-) \\ &+\hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-) - \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) \\ &+\hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) - \hat{R}(h^*, S_+, S_-), \end{aligned}$$

and $\hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) \leq \hat{R}(h^*, S_+, S_-)$, we have

Lemma 1 *For a ranking problem with real-valued function class $\mathcal{H}$, if the ranking uniform convergence holds with rate $\epsilon_{\textit{uni-con}}(m, n)$, then, for an AERM with rate $\epsilon_{\textit{erm}}(m, n)$, this ranking problem is learnable with rate $\epsilon_{\textit{con}}(m, n) = 2\epsilon_{\textit{uni-con}}(m, n) + \epsilon_{\textit{erm}}(m, n)$.*

This lemma states that ranking uniform convergence implies ranking learnability with AERM, and such ranking problem is learnable. However, it is not a necessary condition owing to some "trivial" situations: for a ranking problem with function class $\mathcal{H}$ that could be extremely complex without ranking uniform convergence, we add a single function $\tilde{h}$ such that $c(\tilde{h}, x^+, x^-) < \inf_{h\in\mathcal{H}} c(h, x^+, x^-)$, as shown in Figure 1. This ranking problem is now trivially learnable with ERM which always picks $\tilde{h}$, yet the ranking uniform convergence does not hold. To exclude such cases, one could, inspired from [Vapnik, 1998], introduce stronger notions such as *strict ranking consistency* and *one-side uniform convergence*[1], whereas we will present a different tool to study the ranking learnability in Section 5.2.

In the following of this section, we focus on the bipartite ranking loss given by Eq. (1) and explore a sufficient condition for ranking uniform convergence. Let us recall a combinatorial parameter: bipartite rank-shatter coefficient [Agarwal *et al.*, 2005], defined below with our notations.

Definition 3 (Bipartite rank-shatter coefficient) *For positives m and n, and a ranking problem with function class $\mathcal{H}$, the (m, n)-th bipartite rank-shatter coefficient is*

$$r(\mathcal{H}, m, n) = \max_{S_+\in\mathcal{X}^m, S_-\in\mathcal{X}^n} |\{B_h(S_+, S_-) \colon h \in \mathcal{H}\}|,$$

where $B_h(S_+, S_-)$ is a matrix, whose (i, j) element

$$\begin{aligned} [B_h(S_+, S_-)]_{ij} &= c(h, x_i^+, x_j^-) \\ &= I[h(x_i^+) < h(x_j^-)] + I[h(x_i^+) = h(x_j^-)]/2. \end{aligned}$$

This coefficient plays the similar role as the shatter coefficient plays in binary classification. We present the following sufficient condition for ranking uniform convergence, analogous to the *third milestone* in learning theory [Vapnik, 1998].

Theorem 1 *For a ranking problem with real-valued function class $\mathcal{H}$, the ranking uniform convergence holds if*

$$G(m, n) = \left(\frac{1}{m} + \frac{1}{n}\right) \ln r(\mathcal{H}, 2m, 2n) \to 0$$

as $m, n \to \infty$. Specifically, the ranking uniform convergence holds with rate

$$\epsilon_{uni-con} = \sqrt[4]{\frac{1}{m} + \frac{1}{n}} + 8r(\mathcal{H}, 2m, 2n)e^{-\sqrt{\frac{mn}{8(m+n)}}},$$

if $r(\mathcal{H}, 2m, 2n)$ is finite as $m, n \to \infty$; otherwise

$$\epsilon_{uni-con} = \sqrt{2G(m, n)} + 8/r(\mathcal{H}, 2m, 2n).$$

Proof For any $\epsilon > 0$ and distributions $\mathcal{D}_+$ and $\mathcal{D}_-$, we have

$$P_{S_+\sim\mathcal{D}_+^m, S_-\sim\mathcal{D}_-^n}\left[\sup_{h\in\mathcal{H}} |R(h) - \hat{R}(h, S_+, S_-)| \geq \epsilon\right] \leq 4r(\mathcal{H}, 2m, 2n)e^{-\frac{mn\epsilon^2}{8(m+n)}}$$

from [Agarwal *et al.*, 2005, Theorem 17]. Further, we set $\epsilon = \sqrt[4]{(m+n)/mn}$ if $r(\mathcal{H}, 2m, 2n)$ is finite as $m, n \to \infty$; otherwise, $\epsilon = \sqrt{16G(m, n)}$. By using the formula

$$\begin{aligned} E[X] = E[X|X \geq \epsilon] \Pr[X \geq \epsilon] \\ + E[X|X < \epsilon] \Pr[X < \epsilon], \end{aligned} \tag{2}$$

this theorem follows by simple calculation. ∎

It is easy to observe that $G(m, n) \to 0$ is equivalent to

$$\frac{1}{m} \ln r(\mathcal{H}, 2m, 2n) \to 0 \quad \text{and} \quad \frac{1}{n} \ln r(\mathcal{H}, 2m, 2n) \to 0,$$

as $m, n \to \infty$; therefore, we have

[1]Due to space limit, we omit the detailed discussion on strict ranking consistency and one-side uniform convergence, and also omit some detailed proofs of results such as Lemma 1 and Theorem 1. We will present them in a longer version.

$\boxed{\lim_{m,n\to\infty} G(m,n) = 0} \Rightarrow \boxed{\text{Ranking uniform convergence}}$

$\Downarrow$

$\boxed{\text{learnability}} \Leftarrow \boxed{\text{learnability with AERM}}$

Figure 2: Relationships among $G(m, n)$, ranking learnability and ranking uniform convergence.

Corollary 1 *For a ranking problem with function class $\mathcal{H}$, the ranking uniform convergence holds if, as $m, n \to \infty$,*

$$\frac{1}{m}\ln r(\mathcal{H}, 2m, 2n) \to 0 \quad \textit{and} \quad \frac{1}{n}\ln r(\mathcal{H}, 2m, 2n) \to 0.$$

Combining Lemma 1 with Theorem 1, we establish the relationships among $G(m, n)$, ranking uniform convergence and ranking learnability, as shown in Figure 2. It is, however, not clear whether ranking uniform convergence implies $\lim_{m,n\to\infty} G(m, n) = 0$, and thus we leave the following open problem:

Problem 1 *Does $\lim_{m,n\to\infty} G(m, n)$ equal to zero if the ranking uniform convergence holds?*

5 Ranking Stability and Learnability

Section 5.1 shows that previous uniform stability is not necessary for ranking learnability and introduces average stability. Section 5.2 proves our main result: a ranking problem is learnable if and only if there exists a universally average stable AERM.

5.1 Ranking Stability

Given a sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ with

$$S_+ = \{x_1^+, x_2^+, \cdots, x_m^+\} \text{ and } S_- = \{x_1^-, x_2^-, \cdots, x_n^-\},$$

we define, for $i \in [m]$ and $z^+ \in \mathcal{X}$,

$$S_+^{i,z^+} = \{x_1^+, \cdots, x_{i-1}^+, z^+, x_{i+1}^+, \cdots, x_m^+\}$$

to be the sample with the i-th instance x_i^+ in S_+ replaced by another z^+. Similarly, we can define S_-^{j,z^-}. For real number α, we denote by $\lfloor\alpha\rfloor$ the biggest integer smaller than α.

Several notions of stability have been proposed to study the performance of learning algorithms [Kearns and Ron, 1999; Bousquet and Elisseeff, 2002; Rakhlin *et al.*, 2005; Mukherjee *et al.*, 2006; Gao and Zhou, 2010]. For bipartite ranking, Agarwal and Niyogi [2009] gave the definition of ranking uniform stability, which is equivalent to

Definition 4 *A ranking algorithm $\mathscr{A}$ has ranking uniform stability $\beta(m, n)$ if the it holds that, for any $i \in [m]$, $j \in [n]$, $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, and $x^+, x^-, z^+, z^- \in \mathcal{X}^4$*

$$\left|c(\mathscr{A}_{S_+,S_-}, x^+, x^-) - c(\mathscr{A}_{S_+^{i,z^+}, S_-^{j,z^-}}, x^+, x^-)\right| \le \beta(m, n)$$

where $\beta(m, n) \to 0$ as $m, n \to \infty$.

The ranking uniform stability, which could be a stronger notion, is enough for tight generalization error bounds, and Agarwal and Niyogi [2009] established such stability for some kernel-based ranking algorithms. It is, however, far from necessary for ranking learnability, as shown by the following Example.

Example 1 *There exists a ranking problem with a universally consistent algorithm without ranking uniform stability.*

Proof Let $\mathcal{X} = [-1, 1]$, $\mathcal{H} = \{h_\theta \colon h_\theta(x) = I[x \ge \theta]\}$ for $\theta \in [-1, 1]$, the objective $t(x) = I[x \ge 0]$ and the cost function is given by Eq. (1). Given a sample $(S_+, S_-) \in [0, 1]^m \times [-1, 0)^n$ with $S_- = \{x_1^-, x_2^-, \cdots, x_n^-\}$, we set $\theta^* = \max\{x_1^-, x_2^-, \cdots, x_n^-\}$ and consider the algorithm

$$\mathscr{A}_{S_+,S_-}(x) = h_{\theta^*}(x) = I[x \ge \theta^*].$$

For any $h_\theta \in \mathcal{H}$ and distributions $\mathcal{D}_+$ and $\mathcal{D}_-$, it is easy to derive $h_\theta(x^+) \ge h_\theta(x^-)$ and $R(h_\theta) = \Pr[\theta \le x^- < 0]/2$. This follows that, for any $\epsilon \ge 0$,

$$\begin{aligned}\Pr_{S_+,S_-}[R(\mathscr{A}_{S_+,S_-}) \ge \epsilon] &= \Pr_{S_+,S_-}[\Pr_{x^-}[\theta^* \le x^- < 0] \ge 2\epsilon] \\ &\le (1 - 2\epsilon)^n \le \exp(-2n\epsilon).\end{aligned}$$

It is clear that

$$R(h^*) = \inf_{h_\theta \in \mathcal{H}}\{R(h_\theta)\} = 0,$$

and combining with Eqn. (2) together, the following holds by setting $\epsilon = 1/\sqrt{n}$,

$$E_{S_+,S_-}[R(\mathscr{A}_{S_+,S_-}) - R(h^*)] \le 1/\sqrt{n} + \exp(-2\sqrt{n}),$$

which proves that $\mathscr{A}_{S_+,S_-}$ is universally consistent.

On the other hand, if we choose an instance z^- satisfying $\max\{x_1^-, x_2^-, \cdots, x_n^-\} < z^- < 0$, then, for any instance x^- such that $\max\{x_1^-, x_2^-, \cdots, x_n^-\} < x^- < z^-$, it holds

$$\left|c(\mathscr{A}_{S_+,S_-}, x^+, x^-) - c(\mathscr{A}_{S_+,S_-^{j,z^-}}, x^+, x^-)\right| = 1/2,$$

which implies that this ranking algorithm has no uniform stability, and thus we complete the proof. ■

Our weaker notion of stability, which is equivalent to ranking learnability with AERM (proved later), is

Definition 5 *A ranking algorithm $\mathscr{A}$ has ranking average stability $\beta(m, n)$ under distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ if, for all $i \in [m]$ and $j \in [n]$, the following holds:*

$$\begin{aligned}\big|E_{S_+,x^+,S_-,x^-}[&c(\mathscr{A}_{S_+,S_-}, x^+, x^-) \\ &- c(\mathscr{A}_{S_+^{i,x^+}, S_-^{j,x^-}}, x^+, x^-)]\big| \le \beta(m, n)\end{aligned}$$

where $\beta(m, n) \to 0$ as $m, n \to \infty$.

A ranking algorithm has universally average stability with rate $\beta(m, n)$ if the stability property holds with rate $\beta(m, n)$ under all distributions $\mathcal{D}_+$ and $\mathcal{D}_-$. It is obvious that

Proposition 1 *Ranking uniform stability implies ranking average stability.*

5.2 Sufficient and Necessary Condition for Ranking Learnability

We begin with a useful lemma as follows:

Lemma 2 *For a ranking problem with real-valued function class $\mathcal{H}$, the following holds for every $h \in \mathcal{H}$ independent of sample (S_+, S_-),*

$$E_{S_+\sim\mathcal{D}_+^m, S_-\sim\mathcal{D}_-^n}[|R(h) - \hat{R}(h, S_+, S_-)|] \leq B\sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Proof We first have

$$|\hat{R}(h, S_+, S_-) - \hat{R}(h, S_+^{i,x^+}, S_-)| \leq 2B/m$$

for index $i \in [m]$ and instance x^+ drawn i.i.d. from $\mathcal{D}^+$. In a similar manner,

$$|\hat{R}(h, S_+, S_-) - \hat{R}(h, S_+, S_-^{j,x^-})| \leq 2B/n.$$

Based on [Devroye, 1991, Theorem 3], we have

$$E_{S_+,S_-}[(\hat{R}(h, S_+, S_-) - R(h))^2] \leq B^2/m + B^2/n,$$

which completes the proof by using Jensen's inequality. ■

Based on this lemma, we present the first main result:

Theorem 2 *For an AERM $\mathscr{A}$ with rate $\epsilon_{erm}(m,n)$, if it has ranking average stability with rate $\beta(m,n)$ under distributions $\mathcal{D}^+$ and $\mathcal{D}^-$, then it is consistent with rate*

$$\epsilon_{con}(m,n) = \beta(m,n) + \epsilon_{erm}(m,n) + B\sqrt{(m+n)/mn}.$$

Proof For an AERM $\mathscr{A}$, we have

$$\begin{aligned}
&E[R(\mathscr{A}_{S_+,S_-}) - R(h^*)] \\
&= E[R(\mathscr{A}_{S_+,S_-}) - \hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-)] \\
&\quad + E[\hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-) - \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-)] \\
&\quad + E[\hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) - \hat{R}(h^*, S_+, S_-)] \\
&\quad + E[\hat{R}(h^*, S_+, S_-) - R(h^*)] \\
&\leq E[R(\mathscr{A}_{S_+,S_-}) - \hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-)] \\
&\quad + \epsilon_{\text{erm}}(m,n) + B\sqrt{(m+n)/mn},
\end{aligned}$$

from Lemma 2 and $\hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) \leq \hat{R}(h^*, S_+, S_-)$. For symmetric algorithm, we have, for all $i \in [m]$, $j \in [n]$,

$$\begin{aligned}
&E_{S_+,S_-}[\hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-)] \\
&= E_{S_+,S_-}[c(\mathscr{A}_{S_+,S_-}, x_i^+, x_j^-)] \\
&= E_{S_+,x^+,S_-,x^-}[c(\mathscr{A}_{S_+^{i,x^+},S_-^{j,x^-}}, x^+, x^-)] \qquad (3)
\end{aligned}$$

by interchanging the roles of x_i^+, x_j^- with x^+, x^-, respectively. Similarly, we have

$$\begin{aligned}
&E_{S_+,S_-}[R(\mathscr{A}_{S_+,S_-})] \\
&= E_{S_+,x^+,S_-,x^-}[c(\mathscr{A}_{S_+,S_-}, x^+, x^-)]. \qquad (4)
\end{aligned}$$

Combining Eqs. (3) and (4) gives

$$\begin{aligned}
&|E_{S_+,S_-}[R(\mathscr{A}_{S_+,S_-}) - \hat{R}(\mathscr{A}_{S_+,S_-}, S_+, S_-)]| \\
&= |E_{S_+,x^+,S_-,x^-}[c(\mathscr{A}_{S_+,S_-}, x^+, x^-) \\
&\quad - c(\mathscr{A}_{S_+^{i,x^+},S_-^{j,x^-}}, x^+, x^-)]| \\
&\leq \beta(m,n).
\end{aligned}$$

This completes the proof. ■

This theorem shows that an average stable AERM is sufficient for consistency under each pair of distributions $\mathcal{D}^+$ and $\mathcal{D}^-$, and it does not require the universality. Based on this theorem, we have

Corollary 2 *A ranking problem is learnable if there exists a universally average stable AERM.*

For Theorem 2, we find that the reverse direction does not hold under specific pair of distributions $\mathcal{D}^+$ and $\mathcal{D}^-$, which can be shown by following example:

Example 2 *There exists a ranking problem and distributions $\mathcal{D}_+$ and $\mathcal{D}_-$, such that any ERM (or any AERM) is consistent but not average stable.*

Proof Let the instance space $\mathcal{X} = [-1, 1]$, the objective ranking $t(x) = I[x \geq 0]$, and distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ be any two continuous distributions on $[0, 1]$ and $[-1, 0)$, respectively. We consider the bipartite ranking loss given by Eq. (1) and the hypothesis space

$$\mathcal{H} = \{h_\Delta \colon h_\Delta(x) = I[x \in \Delta] \text{ for } \Delta \in \mathcal{H}_0\}$$

where

$$\mathcal{H}_0 = \{[-1, 0) \cup A \colon A \text{ is a finite subset of } [0, 1]\}.$$

For any $h \in \mathcal{H}$, we have $R(h) = 1$ for the continuous distributions $\mathcal{D}_+$ on $[0, 1]$ and thus this ranking algorithm is consistent. On the other hand, an ERM always achieves $\hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-) = 1/2$ and

$$\begin{aligned}
&\big|E_{S_+,x^+,S_-,x^-}[c(\mathscr{A}_{S_+,S_-}, x^+, x^-) \\
&\qquad - c(\mathscr{A}_{S_+^{i,x^+},S_-^{j,x^-}}, x^+, x^-)]\big| = 1/2,
\end{aligned}$$

which implies that ERM (or any AERM) is not ranking average stable. This completes the proof. ■

Example 2 shows that a consistent ERM does not imply ranking average stability under the specific pair of distributions $\mathcal{D}^+$ and $\mathcal{D}^-$, whereas the universal consistency implies the existence of average stable AERMs as follows:

Theorem 3 *For a ranking problem with a universally consistent algorithm $\mathscr{A}$ with rate $\epsilon_{con}(m,n)$, there must exist a ranking algorithm $\hat{\mathscr{A}}$ such that $\hat{\mathscr{A}}$ is universally an AERM with rate $\epsilon_{erm}(m,n)$ and universally ranking average stable with rate $\hat{\beta}(m,n)$, where*

$$\begin{aligned}
\epsilon_{erm}(m,n) &= \epsilon_{con}(m^{\frac{1}{4}}, n^{\frac{1}{4}}) + 2B/m^{\frac{1}{2}} + 2B/n^{\frac{1}{2}}, \\
\hat{\beta}(m,n) &= \frac{2B}{m^{\frac{3}{4}}n^{\frac{3}{4}}} + \frac{2B}{m^{\frac{3}{4}}} + \frac{2B}{n^{\frac{3}{4}}} + B\sqrt{\frac{2}{m} + \frac{2}{n}}.
\end{aligned}$$

Proof Given a sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, let $(\hat{S}_+, \hat{S}_-) \in \mathcal{X}^{\lfloor m^{\frac{1}{4}} \rfloor} \times \mathcal{X}^{\lfloor n^{\frac{1}{4}} \rfloor}$ be drawn uniformly from S_+ and S_-, respectively, and define $\hat{\mathscr{A}}_{S_+,S_-} = \mathscr{A}_{\hat{S}_+,\hat{S}_-}$. We have

$$\begin{aligned}
&\hat{R}(\hat{\mathscr{A}}_{S_+,S_-}, S_+, S_-) - R(\hat{\mathscr{A}}_{S_+,S_-}) \\
&= \hat{R}(\mathscr{A}_{\hat{S}_+,\hat{S}_-}, S_+, S_-) - R(\mathscr{A}_{\hat{S}_+,\hat{S}_-}) \\
&= \big(\hat{R}(\mathscr{A}_{\hat{S}_+,\hat{S}_-}, \hat{S}_+, \hat{S}_-) - R(\mathscr{A}_{\hat{S}_+,\hat{S}_-})\big)/m^{\frac{3}{4}}n^{\frac{3}{4}} \\
&\quad + \big(\hat{R}(\mathscr{A}_{\hat{S}_+,\hat{S}_-}, S_+\backslash\hat{S}_+, \hat{S}_-) - R(\mathscr{A}_{\hat{S}_+,\hat{S}_-})\big)\tau_1/mn^{\frac{3}{4}} \\
&\quad + \big(\hat{R}(\mathscr{A}_{\hat{S}_+,\hat{S}_-}, \hat{S}_+, S_-\backslash\hat{S}_-) - R(\mathscr{A}_{\hat{S}_+,\hat{S}_-})\big)\tau_2/m^{\frac{3}{4}}n \\
&\quad + \big(\hat{R}(\mathscr{A}_{\hat{S}_+,\hat{S}_-}, S_+\backslash\hat{S}_+, S_-\backslash\hat{S}_-) - R(\mathscr{A}_{\hat{S}_+,\hat{S}_-})\big)\tau_1\tau_2/mn,
\end{aligned}$$

where $\tau_1 = m - \lfloor m^{\frac{1}{4}} \rfloor$ and $\tau_2 = n - \lfloor n^{\frac{1}{4}} \rfloor$. The first three terms are bounded by cost function $|c(h, x^+, x^-)| \leq B$ and the last term can be bounded by Lemma 2 since $\mathscr{A}_{\hat{S}_+,\hat{S}_-}$ is independent of $S_+\backslash\hat{S}_+$ and $S_-\backslash\hat{S}_-$. Thus, we obtain

$$E[|\hat{R}(\hat{\mathscr{A}}_{S_+,S_-}, S_+, S_-) - R(\hat{\mathscr{A}}_{S_+,S_-})|] \leq \beta(m,n), \quad (5)$$

where $\beta(m,n) = \frac{2B}{m^{\frac{3}{4}}n^{\frac{3}{4}}} + \frac{2B}{m^{\frac{3}{4}}} + \frac{2B}{m^{\frac{3}{4}}} + B\sqrt{\frac{2}{m} + \frac{2}{n}}$ for $m \geq 3$ and $n \geq 3$. Combining Eqs. (3) with (4), we have

$$\begin{aligned}
&|E_{S_+,x^+,S_-,x^-}[c(\mathscr{A}_{S_+,S_-}, x^+, x^-) \\
&\qquad - c(\mathscr{A}_{S_+^{i,x^+}, S_-^{j,x^-}}, x^+, x^-)]| \\
&= |E[\hat{R}(\hat{\mathscr{A}}_{S_+,S_-}, S_+, S_-) - R(\hat{\mathscr{A}}_{S_+,S_-})]| \\
&\leq E[|\hat{R}(\hat{\mathscr{A}}_{S_+,S_-}, S_+, S_-) - R(\hat{\mathscr{A}}_{S_+,S_-})|] \\
&\leq \beta(m,n),
\end{aligned}$$

where the last inequality holds from Eq. (5). This shows that $\hat{\mathscr{A}}$ is average stable with rate $\beta(m,n)$.

We will use the universal consistency of $\mathscr{A}$ to prove that $\hat{\mathscr{A}}$ is an AERM. Denote by the random event

$$A = \{S'_+ \text{ or } S'_- \text{ have repeated-index instances from } S_+ \text{ or } S_-, \text{ respectively}\},$$

and it is easy to derive that

$$\begin{aligned}
\Pr[A] &\leq \frac{\sum_{i=1}^{\lfloor m^{\frac{1}{4}} \rfloor}(i-1)}{m} + \frac{\sum_{i=1}^{\lfloor n^{\frac{1}{4}} \rfloor}(i-1)}{n} \\
&\leq \frac{\lfloor m^{\frac{1}{4}} \rfloor^2}{m} + \frac{\lfloor n^{\frac{1}{4}} \rfloor^2}{n} \leq \frac{1}{m^{\frac{1}{2}}} + \frac{1}{n^{\frac{1}{2}}}. \quad (6)
\end{aligned}$$

Also, we denote by the random event

$$B = \{\text{both } S'_+ \text{ and } S'_- \text{ have no repeated-index instances from } S_+ \text{ and } S_-, \text{ respectively}\}.$$

If $\hat{S}_+$ and $\hat{S}_-$ have no repeated instances from S_+ and S_-, then S'_+ and S'_- can be viewed as samples drawn from the uniform distribution over instance in S_+ and S_-, respectively. Therefore, for universally consistent $\mathscr{A}$, we have

$$\begin{aligned}
&E_{S'_+,S'_-}[|\hat{R}(\mathscr{A}_{S'_+,S'_-}, S_+, S_-) \\
&\qquad - \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-)|] \leq \epsilon_{\text{con}}(m^{\frac{1}{4}}, n^{\frac{1}{4}}). \quad (7)
\end{aligned}$$

From the formula $E[X] = E[X|A]\Pr[A] + E[X|B]\Pr[B]$, and by using Eqs. (6), (7) and $|c(h, x^+, x^-)| \leq B$, we have

$$\begin{aligned}
&E[|\hat{R}(\hat{\mathscr{A}}_{S_+,S_-}, S_+, S_-) - \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-)|] \\
&= E[|\hat{R}(\mathscr{A}_{\hat{S}_+,\hat{S}_-}, S_+, S_-) - \hat{R}(\hat{h}_{S_+,S_-}, S_+, S_-)|] \\
&\leq 2B/m^{\frac{1}{2}} + 2B/n^{\frac{1}{2}} + \epsilon_{\text{con}}(m^{\frac{1}{4}}, n^{\frac{1}{4}}).
\end{aligned}$$

This theorem follows as desired. ■

Corollary 2 presents a sufficient condition for ranking learnability, and Theorem 3 shows that such condition is also a necessary condition; therefore, we present a sufficient and necessary condition for ranking learnability as follows:

Theorem 4 *A ranking problem is learnable if and only if there exists a universally average stable AERM.*

6 Conclusion and Future Work

Learning to rank has attracted much attention during the past decade, where bipartite ranking is a commonly studied problem setting. Compared to traditional learning settings such as classification and regression, the main difference of bipartite ranking problems lies in the loss function which is based on pairwise instances from different classes in ranking rather than single instance. Due to the dependence of instance pairs, the characterization of learnability for traditional learning setting could not be applied directly to ranking. Therefore, it is desirable to find sufficient and necessary conditions for characterizing the learnability of ranking, and previous work [Agarwal and Roth, 2005] presented a sufficient condition and a necessary condition for ranking learnability, separately, based on two different measures of function space.

In this paper, we show that the ranking uniform convergence is sufficient but not necessary for ranking learnability with AERM. For the special bipartite ranking loss, we provide a sufficient condition for ranking uniform convergence, that is

$$G(m,n) = \Big(\frac{1}{m} + \frac{1}{n}\Big)\ln r(\mathcal{H}, 2m, 2n) \to 0 \text{ as } m, n \to \infty.$$

This condition can be as analogous to the *third milestone* in learning theory [Vapnik, 1998]. We have not, however, proved that ranking uniform convergence implies $G(m,n) \to 0$, and leave it as an open problem in Problem 1. A possible solution could be motivated from [Blumer *et al.*, 1989] by considering the bipartite rank-shatter coefficient or introducing some new combinatorial parameters.

Considering the ranking uniform convergence is not necessary for ranking learnability, we establish ranking average stability as a sufficient and necessary condition for ranking learnability by proving that ranking learnability is equivalent to ranking average stability with AERM. One important direction for future research is to characterize the learnability of other ranking problems under different settings such as listwise ranking setting [Cao *et al.*, 2007], pointwise ranking setting [Cossock and Zhang, 2008] and the k-height ranking setting [Rudin, 2009].

References

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

[Agarwal and Roth, 2005] S. Agarwal and D. Roth. Learnability of bipartite ranking functions. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 16–31, Bertinoro, Italy, 2005.

[Agarwal *et al.*, 2005] 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*, 6:393–425, 2005.

[Alon *et al.*, 1997] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnablity. *Journal of ACM*, 44(4):615–631, 1997.

[Arrow, 1970] K. J. Arrow. *Social choice and individual values*. Yale University Press, New Haven, 2 edition, 1970.

[Blumer *et al.*, 1989] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of ACM*, 36(4):929–965, 1989.

[Bousquet and Elisseeff, 2002] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[Burges *et al.*, 2005] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, Bonn, Germany, 2005.

[Cao *et al.*, 2007] Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, and H. Li. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th Annual International Conference on Machine Learning*, pages 129–136, Corvallis, Oregon, 2007.

[Clémençon *et al.*, 2008] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36:844–874, 2008.

[Cohen *et al.*, 1999] W. W. Cohen, R. E. Schapire, and Y. Singer. Learning to order things. *Neural Computation*, 10:243–270, 1999.

[Cortes and Mohri, 2003] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2003.

[Cossock and Zhang, 2008] D. Cossock and T. Zhang. Statistical analysis of bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54:5140–5154, 2008.

[Crammer and Singer, 2002] K. Crammer and Y. Singer. Pranking with ranking. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 641–647. MIT Press, Cambridge, MA, 2002.

[Devroye, 1991] L. Devroye. Exponential inequalities in nonparametric estimation. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics, NATO ASI Seires*, pages 31–44. Kluwer Adademic Publisher, 1991.

[Duchi *et al.*, 2010] J. C. Duchi, L. W. Mackey, and M. I. Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th Annual International Conference on Machine Learning*, Haifa, Israel, 2010.

[Freund *et al.*, 2003] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[Gao and Zhou, 2010] W. Gao and Z.-H. Zhou. Approximation stability and boosting. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory*, pages 59–73, Canberra, Australia, 2010.

[Gao and Zhou, 2012] W. Gao and Z.-H. Zhou. On the consistency of AUC optimization. *CoRR*, 1208.0645v3, 2012.

[Herbrich *et al.*, 2000] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, Cambridge, MA, 2000.

[Kearns and Ron, 1999] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11:1427–1453, 1999.

[Kotlowski *et al.*, 2011] W. Kotlowski, K. Dembczynski, and E. Hüllermeier. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1113–1120, Bellevue, WA, 2011.

[Lehmann, 1975] E. L. Lehmann. *Nonparametrics: statistical methods based on ranks*. Holden-Day, San Francisco, 1975.

[Mukherjee *et al.*, 2006] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25:161–193, 2006.

[Rakhlin *et al.*, 2005] A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis Application*, 4:397–417, 2005.

[Rudin and Schapire, 2009] C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research*, 10:161–193, 2009.

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

[Shalev-Shwartz *et al.*, 2010] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010.

[Vapnik, 1998] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.

[Wang *et al.*, 2012] Y. Wang, R. Khardon, D. Pechyony, and R. Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 9.1–9.24, Edinburgh, Scotland, 2012.

[Weng and Lin, 2011] R. S. Weng and C.-J. Lin. A bayesian approximation method for online ranking. *Journal of Machine Learning Research*, 12:267–300, 2011.