Proceedings of the 18th Annual Conference on Learning Theory, 2005

Stability and Generalization of Bipartite Ranking Algorithms

Shivani Agarwal[1] and Partha Niyogi[2]

[1] Department of Computer Science, University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue, Urbana, IL 61801, USA
sagarwal@cs.uiuc.edu
[2] Departments of Computer Science and Statistics, University of Chicago
1100 E. 58th Street, Chicago, IL 60637, USA
niyogi@cs.uchicago.edu

Abstract. The problem of ranking, in which the goal is to learn a real-valued ranking function that induces a ranking or ordering over an instance space, has recently gained attention in machine learning. We study generalization properties of ranking algorithms, in a particular setting of the ranking problem known as the bipartite ranking problem, using the notion of algorithmic stability. In particular, we derive generalization bounds for bipartite ranking algorithms that have good stability properties. We show that kernel-based ranking algorithms that perform regularization in a reproducing kernel Hilbert space have such stability properties, and therefore our bounds can be applied to these algorithms; this is in contrast with previous generalization bounds for ranking, which are based on uniform convergence and in many cases cannot be applied to these algorithms. A comparison of the bounds we obtain with corresponding bounds for classification algorithms yields some interesting insights into the difference in generalization behaviour between ranking and classification.

1 Introduction

A central focus in learning theory research has been the study of generalization properties of learning algorithms. Perhaps the first work in this direction was that of Vapnik and Chervonenkis [1], who derived generalization bounds for classification algorithms based on uniform convergence. Since then, a large number of different tools have been developed for studying generalization, and have been applied successfully to analyze algorithms for both classification (learning of binary-valued functions) and regression (learning of real-valued functions), two of the most well-studied problems in machine learning. Recently, a new learning problem, namely that of *ranking*, has gained attention in the machine learning community [2–5]. In ranking, one learns a real-valued function that assigns scores to instances, but the scores themselves do not matter; instead, what is important is the relative ranking of instances induced by those scores. This problem is distinct from both classification and regression, and it is natural to ask what kinds of generalization properties hold for algorithms for this problem, and in particular, whether tools that have been applied to study generalization properties of classification and regression algorithms can be adapted to study generalization properties of ranking algorithms. It has been shown recently that generalization bounds based

on uniform convergence can be obtained for ranking algorithms in a particular setting of the ranking problem known as the *bipartite* ranking problem [5, 6]. In this paper, we ask whether such a result can be obtained using the notion of algorithmic stability, which has recently been used to derive generalization bounds for classification and regression algorithms, and which offers a different viewpoint than uniform convergence [7, 8].

1.1 Previous Results

The question of the generalization behaviour of ranking algorithms has only recently begun to be addressed. Generalization properties of algorithms for a distinct but closely related problem, namely that of ordinal regression, were considered in [3]. The first study of generalization in ranking was that of Freund et al. [5], in which generalization bounds for the bipartite RankBoost algorithm were derived. These bounds were derived from uniform convergence results for the classification error rate, and were expressed in terms of the VC-dimension of a class of binary classification functions derived from the class of ranking functions searched by RankBoost. More recently, Agarwal et al. [6] have derived a uniform convergence bound for the bipartite ranking error (see Section 2) which is expressed in terms of a new set of combinatorial parameters that measure directly the complexity of the class of ranking functions searched by an algorithm.

Uniform convergence requires the empirical errors of all functions in the searched class to converge to their expected errors. Generalization bounds based on uniform convergence are therefore necessarily loose, as they depend only on properties of the function class being searched, and do not take into account the manner in which the function class is actually searched by the algorithm. In addition, these bounds can be applied only to algorithms that search function classes of bounded complexity.

The notion of algorithmic stability, first studied for learning algorithms by Devroye and Wagner [9], has been used recently to directly obtain generalization bounds, without needing to show uniform convergence, for classification and regression algorithms that satisfy certain stability conditions [7, 8]. In particular, a stable learning algorithm is one whose output does not change much with small changes in the training sample; the above works have shown that classification and regression algorithms that satisfy this condition have good generalization properties. The stability-based bounds depend on properties of the algorithm rather than the function class that is searched, and can be applied also to algorithms that search function classes of unbounded complexity. Algorithms that have been shown to be stable include, for example, kernel-based classification and regression algorithms such as support vector machines (SVMs), which often cannot be analyzed using uniform convergence tools. In this paper, we show that the notion of algorithmic stability can be used also to analyze the generalization behaviour of (bipartite) ranking algorithms.

1.2 Our Results

We define notions of stability for bipartite ranking algorithms, and use these notions to analyze the generalization behaviour of such algorithms. In particular, we derive generalization bounds for bipartite ranking algorithms that exhibit good stability properties.

We show that kernel-based ranking algorithms that perform regularization in a reproducing kernel Hilbert space (RKHS) have such stability properties, and therefore our bounds can be applied to these algorithms; this is in contrast with previous generalization bounds for ranking, which are based on uniform convergence and in many cases cannot be applied to these algorithms. A comparison of the bounds we obtain with corresponding bounds for classification algorithms yields some interesting insights into the difference in generalization behaviour between ranking and classification. In particular, we find that for a training sample of M elements containing m positive and $n = M - m$ negative instances, the sample size M in the classification bounds is replaced with the quantity $mn/(m+n)$ in the ranking bounds. If we define the 'positive skew' of the sample as the proportion of positive examples $\rho = m/(m+n)$, then this means that the'effective' sample size in ranking is reduced from M to $\rho(1-\rho)M$, with the reduction being more drastic as ρ departs from $1/2$, *i.e.*, as the balance between positive and negative examples becomes more uneven. This further corroborates previous observations about the importance of the skew ρ in ranking [10, 6, 11].

1.3 Organization

We describe the bipartite ranking problem in detail in Section 2, and define notions of stability for (bipartite) ranking algorithms in Section 3. Using these notions, we derive generalization bounds for stable ranking algorithms in Section 4. In Section 5 we show stability of kernel-based ranking algorithms that perform regularization in an RKHS, and apply the results of Section 4 to obtain generalization bounds for these algorithms. We conclude with a discussion in Section 6.

2 The Bipartite Ranking Problem

In the bipartite ranking problem [5, 6], instances come from two categories, positive and negative; the learner is given examples of instances labeled as positive or negative, and the goal is to learn a ranking in which positive instances are ranked higher than negative ones. Such problems arise, for example, in information retrieval, where one is interested in retrieving documents from some database that are 'relevant' to a given topic; in this case, the training examples given to the learner consist of documents labeled as relevant (positive) or irrelevant (negative), and the goal is to produce a list of documents that contains relevant documents at the top and irrelevant documents at the bottom – in other words, one wants a ranking of the documents such that relevant documents are ranked higher than irrelevant documents.

Formally, the setting of the bipartite ranking problem can be described as follows. There is an instance space $\mathcal{X}$ from which instances are drawn, and the learner is given a training sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ consisting of a sequence of positive training examples $S_+ = (x_1^+, \ldots, x_m^+)$ and a sequence of negative training examples $S_- = (x_1^-, \ldots, x_n^-)$. The goal is to learn from these examples a real-valued ranking function $f : \mathcal{X} \to \mathbb{R}$ that ranks future positive instances higher than negative ones, where f is considered to rank an instance x higher than an instance x' if $f(x) > f(x')$ and is considered to rank x lower than x' if $f(x) < f(x')$. We assume that positive instances

are drawn randomly and independently according to some (unknown) distribution $\mathcal{D}_+$ on the instance space $\mathcal{X}$, and that negative instances are drawn randomly and independently according to some (unknown) distribution $\mathcal{D}_-$ on $\mathcal{X}$. The quality of a ranking function $f : \mathcal{X} \to \mathbb{R}$ is then measured by its *expected ranking error*, denoted by $R(f)$ and defined as follows:

$$R(f) = \mathbf{E}_{x^+ \sim \mathcal{D}_+, x^- \sim \mathcal{D}_-} \left\{ \mathbf{I}_{\{f(x^+) < f(x^-)\}} + \frac{1}{2} \mathbf{I}_{\{f(x^+) = f(x^-)\}} \right\}, \tag{1}$$

where $\mathbf{I}_{\{\cdot\}}$ denotes the indicator variable whose value is one if its argument is true and zero otherwise. The expected error $R(f)$ is the probability that a positive instance drawn randomly according to $\mathcal{D}_+$ is ranked lower by f than a negative instance drawn randomly according to $\mathcal{D}_-$, assuming that ties are broken uniformly at random. In practice, since the distributions $\mathcal{D}_+$ and $\mathcal{D}_-$ are unknown, the expected error of a ranking function f must be estimated from an empirically observable quantity such as its *empirical ranking error* with respect to a sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, denoted by $\hat{R}(f; S_+, S_-)$ and defined as follows:

$$\hat{R}(f; S_+, S_-) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \left\{ \mathbf{I}_{\{f(x_i^+) < f(x_j^-)\}} + \frac{1}{2} \mathbf{I}_{\{f(x_i^+) = f(x_j^-)\}} \right\}. \tag{2}$$

This is simply the fraction of positive-negative pairs in (S_+, S_-) that are ranked incorrectly by f, assuming that ties are broken uniformly at random.

Although the bipartite ranking problem shares similarities with the binary classification problem, it should be noted that the two problems are in fact distinct. In particular, it is possible for binary-valued functions obtained by thresholding different real-valued functions to have the same classification errors, while the ranking errors of the real-valued functions may differ significantly. For example, consider the following two rankings on a sample consisting of 4 positive and 4 negative examples:

In both cases, the error of the best classification function that can be obtained by applying a threshold is $2/8$. However, the ranking error of f_1 is $4/16$, whereas that of f_2 is $8/16$. For a detailed analysis of this distinction, see [10][3].

A bipartite ranking algorithm takes as input a training sample $(S_+, S_-) \in \left(\bigcup_{m=1}^{\infty} \mathcal{X}^m\right) \times \left(\bigcup_{n=1}^{\infty} \mathcal{X}^n\right)$ and returns as output a ranking function $f_{S_+,S_-} : \mathcal{X} \to \mathbb{R}$. For simplicity, we consider only deterministic algorithms. We are concerned in this paper with generalization properties of such algorithms; in particular, we are interested in bounding the expected error of a learned ranking function in terms of an empirically observable quantity such as its empirical error on the training sample from which it is learned. The following definitions will be useful in our study.

Definition 1 (Ranking loss function). *A* ranking loss function *is a function $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ that assigns, for each $f : \mathcal{X} \to \mathbb{R}$ and $x, x' \in \mathcal{X}$, a non-negative real number $\ell(f, x, x')$ interpreted as the loss of f in its relative ranking of x and x'.*

[3] In [10], the performance of a ranking function is measured in terms of the area under the ROC curve (AUC); this quantity is simply equal to one minus the empirical ranking error.

Definition 2 (Expected ℓ-error). *Let $f : \mathcal{X} \to \mathbb{R}$ be a ranking function on $\mathcal{X}$. Let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss function. Define the* expected ℓ-error *of f, denoted by $R_\ell(f)$, as follows:*

$$R_\ell(f) = \mathbf{E}_{x^+ \sim \mathcal{D}_+, x^- \sim \mathcal{D}_-} \left\{ \ell(f, x^+, x^-) \right\} .$$

Definition 3 (Empirical ℓ-error). *Let $f : \mathcal{X} \to \mathbb{R}$ be a ranking function on $\mathcal{X}$, and let $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ be a finite sample. Let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss function. Define the* empirical ℓ-error *of f with respect to S_+ and S_-, denoted by $\hat{R}_\ell(f; S_+, S_-)$, as follows:*

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

Comparing with Eqs. (1-2), we see that the ranking error can be expressed as the ℓ_b-error, *i.e.*, $R \equiv R_{\ell_b}$ and $\hat{R} \equiv \hat{R}_{\ell_b}$, where ℓ_b is the *bipartite ranking loss* given by

$$\ell_b(f, x, x') = \mathbf{I}_{\{f(x) < f(x')\}} + \frac{1}{2} \mathbf{I}_{\{f(x) = f(x')\}} . \tag{3}$$

3 Stability of (Bipartite) Ranking Algorithms

A stable algorithm is one whose output does not change significantly with small changes in the input. The input to a ranking algorithm is a training sample of the form $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ for some $m, n \in \mathbb{N}$; we consider changes to such a sample that consist of replacing a single element of the sample with a new instance. For any $i \in \{1, \ldots, m\}$ and $z \in \mathcal{X}$, we use $S_+^{i,z}$ to denote the sequence obtained from S_+ by replacing x_i^+ with z; similarly, for any $j \in \{1, \ldots, n\}$ and $z \in \mathcal{X}$, we use $S_-^{j,z}$ to denote the sequence obtained from S_- by replacing x_j^- with z.

Several different notions of stability have been used in the study of classification and regression algorithms [9, 12, 7, 8, 13]. The notions of stability that we define for ranking algorithms below are most closely related to those used by Bousquet and Elisseeff [7].

Definition 4 (Uniform loss stability). *Let L be a bipartite ranking algorithm whose output on a training sample (S_+, S_-) we denote by f_{S_+,S_-}, and let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss function. Let $\alpha : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, $\beta : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$. We say that L has* uniform loss stability (α, β) *with respect to ℓ if for all $m, n \in \mathbb{N}$, $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, $z \in \mathcal{X}$, $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, we have for all $x^+, x^- \in \mathcal{X}$,*

$$\left| \ell(f_{S_+,S_-}, x^+, x^-) - \ell(f_{S_+^{i,z},S_-}, x^+, x^-) \right| \leq \alpha(m, n) ,$$
$$\left| \ell(f_{S_+,S_-}, x^+, x^-) - \ell(f_{S_+,S_-^{j,z}}, x^+, x^-) \right| \leq \beta(m, n) .$$

Definition 5 (Uniform score stability). *Let L be a bipartite ranking algorithm whose output on a training sample (S_+, S_-) we denote by f_{S_+,S_-}. Let $\mu : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, $\nu : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$. We say that L has* uniform score stability (μ, ν) *if for all $m, n \in \mathbb{N}$, $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, $z \in \mathcal{X}$, $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, we have for all $x \in \mathcal{X}$,*

$$\left| f_{S_+,S_-}(x) - f_{S_+^{i,z},S_-}(x) \right| \leq \mu(m, n) ,$$
$$\left| f_{S_+,S_-}(x) - f_{S_+,S_-^{j,z}}(x) \right| \leq \nu(m, n) .$$

4 Generalization Bounds for Stable Ranking Algorithms

In this section we derive generalization bounds for ranking algorithms that exhibit good stability properties. Our methods are based on those of Bousquet and Elisseeff [7], who derived such bounds for classification and regression algorithms. We start with the following technical lemma.

Lemma 1. *Let L be a symmetric bipartite ranking algorithm*[4] *whose output on a training sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ we denote by f_{S_+,S_-}, and let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss function. Then for all $i \in \{1,\ldots,m\}$, $j \in \{1,\ldots,n\}$, we have*

$$\begin{aligned}&\mathbf{E}_{S_+,S_-}\Big\{R_\ell(f_{S_+,S_-}) - \hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\Big\}\\ &\qquad = \mathbf{E}_{S_+,S_-,x^+,x^-}\Big\{\ell(f_{S_+,S_-}, x^+, x^-) - \ell(f_{S_+^{i,x^+},S_-^{j,x^-}}, x^+, x^-)\Big\}.\end{aligned}$$

Proof. We have,

$$\mathbf{E}_{S_+,S_-}\Big\{\hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\Big\} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{E}_{S_+,S_-}\Big\{\ell(f_{S_+,S_-}, x_i^+, x_j^-)\Big\}.$$

By symmetry, the term in the summation is the same for all i, j. Therefore, for each $i \in \{1,\ldots,m\}$ and $j \in \{1,\ldots,n\}$, we get

$$\begin{aligned}\mathbf{E}_{S_+,S_-}\Big\{\hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\Big\} &= \mathbf{E}_{S_+,S_-}\Big\{\ell(f_{S_+,S_-}, x_i^+, x_j^-)\Big\}\\ &= \mathbf{E}_{S_+,S_-,x^+,x^-}\Big\{\ell(f_{S_+,S_-}, x_i^+, x_j^-)\Big\}.\end{aligned}$$

Interchanging the roles of x_i^+ with x^+ and x_j^- with x^-, we get

$$\mathbf{E}_{S_+,S_-}\Big\{\hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\Big\} = \mathbf{E}_{S_+,S_-,x^+,x^-}\Big\{\ell(f_{S_+^{i,x^+},S_-^{j,x^-}}, x^+, x^-)\Big\}.$$

Since by definition

$$\mathbf{E}_{S_+,S_-}\Big\{R_\ell(f_{S_+,S_-})\Big\} = \mathbf{E}_{S_+,S_-,x^+,x^-}\Big\{\ell(f_{S_+,S_-}, x^+, x^-)\Big\},$$

the result follows. □

Our main tool will be the following powerful concentration inequality of McDiarmid [14], which bounds the deviation of any function of a sample for which a single change in the sample has limited effect.

Theorem 1 (McDiarmid [14]). *Let $X_1,\ldots,X_N$ be independent random variables, each taking values in a set A. Let $\phi : A^N \to \mathbb{R}$ be such that for each $k \in \{1,\ldots,N\}$, there exists $c_k > 0$ such that*

$$\sup_{x_1,\ldots,x_N \in A, x_k' \in A}\big|\phi(x_1,\ldots,x_N) - \phi(x_1,\ldots,x_{k-1},x_k',x_{k+1},\ldots,x_N)\big| \le c_k.$$

Then for any $\epsilon > 0$,

$$\mathbf{P}\big\{\phi(X_1,\ldots,X_N) - \mathbf{E}\big\{\phi(X_1,\ldots,X_N)\big\} \ge \epsilon\big\} \le e^{-2\epsilon^2/\sum_{k=1}^{N} c_k^2}.$$

We are now ready to give our main result, which bounds the expected ℓ-error of a ranking function learned by an algorithm with good uniform loss stability in terms of its empirical ℓ-error on the training sample.

[4] A symmetric bipartite ranking algorithm is one whose output is independent of the order of elements in the training sequences S_+ and S_-.

Theorem 2. *Let L be a symmetric bipartite ranking algorithm whose output on a training sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ we denote by f_{S_+,S_-}, and let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss function such that $0 \leq \ell(f, x, x') \leq B$ for all $f : \mathcal{X} \to \mathbb{R}$ and $x, x' \in \mathcal{X}$. Let $\alpha : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, $\beta : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ be such that L has uniform loss stability (α, β) with respect to ℓ. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$ over the draw of (S_+, S_-),*

$$R_\ell(f_{S_+,S_-}) < \hat{R}_\ell(f_{S_+,S_-}; S_+, S_-) + \alpha(m,n) + \beta(m,n) + \sqrt{\frac{\left\{n(2m\,\alpha(m,n) + B)^2 + m(2n\,\beta(m,n) + B)^2\right\} \ln(1/\delta)}{2mn}}\,.$$

Proof. Let $\phi : \mathcal{X}^m \times \mathcal{X}^n \to \mathbb{R}$ be defined as follows:

$$\phi(S_+, S_-) = R_\ell(f_{S_+,S_-}) - \hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\,.$$

We shall show that ϕ satisfies the conditions of McDiarmid's inequality (Theorem 1). Let $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, and let $z \in \mathcal{X}$. For each $i \in \{1, \ldots, m\}$, we have

$$\begin{aligned}
\left|R_\ell(f_{S_+,S_-}) - R_\ell(f_{S_+^{i,z},S_-})\right| &= \left|\mathbf{E}_{x^+,x^-}\left\{\ell(f_{S_+,S_-}, x^+, x^-) - \ell(f_{S_+^{i,z},S_-}, x^+, x^-)\right\}\right| \\
&\leq \mathbf{E}_{x^+,x^-}\left\{\left|\ell(f_{S_+,S_-}, x^+, x^-) - \ell(f_{S_+^{i,z},S_-}, x^+, x^-)\right|\right\} \\
&\leq \alpha(m,n)\,,
\end{aligned}$$

and

$$\begin{aligned}
&\left|\hat{R}_\ell(f_{S_+,S_-}; S_+, S_-) - \hat{R}_\ell(f_{S_+^{i,z},S_-}; S_+^{i,z}, S_-)\right| \\
&\leq \frac{1}{mn}\sum_{i' \neq i}\sum_{j=1}^{n}\left|\ell(f_{S_+,S_-}, x_{i'}^+, x_j^-) - \ell(f_{S_+^{i,z},S_-}, x_{i'}^+, x_j^-)\right| \\
&\quad + \frac{1}{mn}\sum_{j=1}^{n}\left|\ell(f_{S_+,S_-}, x_i^+, x_j^-) - \ell(f_{S_+^{i,z},S_-}, z, x_j^-)\right| \\
&\leq \alpha(m,n) + \frac{B}{m}\,.
\end{aligned}$$

This gives

$$\left|\phi(S_+, S_-) - \phi(S_+^{i,z}, S_-)\right| \leq 2\alpha(m,n) + \frac{B}{m}\,.$$

Similarly, it can be shown that for each $j \in \{1, \ldots, n\}$,

$$\left|\phi(S_+, S_-) - \phi(S_+, S_-^{j,z})\right| \leq 2\beta(m,n) + \frac{B}{n}\,.$$

Thus, applying McDiarmid's inequality to ϕ, we get for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbf{P}_{S_+,S_-}\left\{\left\{R_\ell(f_{S_+,S_-}) - \hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\right\}\right. \\
&\qquad\qquad \left. - \mathbf{E}_{S_+,S_-}\left\{R_\ell(f_{S_+,S_-}) - \hat{R}_\ell(f_{S_+,S_-}; S_+, S_-)\right\} \geq \epsilon\right\} \\
&\leq e^{-2\epsilon^2/(m(2\alpha(m,n)+B/m)^2 + n(2\beta(m,n)+B/n)^2)} \\
&= e^{-2mn\epsilon^2/(n(2m\alpha(m,n)+B)^2 + m(2n\beta(m,n)+B)^2)}\,.
\end{aligned}$$

Now, by Lemma 1, we have

$$\begin{aligned}
&\mathbf{E}_{S_+,S_-}\left\{R_\ell(f_{S_+,S_-}) - \hat{R}_\ell(f_{S_+,S_-};S_+,S_-)\right\} \\
&\quad= \mathbf{E}_{S_+,S_-,x^+,x^-}\left\{\ell(f_{S_+,S_-},x^+,x^-) - \ell(f_{S_+^{i,x^+},S_-^{j,x^-}},x^+,x^-)\right\} \\
&\quad= \mathbf{E}_{S_+,S_-,x^+,x^-}\Big\{\ell(f_{S_+,S_-},x^+,x^-) - \ell(f_{S_+^{i,x^+},S_-},x^+,x^-) \\
&\qquad\qquad + \ell(f_{S_+^{i,x^+},S_-},x^+,x^-) - \ell(f_{S_+^{i,x^+},S_-^{j,x^-}},x^+,x^-)\Big\} \\
&\quad\le \mathbf{E}_{S_+,S_-,x^+,x^-}\left\{\left|\ell(f_{S_+,S_-},x^+,x^-) - \ell(f_{S_+^{i,x^+},S_-},x^+,x^-)\right|\right\} \\
&\qquad + \mathbf{E}_{S_+,S_-,x^+,x^-}\left\{\left|\ell(f_{S_+^{i,x^+},S_-},x^+,x^-) - \ell(f_{S_+^{i,x^+},S_-^{j,x^-}},x^+,x^-)\right|\right\} \\
&\quad\le \alpha(m,n) + \beta(m,n)\,.
\end{aligned}$$

Thus we get for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbf{P}_{S_+,S_-}\left\{R_\ell(f_{S_+,S_-}) - \hat{R}_\ell(f_{S_+,S_-};S_+,S_-) - \big(\alpha(m,n)+\beta(m,n)\big) \ge \epsilon\right\} \\
&\le e^{-2mn\epsilon^2/(n(2m\alpha(m,n)+B)^2+m(2n\beta(m,n)+B)^2)}\,.
\end{aligned}$$

The result follows by setting the right hand side equal to δ and solving for ϵ. □

Theorem 2 gives meaningful bounds when $\alpha(m,n) = o(1/\sqrt{m})$ and $\beta(m,n) = o(1/\sqrt{n})$. This means the theorem cannot be applied directly to obtain bounds on the expected ranking error, since it is not possible to have non-trivial uniform loss stability with respect to the bipartite ranking loss ℓ_b (except by an algorithm that picks the same ranking function for all training samples of a given size m, n). However, for any ranking loss ℓ that satisfies $\ell_b \le \ell$, Theorem 2 can be applied to ranking algorithms that have good uniform loss stability with respect to ℓ to obtain bounds on the expected ℓ-error; since in this case $R \le R_\ell$, these bounds apply also to the expected ranking error. We consider below a specific ranking loss that satisfies this condition.

For $\gamma > 0$, let the γ *ranking loss*, denoted by ℓ_γ, be defined as follows:

$$\ell_\gamma(f,x,x') = \begin{cases} 1 & \text{if } (f(x) - f(x')) \le 0 \\ 1 - \frac{(f(x)-f(x'))}{\gamma} & \text{if } 0 < (f(x) - f(x')) < \gamma \\ 0 & \text{if } (f(x) - f(x')) \ge \gamma \end{cases}. \tag{4}$$

Clearly, for all $\gamma > 0$, we have $\ell_b \le \ell_\gamma$. Therefore, for any ranking algorithm that has good uniform loss stability with respect to ℓ_γ for some $\gamma > 0$, Theorem 2 can be applied to bound the expected ranking error of a learned ranking function in terms of its empirical ℓ_γ-error on the training sample. The following lemma shows that, for every $\gamma > 0$, a ranking algorithm that has good uniform score stability also has good uniform loss stability with respect to ℓ_γ.

Lemma 2. *Let L be a bipartite ranking algorithm whose output on a training sample (S_+, S_-) we denote by f_{S_+,S_-}. Let $\mu : \mathbb{N}\times\mathbb{N}\to\mathbb{R}$, $\nu : \mathbb{N}\times\mathbb{N}\to\mathbb{R}$ be such that L has uniform score stability (μ,ν). Then for every $\gamma > 0$, L has uniform loss stability $(\alpha_\gamma, \beta_\gamma)$ with respect to the γ ranking loss ℓ_γ, where for all $m, n \in \mathbb{N}$,*

$$\alpha_\gamma(m,n) = \frac{2\mu(m,n)}{\gamma}\,, \qquad \beta_\gamma(m,n) = \frac{2\nu(m,n)}{\gamma}\,.$$

Proof. By the definition of ℓ_γ in Eq. (4), we have that

$$\ell_\gamma(f, x, x') \leq 1 - \tfrac{(f(x)-f(x')}{\gamma} \quad \text{if } (f(x) - f(x')) \leq 0\,, \tag{5}$$

$$\ell_\gamma(f, x, x') \geq 1 - \tfrac{(f(x)-f(x')}{\gamma} \quad \text{if } (f(x) - f(x')) \geq \gamma\,. \tag{6}$$

Now, let $m, n \in \mathbb{N}$, $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, $z \in \mathcal{X}$, $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, and let $x^+, x^- \in \mathcal{X}$. The case $\ell_\gamma(f_{S_+,S_-}, x^+, x^-) = \ell_\gamma(f_{S_+^{i,z},S_-}, x^+, x^-)$ is trivial. Assume $\ell_\gamma(f_{S_+,S_-}, x^+, x^-) \neq \ell_\gamma(f_{S_+^{i,z},S_-}, x^+, x^-)$. Then, using the observations in Eqs. (5-6), it can be verified that

$$\begin{aligned}
&\left|\ell_\gamma(f_{S_+,S_-}, x^+, x^-) - \ell_\gamma(f_{S_+^{i,z},S_-}, x^+, x^-)\right| \\
&\leq \left|\left(1 - \frac{(f_{S_+,S_-}(x^+) - f_{S_+,S_-}(x^-))}{\gamma}\right) - \left(1 - \frac{(f_{S_+^{i,z},S_-}(x^+) - f_{S_+^{i,z},S_-}(x^-))}{\gamma}\right)\right| \\
&\leq \frac{1}{\gamma}\left(\left|f_{S_+,S_-}(x^+) - f_{S_+^{i,z},S_-}(x^+)\right| + \left|f_{S_+,S_-}(x^-) - f_{S_+^{i,z},S_-}(x^-)\right|\right) \\
&\leq \frac{2\mu(m,n)}{\gamma}\,.
\end{aligned}$$

Similarly, it can be shown that

$$\left|\ell_\gamma(f_{S_+,S_-}, x^+, x^-) - \ell_\gamma(f_{S_+,S_-^{j,z}}, x^+, x^-)\right| \leq \frac{2\nu(m,n)}{\gamma}\,.$$

The result follows. □

Putting everything together, we thus get the following result which bounds the expected ranking error of a learned ranking function in terms of its empirical ℓ_γ-error for any ranking algorithm that has good uniform score stability.

Theorem 3. *Let L be a symmetric bipartite ranking algorithm whose output on a training sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$ we denote by f_{S_+,S_-}. Let $\mu : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, $\nu : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ be such that L has uniform score stability (μ, ν), and let $\gamma > 0$. Then for any $0 < \delta < 1$, with probability at least $1 - \delta$ over the draw of (S_+, S_-),*

$$\begin{aligned}
R(f_{S_+,S_-}) < \hat{R}_{\ell_\gamma}(f_{S_+,S_-}; S_+, S_-) + \frac{2\mu(m,n)}{\gamma} + \frac{2\nu(m,n)}{\gamma} \\
+ \sqrt{\frac{\left\{n\left(\frac{4m\,\mu(m,n)}{\gamma} + 1\right)^2 + m\left(\frac{4n\,\nu(m,n)}{\gamma} + 1\right)^2\right\}\ln(1/\delta)}{2mn}}\,.
\end{aligned}$$

Proof. The result follows by applying Theorem 2 to L with the ranking loss ℓ_γ (using Lemma 2), which satisfies $0 \leq \ell_\gamma \leq 1$, and from the fact that $R \leq R_{\ell_\gamma}$. □

We note that although our bounds above are derived for the case when a fixed number m of positive examples are drawn i.i.d. from $\mathcal{D}_+$ and a fixed number n of negative examples are drawn i.i.d. from $\mathcal{D}_-$, the bounds can be extended easily to the case when

M examples are drawn i.i.d. from a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$. In particular, using exactly the same techniques as above, the same confidence intervals can be derived for a draw conditioned on any fixed label sequence that contains m positive and $n = M - m$ negative labels. The conditioning can then be removed using an expectation trick (see [6, Theorems 8 and 19]); in the resulting confidence intervals, the numbers m and n are replaced by functions of the (random) label sequence that correspond to the numbers of positive and negative labels drawn.

5 Stable Ranking Algorithms

In this section we show stability of certain ranking algorithms that select a ranking function by minimizing a regularized objective function. We start by deriving a general result for regularization-based ranking algorithms in Section 5.1. In Section 5.2 we use this result to show stability of kernel-based ranking algorithms that perform regularization in a reproducing kernel Hilbert space (RKHS). We show, in particular, stability of an SVM-like ranking algorithm, and apply the results of Section 4 to obtain a generalization bound for this algorithm. A comparison with the uniform convergence bound of [6] demonstrates the benefit of the stability analysis. Again, our methods are based on those of Bousquet and Elisseeff [7], who showed similar results for classification and regression algorithms.

5.1 General Regularizers

Given a ranking loss function $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$, a class $\mathcal{F}$ of real-valued functions on $\mathcal{X}$, and a regularization functional $N : \mathcal{F} \to \mathbb{R}^+ \cup \{0\}$, consider the following regularized empirical ℓ-error of a ranking function $f \in \mathcal{F}$ (with respect to a sample $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$), with regularization parameter $\lambda > 0$:

$$\hat{R}^{\lambda}_{\ell}(f; S_+, S_-) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \ell(f, x_i^+, x_j^-) + \lambda N(f) \,. \tag{7}$$

We consider bipartite ranking algorithms that minimize such a regularized objective function, *i.e.*, ranking algorithms that, given a training sample (S_+, S_-), output a ranking function $f_{S_+,S_-} \in \mathcal{F}$ that satisfies

$$\begin{aligned} f_{S_+,S_-} &= \arg\min_{f \in \mathcal{F}} \; \hat{R}^{\lambda}_{\ell}(f; S_+, S_-) \\ &= \arg\min_{f \in \mathcal{F}} \; \left\{ \hat{R}_{\ell}(f; S_+, S_-) + \lambda N(f) \right\}, \end{aligned} \tag{8}$$

for some fixed choice of ranking loss ℓ, function class $\mathcal{F}$, regularizer N, and regularization parameter λ. We derive below a general result that will be useful for showing stability of such regularization-based algorithms.

Definition 6 (σ-admissibility). *Let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss and $\mathcal{F}$ a class of real-valued functions on $\mathcal{X}$. Let $\sigma > 0$. We say that ℓ is* σ-admissible *with respect to $\mathcal{F}$ if for all $f_1, f_2 \in \mathcal{F}$ and all $x, x' \in \mathcal{X}$, we have*

$$\left|\ell(f_1, x, x') - \ell(f_2, x, x')\right| \leq \sigma\Big(\left|f_1(x) - f_2(x)\right| + \left|f_1(x') - f_2(x')\right|\Big) \,.$$

Lemma 3. *Let $\ell : \mathbb{R}^{\mathcal{X}} \times \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss such that $\ell(f, x, x')$ is convex in f. Let $\mathcal{F}$ be a convex class of real-valued functions on $\mathcal{X}$, and let $\sigma > 0$ be such that ℓ is σ-admissible with respect to $\mathcal{F}$. Let $\lambda > 0$, and let $N : \mathcal{F} \to \mathbb{R}^+ \cup \{0\}$ be a functional defined on $\mathcal{F}$ such that for all samples $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, the regularized empirical ℓ-error $\hat{R}^{\lambda}_{\ell}(f; S_+, S_-)$ has a minimum (not necessarily unique) in $\mathcal{F}$. Let L be a ranking algorithm defined by Eq. (8), and let $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, $z \in \mathcal{X}$, $i \in \{1, \ldots, m\}$, and $j \in \{1, \ldots, n\}$. For brevity, denote*

$$f \;\equiv\; f_{S_+,S_-}, \qquad f_+^{i,z} \;\equiv\; f_{S_+^{i,z},S_-}, \qquad f_-^{j,z} \;\equiv\; f_{S_+,S_-^{j,z}},$$

and let

$$\Delta f_+ \;=\; \left(f_+^{i,z} - f\right), \qquad \Delta f_- \;=\; \left(f_-^{j,z} - f\right).$$

Then we have that for any $t \in [0, 1]$,

$$\begin{aligned} &N(f) - N(f + t\Delta f_+) + N(f_+^{i,z}) - N(f_+^{i,z} - t\Delta f_+) \\ &\qquad\qquad \leq \frac{t\sigma}{\lambda mn} \sum_{j=1}^{n} \left(|\Delta f_+(x_i^+)| + 2|\Delta f_+(x_j^-)| + |\Delta f_+(z)| \right), \\ &N(f) - N(f + t\Delta f_-) + N(f_-^{j,z}) - N(f_-^{j,z} - t\Delta f_-) \\ &\qquad\qquad \leq \frac{t\sigma}{\lambda mn} \sum_{i=1}^{m} \left(|\Delta f_-(x_j^-)| + 2|\Delta f_-(x_i^+)| + |\Delta f_-(z)| \right). \end{aligned}$$

The proof of this result makes use of techniques similar to those used in [7], and is omitted for lack of space (see [15] for details). As we show below, this result can be used to establish stability of certain regularization-based ranking algorithms.

5.2 Regularization in Hilbert Spaces

Let $\mathcal{F}$ be an RKHS with kernel K. Then from the properties of an RKHS (see, for example, [16]), we have for all $f \in \mathcal{F}$ and all $x \in \mathcal{X}$,

$$|f(x)| \leq \|f\|_K \sqrt{K(x, x)}\,. \tag{9}$$

Let $N : \mathcal{F} \to \mathbb{R}^+ \cup \{0\}$ be the regularizer defined by

$$N(f) = \|f\|_K^2\,. \tag{10}$$

We show below that, if the kernel K is such that $K(x, x)$ is bounded for all $x \in \mathcal{X}$, then a ranking algorithm that minimizes an appropriate regularized error over $\mathcal{F}$, with regularizer N as defined above, has good uniform score stability.

Theorem 4. *Let $\mathcal{F}$ be an RKHS with kernel K such that for all $x \in \mathcal{X}$, $K(x, x) \leq \kappa^2 < \infty$. Let ℓ be a ranking loss such that $\ell(f, x, x')$ is convex in f and ℓ is σ-admissible with respect to $\mathcal{F}$. Let $\lambda > 0$, and let N be given by Eq. (10). Let L be a ranking algorithm defined by Eq. (8). Then L has uniform score stability (μ, ν), where for all $m, n \in \mathbb{N}$,*

$$\mu(m, n) = \frac{4\sigma\kappa^2}{\lambda m}\,, \qquad \nu(m, n) = \frac{4\sigma\kappa^2}{\lambda n}\,.$$

Proof. Let $m, n \in \mathbb{N}$, $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, $z \in \mathcal{X}$, and $i \in \{1, \ldots, m\}$. Since $\mathcal{F}$ is a vector space, we have (using the notation of Lemma 3) that $\Delta f_+ \in \mathcal{F}$. Applying Lemma 3 with $t = 1/2$, we get that

$$\frac{1}{2}\|\Delta f_+\|_K^2 \leq \frac{\sigma}{2\lambda mn} \sum_{j=1}^{n} \Big(|\Delta f_+(x_i^+)| + 2|\Delta f_+(x_j^-)| + |\Delta f_+(z)| \Big) .$$

By Eq. (9), we thus get that

$$\|\Delta f_+\|_K^2 \leq \frac{4\sigma\kappa}{\lambda m} \|\Delta f_+\|_K ,$$

which gives

$$\|\Delta f_+\|_K \leq \frac{4\sigma\kappa}{\lambda m} . \tag{11}$$

Thus, by Eqs. (9) and (11), we have for all $x \in \mathcal{X}$,

$$\left| f_{S_+,S_-}(x) - f_{S_+^{i,z},S_-}(x) \right| = |\Delta f_+(x)| \leq \frac{4\sigma\kappa^2}{\lambda m} .$$

Similarly, for each $j \in \{1, \ldots, n\}$, we can show that

$$\left| f_{S_+,S_-}(x) - f_{S_+,S_-^{j,z}}(x) \right| \leq \frac{4\sigma\kappa^2}{\lambda n} .$$

The result follows. □

Consider now the following ranking loss function, which we refer to as the *hinge ranking loss* due to its similarity to the hinge loss in classification:

$$\ell_h(f, x, x') = \begin{cases} 1 - (f(x) - f(x')) & \text{if } (f(x) - f(x')) \ < \ 1 \\ 0 & \text{if } (f(x) - f(x')) \ \geq \ 1 \end{cases} . \tag{12}$$

We consider a ranking algorithm L that minimizes the regularized ℓ_h-error in an RKHS $\mathcal{F}$. Specifically, let L be a ranking algorithm which, given a training sample (S_+, S_-), outputs a ranking function $f_{S_+,S_-} \in \mathcal{F}$ that satisfies (for some fixed $\lambda > 0$)

$$f_{S_+,S_-} = \arg\min_{f \in \mathcal{F}} \left\{ \hat{R}_{\ell_h}(f; S_+, S_-) + \lambda \|f\|_K^2 \right\} . \tag{13}$$

We note that this algorithm has an equivalent quadratic programming formulation similar to SVMs in the case of classification (see [17, 15]). It can be verified that $\ell_h(f, x, x')$ is convex in f, and that ℓ_h is 1-admissible with respect to $\mathcal{F}$. Thus, if $K(x, x) \leq \kappa^2$ for all $x \in \mathcal{X}$, then from Theorem 4 we get that L has uniform score stability (μ, ν), where for all $m, n \in \mathbb{N}$,

$$\mu(m, n) = \frac{4\kappa^2}{\lambda m} , \qquad \nu(m, n) = \frac{4\kappa^2}{\lambda n} .$$

Applying Theorem 3 with $\gamma = 1$, we then get that for any $0 < \delta < 1$, with probability at least $1 - \delta$ over the draw of $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, the expected ranking error of the ranking function f_{S_+,S_-} learned by the above algorithm L is bounded by

$$R(f_{S_+,S_-}) < \hat{R}_{\ell_1}(f_{S_+,S_-}; S_+, S_-) + \frac{8\kappa^2}{\lambda} \Big(\frac{m+n}{mn} \Big) + \Big(1 + \frac{16\kappa^2}{\lambda} \Big) \sqrt{\frac{(m+n)\ln(1/\delta)}{2mn}} . \tag{14}$$

In particular, for the RKHS corresponding to the linear kernel defined on the unit ball in $\mathbb{R}^d$, so that $K(\mathbf{x}, \mathbf{x}) \leq 1$ for all $\mathbf{x}$, we have that with probability at least $1 - \delta$ over the draw of $(S_+, S_-) \in \mathcal{X}^m \times \mathcal{X}^n$, the ranking function f_{S_+,S_-} learned by the above algorithm (defined by Eq. (13)) satisfies

$$R(f_{S_+,S_-}) < \hat{R}_{\ell_1}(f_{S_+,S_-}; S_+, S_-) + \frac{8}{\lambda}\Big(\frac{m+n}{mn}\Big) + \Big(1 + \frac{16}{\lambda}\Big)\sqrt{\frac{(m+n)\ln(1/\delta)}{2mn}}\,.$$

On the other hand, the confidence interval obtained for this algorithm using the uniform convergence bound of [6] gives that, with probability at least $1 - \delta$,

$$R(f_{S_+,S_-}) < \hat{R}(f_{S_+,S_-}; S_+, S_-) + \sqrt{\frac{8(m+n)\big(d(\ln(8mn/d)+1) + \ln(4/\delta)\big)}{mn}}\,.$$

The above bounds are plotted in Figure 1 for $\delta = 0.01$, $\lambda = 1$, and various values of d and $m/(m+n)$. As can be seen, directly analyzing stability properties of the algorithm gives considerable benefit over the general uniform convergence based analysis. In particular, since the uniform convergence bound depends on the complexity of the function class that is searched, the bound quickly becomes uninformative in high dimensions; on the other hand, the stability bound is independent of the dimensionality of the space. In the case of kernel spaces whose complexity cannot be bounded, *e.g.*, the RKHS corresponding to the Gaussian kernel, the uniform convergence bound cannot be applied at all, while the stability analysis continues to hold.

Comparing the bound derived in Eq. (14) to the corresponding bound for classification derived by Bousquet and Elisseeff [7], we find that if the total number of training examples is denoted by $M = m + n$, then the sample size M in their bound is replaced by the quantity $mn/(m+n)$ in our bound.[5] If we define the 'positive skew' of the sample as the proportion of positive examples $\rho = m/(m+n)$, then this is equivalent to replacing M in the classification bound with $\rho(1-\rho)M$ in our bound. The 'effective' sample size in ranking is thus reduced from M to $\rho(1-\rho)M$, the reduction being more drastic as the skew ρ departs from $1/2$. Interestingly, a similar observation holds for the uniform convergence and large deviation bounds for the ranking error derived in [6] when compared to corresponding bounds for the classification error.

As in the case of classification [7], the above results show that a larger regularization parameter λ leads to better stability and, therefore, a tighter confidence interval in the resulting generalization bound. In particular, one must have $\lambda \gg \sqrt{(m+n)/mn}$ in order for the above bound to be meaningful.

6 Discussion

The main difference in the mathematical formulation of the (bipartite) ranking problem as compared to the classification problem is that the loss function in ranking is 'pairwise' rather than 'point-wise'. The general analysis of ranking is otherwise similar to

[5] The difference in constants in the two bounds is due in part to the difference in loss functions in ranking and classification, and in part to a slight difference in definitions of stability; in particular, our definitions are in terms of changes to a training sample that consist of replacing one element in the sample with a new one, while the definitions of Bousquet and Elisseeff are in terms of changes that consist of removing one element from the sample.

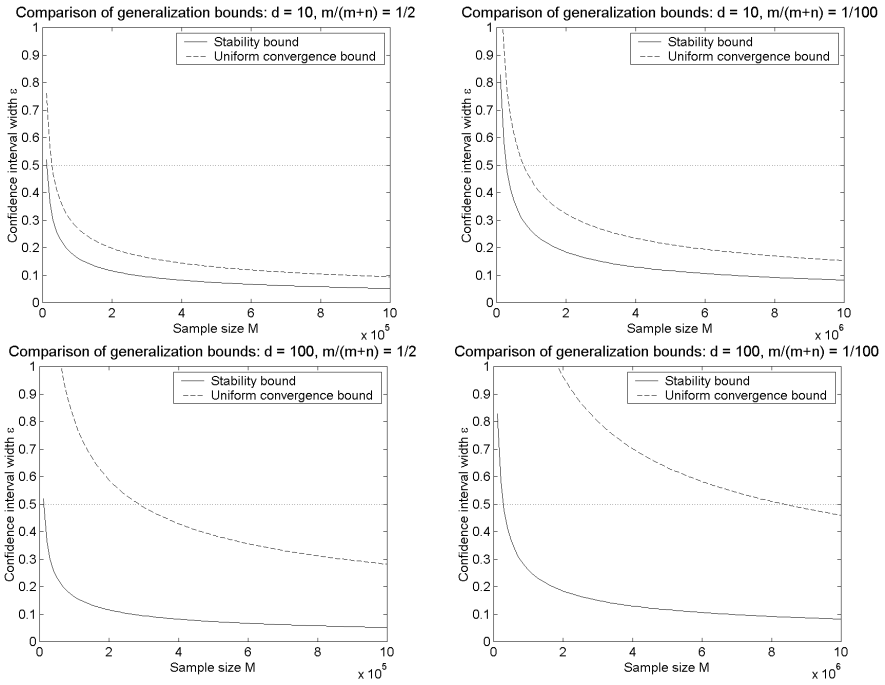

Fig. 1. A comparison of our stability bound with the uniform convergence bound of [6] for the kernel-based algorithm described in Section 5.2, with a linear kernel over the unit ball in $\mathbb{R}^d$. The plots are for $\delta = 0.01$, $\lambda = 1$, and show how the confidence interval size ϵ given by the two bounds varies with the sample size $M = m + n$, for various values of d and $m/(m + n)$.

that for classification, and indeed, ranking algorithms often resemble 'classification on pairs'. However, generalization bounds from classification *cannot* be applied directly to ranking, due to dependences among the instance pairs. Indeed, the bounds we have obtained for ranking suggest that the effective sample size in ranking is not only smaller than the number of positive-negative pairs mn, but is even smaller than the number of instances $M = m+n$; the dependences reduce the effective sample size to $\rho(1-\rho)M$, where $\rho = m/(m+n)$ is the 'positive skew' of the sample.

The notions of uniform stability studied in this paper correspond most closely to those studied by Bousquet and Elisseeff [7]. These notions are strict in that they require changes in a sample to have bounded effect uniformly over all samples and replacements. Kutin and Niyogi [8] have derived generalization bounds (for classification and regression algorithms) using a less strict notion of stability termed 'almost-everywhere' stability; this requires changes in a sample to have bounded effect only with high probability (over the draw of the sample and the replacement element). The notion of almost-everywhere stability leads to a distribution-dependent treatment as opposed to the distribution-free treatment obtained with uniform stability, and it would be particularly interesting to see if making distributional assumptions in ranking can mitigate the reduced sample size effect discussed above.

An open question concerns the analysis of other ranking algorithms using the algorithmic stability framework. It has been shown [18] that AdaBoost is stability-preserving, in the sense that stability of base classifiers implies stability of the final learned classifier. It would be interesting if a similar result could be shown for the bipartite RankBoost algorithm [5], which is based on the same principles of boosting as AdaBoost.

Finally, it is also an open question to analyze generalization properties of ranking algorithms in other settings of the ranking problem (*i.e.*, other than bipartite).

Acknowledgments

S. A. was supported in part through National Science Foundation (NSF) ITR grants IIS 00-85980 and IIS 00-85836. P. N. thanks the NSF for financial support.

References

1. Vapnik, V.N., Chervonenkis, A.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16** (1971) 264–280
2. Cohen, W.W., Schapire, R.E., Singer, Y.: Learning to order things. Journal of Artificial Intelligence Research **10** (1999) 243–270
3. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers (2000) 115–132
4. Crammer, K., Singer, Y.: Pranking with ranking. In: Advances in Neural Information Processing Systems 14, MIT Press (2002) 641–647
5. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research **4** (2003) 933–969
6. Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D.: Generalization bounds for the area under the ROC curve. Journal of Machine Learning Research **6** (2005) 393–425
7. Bousquet, O., Elisseeff, A.: Stability and generalization. Journal of Machine Learning Research **2** (2002) 499–526
8. Kutin, S., Niyogi, P.: Almost-everywhere algorithmic stability and generalization error. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. (2002)
9. Devroye, L., Wagner, T.: Distribution-free performance bounds for potential function rules. IEEE Transactions on Information Theory **25** (1979) 601–604
10. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In: Advances in Neural Information Processing Systems 16, MIT Press (2004)
11. Agarwal, S., Roth, D.: Learnability of bipartite ranking functions. In: Proceedings of the 18th Annual Conference on Learning Theory. (2005)
12. Kearns, M., Ron, D.: Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural Computation **11** (1999) 1427–1453
13. Poggio, T., Rifkin, R., Mukherjee, S., Niyogi, P.: General conditions for predictivity in learning theory. Nature **428** (2004) 419–422
14. McDiarmid, C.: On the method of bounded differences. In: Surveys in Combinatorics 1989, Cambridge University Press (1989) 148–188
15. Agarwal, S.: A Study of the Bipartite Ranking Problem in Machine Learning. PhD thesis, University of Illinois at Urbana-Champaign (2005)
16. Evgeniou, T., Pontil, M., Poggio, T.: Regularization networks and support vector machines. Advances in Computational Mathematics **13** (2000) 1–50
17. Rakotomamonjy, A.: SVMs and area under ROC curves. Technical report, PSI- INSA de Rouen (2004)
18. Kutin, S., Niyogi, P.: The interaction of stability and weakness in AdaBoost. Technical Report TR-2001-30, Computer Science Department, University of Chicago (2001)