Journal of Machine Learning Research 10 (2009) 441-474 Submitted 10/07; Revised 12/08; Published 2/09

Generalization Bounds for Ranking Algorithms via Algorithmic Stability*

Shivani Agarwal SHIVANI@MIT.EDU
Department of Electrical Engineering & Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Partha Niyogi NIYOGI@CS.UCHICAGO.EDU
Departments of Computer Science and Statistics
University of Chicago
Chicago, IL 60637, USA

Editor: Andre Elisseeff

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 much attention in machine learning. We study generalization properties of ranking algorithms using the notion of algorithmic stability; in particular, we derive generalization bounds for 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 generalization bounds based on uniform convergence, which in many cases cannot be applied to these algorithms. Our results generalize earlier results that were derived in the special setting of bipartite ranking (Agarwal and Niyogi, 2005) to a more general setting of the ranking problem that arises frequently in applications.

Keywords: ranking, generalization bounds, algorithmic stability

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 (1971), 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.

In recent years, a new learning problem, namely that of *ranking*, has gained attention in machine learning (Cohen et al., 1999; Herbrich et al., 2000; Crammer and Singer, 2002; Joachims, 2002; Freund et al., 2003; Agarwal et al., 2005; Rudin et al., 2005; Burges et al., 2005; Cossock and Zhang, 2006; Cortes et al., 2007; Clemencon et al., 2008). In ranking, one learns a real-valued function that assigns scores to instances, but the scores themselves do not matter; instead, what is

*. A preliminary version of this paper (which focused on the special setting of bipartite ranking) appeared in the Proceedings of the 18th Annual Conference on Learning Theory (COLT) in 2005.

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.

Although there have been several recent advances in developing algorithms for various settings of the ranking problem, the study of generalization properties of ranking algorithms has been largely limited to the special setting of bipartite ranking (Freund et al., 2003; Agarwal et al., 2005). In this paper, we study generalization properties of ranking algorithms in a more general setting of the ranking problem that arises frequently in applications. Our generalization bounds are derived using the notion of algorithmic stability; we show that a number of practical ranking algorithms satisfy the required stability conditions, and therefore can be analyzed using our bounds.

1.1 Previous Results

While ranking has been studied extensively in some form or another in fields as diverse as social choice theory (Arrow, 1970), statistics (Lehmann, 1975), and mathematical economics (Chiang and Wainwright, 2005), the study of ranking in machine learning is relatively new: the first paper on the subject appeared less than a decade ago (Cohen et al., 1999). Since then, however, the number of domains in which ranking has found applications has grown quickly, and as a result, ranking has gained considerable attention in machine learning and learning theory in recent years.

Some of the earlier work, by Herbrich et al. (2000) and by Crammer and Singer (2002), focused on the closely related but distinct problem of ordinal regression. Freund et al. (2003) gave one of the first learning algorithms for ranking, termed RankBoost, which was based on the principles of boosting. Since then there have been many other algorithmic developments: for example, Radlinski and Joachims (2005) have developed an algorithmic framework for ranking in information retrieval applications; Burges et al. (2005) have developed a neural network based algorithm for ranking; and Agarwal (2006) has developed an algorithmic framework for ranking in a graph-based transductive setting. More recently, there has been some interest in learning ranking functions that emphasize accuracy at the top of the ranked list; work by Rudin (2006), Cossock and Zhang (2006) and Clemencon and Vayatis (2007) falls in this category. There has also been interest in statistical analysis of ranking; in recent work, Clemencon et al. (2008) have studied statistical convergence properties of ranking algorithms—specifically, ranking algorithms based on empirical and convex risk minimization—using the theory of U-statistics.

In the paper that developed the RankBoost algorithm, Freund et al. (2003) also gave a basic generalization bound for the algorithm in the bipartite setting. Their bound was derived from uniform convergence results for the binary classification error, and was expressed in terms of the VC-dimension of a class of binary classification functions derived from the class of ranking functions searched by the algorithm. Agarwal et al. (2005) also gave a generalization bound for bipartite ranking algorithms based on uniform convergence; in this case, the uniform convergence result was derived directly for the bipartite ranking error, and the resulting generalization bound was expressed in terms of a new set of combinatorial parameters that measure directly the complexity of the class of ranking functions searched by a bipartite ranking algorithm. Agarwal and Niyogi (2005) used a different tool, namely that of algorithmic stability (Rogers and Wagner, 1978; Bousquet and Elisseeff, 2002), to obtain generalization bounds for bipartite ranking algorithms that have good stability properties. Unlike bounds based on uniform convergence, the stability-based bounds depend on

properties of the algorithm rather than the function class being searched, and can be applied also to algorithms that search function classes of unbounded complexity.

As can be noted from the above discussion, the question of generalization properties of ranking algorithms has so far been investigated mainly in the special setting of bipartite ranking. There have been limited studies of generalization properties in more general settings. For example, Rudin et al. (2005) derived a margin-based bound which is expressed in terms of covering numbers and relies ultimately on a uniform convergence result; this bound is derived for a non-bipartite setting of the ranking problem, but under the restrictive distributional assumption of a "truth" function. Cortes et al. (2007) consider a different setting of the ranking problem and derive stability-based generalization bounds for algorithms in this setting. However, they also implicitly assume a "truth" function. In addition, as we discuss later in the paper, the results of Cortes et al. as stated involve some strong assumptions about the function class searched by an algorithm. These assumptions rarely hold for practical ranking algorithms, which prevents the direct application of their results. We shall discuss in Section 6 how this can be remedied.

1.2 Our Results

We use the notion of algorithmic stability to study generalization properties of ranking algorithms in a more general setting of the ranking problem than has been considered previously, and that arises frequently in applications. The notion of algorithmic stability, first studied for learning algorithms by Rogers and Wagner (1978), has been used to obtain generalization bounds for classification and regression algorithms that satisfy certain stability conditions (Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002). Here we show that algorithmic stability can be useful also in analyzing generalization properties of ranking algorithms in the setting we consider; in particular, we derive generalization bounds for ranking algorithms that have 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. Our techniques are based on those of Bousquet and Elisseeff (2002); indeed, we show that the ranking error in our setting satisfies the same conditions that were used to establish the classification and regression bounds of Bousquet and Elisseeff (2002), and therefore essentially the same proof techniques can be used to analyze the ranking problem we consider. Our results generalize those of Agarwal and Niyogi (2005), which focused on bipartite ranking.

We describe the general ranking problem and the setting we consider in detail in Section 2, and define notions of stability for ranking algorithms in this setting 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. Section 6 provides comparisons with related work; we conclude with a discussion in Section 7.

2. The Ranking Problem

In the problem of ranking, one is given a finite number of examples of order relationships among instances in some instance space $\mathcal{X}$, and the goal is to learn from these examples a ranking or ordering over $\mathcal{X}$ that ranks accurately future instances. Examples of ranking problems arise in a variety of domains: in information retrieval, one wants to rank documents according to relevance to some query or topic; in user-preference modeling, one wants to rank books or movies according

to a user's likes and dislikes; in computational biology, one wants to rank genes according to their relevance to some disease.

In the most general setting of the ranking problem, the learner is given training examples in the form of ordered pairs of instances $(x,x') \in \mathcal{X} \times \mathcal{X}$ labeled with a ranking preference $r \in \mathbb{R}$, with the interpretation that x is to be ranked higher than (preferred over) x' if $r > 0$, and lower than x' if $r < 0$ ($r = 0$ indicates no ranking preference between the two instances); the penalty for mis-ordering such a pair is proportional to $|r|$. Given a finite number of such examples $((x_1,x'_1,r_1),\ldots,(x_m,x'_m,r_m))$, the goal is to learn a real-valued ranking function $f : \mathcal{X} \to \mathbb{R}$ that ranks accurately future instances; f is considered to rank an instance $x \in \mathcal{X}$ higher than an instance $x' \in \mathcal{X}$ if $f(x) > f(x')$, and lower than x' if $f(x) < f(x')$. Thus, assuming that ties are broken uniformly at random, the expected penalty (or loss) incurred by a ranking function f on a pair of instances (x,x') labeled by r can be written as

$$|r| \left(\mathbf{I}_{\{r(f(x)-f(x'))<0\}} + \frac{1}{2}\mathbf{I}_{\{f(x)=f(x')\}} \right),$$

where $\mathbf{I}_{\{\phi\}}$ is 1 if ϕ is true and 0 otherwise.

A particular setting of the ranking problem that has been investigated in some detail in recent years is the *bipartite* setting (Freund et al., 2003; Agarwal et al., 2005). In the bipartite ranking problem, 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. Formally, the learner is given a training sample (S^+,S^-) consisting of a sequence of 'positive' examples $S^+ = (x_1^+,\ldots,x_m^+)$ and a sequence of 'negative' examples $S^- = (x_1^-,\ldots,x_n^-)$, the x_i^+ and x_j^- being instances in some instance space $\mathcal{X}$, and the goal is to learn a real-valued ranking function $f : \mathcal{X} \to \mathbb{R}$ that ranks future positive instances higher than negative ones. The bipartite ranking problem is easily seen to be a special case of the general ranking problem described above, since a training sample $(S^+,S^-) \in \mathcal{X}^m \times \mathcal{X}^n$ in the bipartite setting can be viewed as consisting of mn examples of the form $(x_i^+,x_j^-,1)$, for $1 \le i \le m$, $1 \le j \le n$; in other words, mis-ranking any positive-negative pair of instances incurs a constant penalty of 1. Thus, assuming again that ties are broken uniformly at random, the expected penalty incurred by f on a positive-negative pair (x^+,x^-) is simply

$$\mathbf{I}_{\{f(x^+)-f(x^-)<0\}} + \frac{1}{2}\mathbf{I}_{\{f(x^+)=f(x^-)\}};$$

the penalty incurred on a pair of positive instances or a pair of negative instances is zero.

In this paper, we consider a more general setting: the learner is given examples of instances labeled by real numbers, and the goal is to learn a ranking in which instances labeled by larger numbers are ranked higher than instances labeled by smaller numbers. Such ranking problems arise frequently in practice: for example, in information retrieval, one is often given examples of documents with real-valued relevance scores for a particular topic or query; similarly, in computational biology, one often receives examples of molecular structures with real-valued biological activity scores with respect to a particular target.

Formally, the setting we consider can be described as follows. The learner is given a finite sequence of labeled training examples $S = ((x_1,y_1),\ldots,(x_m,y_m))$, where the x_i are instances in some instance space $\mathcal{X}$ and the y_i are real-valued labels in some bounded set $\mathcal{Y} \subseteq \mathbb{R}$ which we take without loss of generality to be $\mathcal{Y} = [0,M]$ for some $M > 0$, and the goal is to learn a real-valued function $f : \mathcal{X} \to \mathbb{R}$ that ranks future instances with larger labels higher than those with smaller

labels. The penalty for mis-ranking a pair of instances could again be taken to be constant for all pairs; here we consider the more general case where the penalty is larger for mis-ranking a pair of instances with a greater difference between their real-valued labels. In particular, in our setting, the penalty for mis-ranking a pair of instances is proportional to the absolute difference between their real-valued labels. Thus, assuming again that ties are broken uniformly at random, the expected penalty incurred by f on a pair of instances (x,x') with corresponding real-valued labels y and y' can be written as

$$|y-y'|\left(\mathbf{I}_{\{(y-y')(f(x)-f(x'))<0\}}+\frac{1}{2}\mathbf{I}_{\{f(x)=f(x')\}}\right).$$

This problem can also be seen to be a special case of the general ranking problem described above; a training sample $S=((x_1,y_1),\ldots,(x_m,y_m))\in(\mathcal{X}\times\mathcal{Y})^m$ in this setting can be viewed as consisting of $\binom{m}{2}$ examples of the form (x_i,x_j,y_i-y_j), for $1\leq i<j\leq m$.

In studying generalization properties of learning algorithms, one usually assumes that both training examples and future, unseen examples are generated according to some underlying random process. We shall assume in our setting that all examples (x,y) (both training examples and future, unseen examples) are drawn randomly and independently according to some (unknown) distribution $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$.[1] The quality of a ranking function $f:\mathcal{X}\to\mathbb{R}$ can then be measured by its *expected ranking error*, which we denote by $R(f)$ and define as follows:

$$R(f) \;=\; \mathbf{E}_{((X,Y),(X',Y'))\sim\mathcal{D}\times\mathcal{D}}\left[|Y-Y'|\left(\mathbf{I}_{\{(Y-Y')(f(X)-f(X'))<0\}}+\frac{1}{2}\mathbf{I}_{\{f(X)=f(X')\}}\right)\right]. \qquad (1)$$

Note that the expected error $R(f)$ is simply the expected mis-ranking penalty of f on a pair of examples drawn randomly and independently according to $\mathcal{D}$, assuming that ties are broken uniformly at random. In practice, since the distribution $\mathcal{D}$ is 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 $S=((x_1,y_1),\ldots,(x_m,y_m))\in(\mathcal{X}\times\mathcal{Y})^m$, which we denote by $\widehat{R}(f;S)$ and define as follows:

$$\widehat{R}(f;S) \;=\; \frac{1}{\binom{m}{2}}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}|y_i-y_j|\left(\mathbf{I}_{\{(y_i-y_j)(f(x_i)-f(x_j))<0\}}+\frac{1}{2}\mathbf{I}_{\{f(x_i)=f(x_j)\}}\right). \qquad (2)$$

This is simply the average mis-ranking penalty incurred by f on the $\binom{m}{2}$ pairs (x_i,x_j), where $1\leq i<j\leq m$, assuming that ties are broken uniformly at random.

A learning algorithm for the ranking problem described above takes as input a training sample $S\in\bigcup_{m=1}^{\infty}(\mathcal{X}\times\mathcal{Y})^m$ and returns as output a ranking function $f_S:\mathcal{X}\to\mathbb{R}$. 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) *Define a* ranking loss function *to be a function* $\ell:\mathbb{R}^{\mathcal{X}}\times(\mathcal{X}\times\mathcal{Y})\times(\mathcal{X}\times\mathcal{Y})\to\mathbb{R}^+\cup\{0\}$ *that assigns to each* $f:\mathcal{X}\to\mathbb{R}$ *and* $(x,y),(x',y')\in(\mathcal{X}\times\mathcal{Y})$ *a non-negative real number* $\ell(f,(x,y),(x',y'))$*, interpreted as the penalty or loss of* f *in its relative*

1. Cortes et al. (2007) consider a similar setting as ours; however, they assume that only instances $x\in\mathcal{X}$ are drawn randomly, and that labels $y\in\mathcal{Y}$ are then determined according to a "truth" function $f^*:\mathcal{X}\to\mathcal{Y}$.

ranking of x and x' given corresponding labels y and y'. We shall require that ℓ be symmetric with respect to (x,y) and (x',y'), that is, that $\ell(f,(x,y),(x',y')) = \ell(f,(x',y'),(x,y))$ for all $f,(x,y),(x',y')$.

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

$$R_\ell(f) \quad = \quad \mathbf{E}_{((X,Y),(X',Y'))\sim\mathcal{D}\times\mathcal{D}}\left[\ell(f,(X,Y),(X',Y'))\right] .$$

Definition 3 (Empirical ℓ-error) *Let $f : \mathcal{X} \to \mathbb{R}$ be a ranking function on $\mathcal{X}$, and let $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$. Let $\ell : \mathbb{R}^{\mathcal{X}} \times (\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}^+ \cup \{0\}$ be a ranking loss function. Define the* empirical ℓ-error *of f with respect to S, denoted by $\widehat{R}_\ell(f;S)$, as*

$$\widehat{R}_\ell(f;S) \quad = \quad \frac{1}{\binom{m}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \ell(f,(x_i,y_i),(x_j,y_j)) .$$

As mentioned above, one choice for a ranking loss function could be a 0-1 loss that simply assigns a constant penalty of 1 to any mis-ranked pair:

$$\ell_{0\text{-}1}(f,(x,y),(x',y')) \quad = \quad \mathbf{I}_{\{(y-y')(f(x)-f(x'))<0\}} + \frac{1}{2}\mathbf{I}_{\{f(x)=f(x')\}} .$$

The (empirical) $\ell_{0\text{-}1}$-error then simply counts the fraction of mis-ranked pairs, which corresponds to the well-known Kendall τ measure. However, the 0-1 loss effectively uses only the sign of the difference $y-y'$ between labels, and ignores the magnitude of this difference; the loss function we use, which we term the *discrete ranking loss* and denote as

$$\ell_{\text{disc}}(f,(x,y),(x',y')) \quad = \quad |y-y'| \left(\mathbf{I}_{\{(y-y')(f(x)-f(x'))<0\}} + \frac{1}{2}\mathbf{I}_{\{f(x)=f(x')\}} \right) , \tag{3}$$

takes the magnitude of this difference into account. Comparing with Eqs. (1-2) above, we see that the expected and empirical ranking errors we have defined are simply the corresponding ℓ_{disc}-errors:

$$R(f) \equiv R_{\ell_{\text{disc}}}(f); \;\; \widehat{R}(f;S) \equiv \widehat{R}_{\ell_{\text{disc}}}(f;S) .$$

While our focus will be on bounding the expected (ℓ_{disc}-)ranking error of a learned ranking function, our results can be used also to bound the expected $\ell_{0\text{-}1}$-error.

Several other ranking loss functions will be useful in our study; these will be introduced in the following sections as needed.

3. Stability of 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 in our setting is a training sample of the form $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ for some $m \in \mathbb{N}$; we consider changes to such a sample that consist of replacing a single example in the sequence with a new example. For $1 \le i \le m$ and $(x_i',y_i') \in (\mathcal{X} \times \mathcal{Y})$, we use S^i to denote the sequence obtained from S by replacing (x_i,y_i) with (x_i',y_i').

Several different notions of stability have been used in the study of classification and regression algorithms (Rogers and Wagner, 1978; Devroye and Wagner, 1979; Kearns and Ron, 1999; Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002; Poggio et al., 2004). The notions of stability that we define below for ranking algorithms in our setting are based on those defined earlier for bipartite ranking algorithms (Agarwal and Niyogi, 2005) and are most closely related to the notions of stability used by Bousquet and Elisseeff (2002).

Definition 4 (Uniform loss stability) *Let $\mathcal{A}$ be a ranking algorithm whose output on a training sample S we denote by f_S, and let ℓ be a ranking loss function. Let $\beta : \mathbb{N} \to \mathbb{R}$. We say that $\mathcal{A}$ has* uniform loss stability β *with respect to ℓ if for all $m \in \mathbb{N}$, $S \in (\mathcal{X} \times \mathcal{Y})^m$, $1 \leq i \leq m$ and $(x_i', y_i') \in (\mathcal{X} \times \mathcal{Y})$, we have for all $(x,y),(x',y') \in (\mathcal{X} \times \mathcal{Y})$,*

$$\left|\ell(f_S,(x,y),(x',y')) - \ell(f_{S^i},(x,y),(x',y'))\right| \quad \leq \quad \beta(m)\,.$$

Definition 5 (Uniform score stability) *Let $\mathcal{A}$ be a ranking algorithm whose output on a training sample S we denote by f_S. Let $\nu : \mathbb{N} \to \mathbb{R}$. We say that $\mathcal{A}$ has* uniform score stability ν *if for all $m \in \mathbb{N}$, $S \in (\mathcal{X} \times \mathcal{Y})^m$, $1 \leq i \leq m$ and $(x_i', y_i') \in (\mathcal{X} \times \mathcal{Y})$, we have for all $x \in \mathcal{X}$,*

$$\left|f_S(x) - f_{S^i}(x)\right| \quad \leq \quad \nu(m)\,.$$

If a ranking algorithm has uniform loss stability β with respect to ℓ, then changing an input training sample of size m by a single example leads to a difference of at most $\beta(m)$ in the ℓ-loss incurred by the output ranking function on any pair of examples $(x,y),(x',y')$. Therefore, a smaller value of $\beta(m)$ corresponds to greater loss stability. Similarly, if a ranking algorithm has uniform score stability ν, then changing an input training sample of size m by a single example leads to a difference of at most $\nu(m)$ in the score assigned by the output ranking function to any instance x. A smaller value of $\nu(m)$ therefore corresponds to greater score stability.

The term 'uniform' in the above definitions refers to the fact that the bounds on the difference in loss or score are required to hold uniformly for all training samples S (and all single-example changes to them) and for all examples $(x,y),(x',y')$ or instances x. This is arguably a strong requirement; one can define weaker notions of stability, analogous to the hypothesis stability considered by Devroye and Wagner (1979), Kearns and Ron (1999), and Bousquet and Elisseeff (2002), or the almost-everywhere stability considered by Kutin and Niyogi (2002), which would require the bounds to hold only in expectation or with high probability (and would therefore depend on the distribution $\mathcal{D}$ governing the data). However, we do not consider such weaker notions of stability in this paper; we shall show later (Section 5) that several practical ranking algorithms in fact exhibit good uniform stability properties.

4. Generalization Bounds for Stable Ranking Algorithms

In this section we give generalization bounds for ranking algorithms that exhibit good (uniform) stability properties. The methods we use are based on those of Bousquet and Elisseeff (2002), who derived such bounds for classification and regression algorithms. Our main result is the following, 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.

Theorem 6 *Let $\mathcal{A}$ be a symmetric ranking algorithm*[2] *whose output on a training sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ we denote by f_S, and let ℓ be a bounded ranking loss function such that $0 \leq \ell(f,(x,y),(x',y')) \leq B$ for all $f : \mathcal{X} \to \mathbb{R}$ and $(x,y),(x',y') \in (\mathcal{X} \times \mathcal{Y})$. Let $\beta : \mathbb{N} \to \mathbb{R}$ be such that $\mathcal{A}$ has uniform loss stability β with respect to ℓ. Then for any $0 < \delta < 1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$),*

$$R_\ell(f_S) \quad < \quad \widehat{R}_\ell(f_S;S) + 2\beta(m) + (m\beta(m)+B)\sqrt{\frac{2\ln(1/\delta)}{m}}.$$

The proof follows the proof of a similar result for classification and regression algorithms by Bousquet and Elisseeff (2002). In particular, the random variable $R_\ell(f_S) - \widehat{R}_\ell(f_S;S)$, representing the difference between the expected and empirical ℓ-errors of the ranking function f_S learned from the random sample S, is shown to satisfy the conditions of McDiarmid's inequality (McDiarmid, 1989), as a result of which the deviation of this difference from its expected value $\mathbf{E}_{S\sim\mathcal{D}^m}[R_\ell(f_S) - \widehat{R}_\ell(f_S;S)]$ can be bounded with high probability; a bound on this expected value then allows the above result to be established. Details of the proof are provided in Appendix A.

A few remarks on the significance of the above result are in order, especially in relation to the results of Bousquet and Elisseeff (2002). As can be seen from the definitions given in the preceding two sections, 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. Indeed, this is the reason that in deriving uniform convergence bounds for the ranking error, the standard Hoeffding inequality used to obtain such bounds in classification and regression can no longer be applied (Agarwal et al., 2005). It may initially come as a bit of a surprise, therefore, that the proof methods of Bousquet and Elisseeff (2002) carry through for ranking without siginificant change. The reason for this is that they rely on the more general inequality of McDiarmid which, in Bousquet and Elisseeff's work, is used to capture the effect of stability, but which is also powerful enough to capture the structure of the ranking error; indeed, the uniform convergence bound for the (bipartite) ranking error derived by Agarwal et al. (2005) also made use of McDiarmid's inequality. Thus, in general, any learning problem in which the empirical performance measure fits the conditions of McDiarmid's inequality should be amenable to a stability analysis similar to Bousquet and Elisseeff's, provided of course that appropriate notions of stability are defined.

Theorem 6 gives meaningful bounds when $\beta(m) = o(1/\sqrt{m})$. 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 discrete ranking loss ℓ_{disc} defined in Eq. (3). (To see this, note that unless an algorithm picks essentially the same ranking function that orders all instances the same way for all training samples of a given size m, in which case the algorithm trivially has uniform loss stability $\beta(m) = 0$ with respect to ℓ_{disc}, there must be some $S \in (\mathcal{X} \times \mathcal{Y})^m$, $1 \leq i \leq m$, $(x_i',y_i') \in (\mathcal{X} \times \mathcal{Y})$ and some $x,x' \in \mathcal{X}$ such that f_S and f_{S^i} order x,x' differently. In this case, for $y = 0$ and $y' = M$, we get $|\ell_{\text{disc}}(f_S,(x,y),(x',y')) - \ell_{\text{disc}}(f_{S^i},(x,y),(x',y'))| = M$, giving loss stability $\beta(m) = M$ with respect to ℓ_{disc}.) However, for any ranking loss ℓ that satisfies $\ell_{\text{disc}} \leq \ell$,

2. A symmetric ranking algorithm is one whose output is independent of the order of elements in the training sequence S.

Theorem 6 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 $R \leq R_\ell$, these bounds apply also to the expected ranking error. We consider below a specific ranking loss that satisfies this condition, and with respect to which we will be able to show good loss stability of certain ranking algorithms; other ranking losses which can also be used in this manner will be discussed in later sections.

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

$$\ell_\gamma(f,(x,y),(x',y')) = \begin{cases} |y-y'| & \text{if } \frac{(f(x)-f(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} \leq 0 \\ |y-y'| - \frac{(f(x)-f(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} & \text{if } 0 < \frac{(f(x)-f(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} < |y-y'| \\ 0 & \text{if } \frac{(f(x)-f(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} \geq |y-y'|, \end{cases}$$

where for $u \in \mathbb{R}$,

$$\mathrm{sgn}(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u = 0 \\ -1 & \text{if } u < 0. \end{cases}$$

Clearly, for all $\gamma > 0$, we have $\ell_{\mathrm{disc}} \leq \ell_\gamma$. Therefore, for any ranking algorithm that has good uniform loss stability with respect to ℓ_γ for some $\gamma > 0$, Theorem 6 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 7 *Let $\mathcal{A}$ be a ranking algorithm whose output on a training sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ we denote by f_S. Let $\nu : \mathbb{N} \to \mathbb{R}$ be such that $\mathcal{A}$ has uniform score stability ν. Then for every $\gamma > 0$, $\mathcal{A}$ has uniform loss stability β_γ with respect to the γ ranking loss ℓ_γ, where for all $m \in \mathbb{N}$,*

$$\beta_\gamma(m) = \frac{2\nu(m)}{\gamma}.$$

Proof Let $m \in \mathbb{N}$, $S \in (\mathcal{X} \times \mathcal{Y})^m$, $1 \leq i \leq m$ and $(x_i', y_i') \in (\mathcal{X} \times \mathcal{Y})$. Let $(x,y),(x',y') \in (\mathcal{X} \times \mathcal{Y})$. We want to show

$$\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right| \leq \frac{2\nu(m)}{\gamma}.$$

Now, if $\ell_\gamma(f_S,(x,y),(x',y')) = \ell_\gamma(f_{S^i},(x,y),(x',y'))$, then trivially

$$\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right| = 0 \leq \frac{2\nu(m)}{\gamma},$$

and there is nothing to prove. Therefore assume $\ell_\gamma(f_S,(x,y),(x',y')) \neq \ell_\gamma(f_{S^i},(x,y),(x',y'))$. Without loss of generality, let $\ell_\gamma(f_S,(x,y),(x',y')) > \ell_\gamma(f_{S^i},(x,y),(x',y'))$. There are four possibilities:

(i) $\frac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} \leq 0$ and $0 < \frac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} < |y-y'|$.

In this case, we have

$$\begin{aligned}
&\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right| \\
&= |y-y'| - \left(|y-y'| - \tfrac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) \\
&\leq \left(|y-y'| - \tfrac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) - \left(|y-y'| - \tfrac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) \\
&\leq \frac{1}{\gamma}\left(|f_S(x) - f_{S^i}(x)| + \left|f_S(x') - f_{S^i}(x')\right|\right) \\
&\leq \frac{2\nu(m)}{\gamma}.
\end{aligned}$$

(ii) $0 < \frac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} < |y-y'|$ and $\frac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} \geq |y-y'|$.

In this case, we have

$$\begin{aligned}
&\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right| \\
&= \left(|y-y'| - \tfrac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) - 0 \\
&\leq \left(|y-y'| - \tfrac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) - \left(|y-y'| - \tfrac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) \\
&\leq \frac{1}{\gamma}\left(|f_S(x) - f_{S^i}(x)| + \left|f_S(x') - f_{S^i}(x')\right|\right) \\
&\leq \frac{2\nu(m)}{\gamma}.
\end{aligned}$$

(iii) $\frac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} \leq 0$ and $\frac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} \geq |y-y'|$.

In this case, we have

$$\begin{aligned}
&\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right| \\
&= |y-y'| - 0 \\
&\leq \left(|y-y'| - \tfrac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) - \left(|y-y'| - \tfrac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) \\
&\leq \frac{1}{\gamma}\left(|f_S(x) - f_{S^i}(x)| + \left|f_S(x') - f_{S^i}(x')\right|\right) \\
&\leq \frac{2\nu(m)}{\gamma}.
\end{aligned}$$

(iv) $0 < \frac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} < |y-y'|$ and $0 < \frac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma} < |y-y'|$.

In this case, we have

$$\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right|$$

$$
\begin{aligned}
&= \left(|y-y'| - \tfrac{(f_S(x)-f_S(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) - \left(|y-y'| - \tfrac{(f_{S^i}(x)-f_{S^i}(x'))\cdot \mathrm{sgn}(y-y')}{\gamma}\right) \\
&\leq \frac{1}{\gamma}\left(|f_S(x) - f_{S^i}(x)| + \left|f_S(x') - f_{S^i}(x')\right|\right) \\
&\leq \frac{2\nu(m)}{\gamma}.
\end{aligned}
$$

Thus in all cases, $\left|\ell_\gamma(f_S,(x,y),(x',y')) - \ell_\gamma(f_{S^i},(x,y),(x',y'))\right| \leq \frac{2\nu(m)}{\gamma}$. ■

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 8 *Let $\mathcal{A}$ be a symmetric ranking algorithm whose output on a training sample $S \in (\mathcal{X}\times\mathcal{Y})^m$ we denote by f_S. Let $\nu : \mathbb{N}\to\mathbb{R}$ be such that $\mathcal{A}$ has uniform score stability ν, and let $\gamma > 0$. Then for any $0 < \delta < 1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$),*

$$
R(f_S) \quad < \quad \widehat{R}_{\ell_\gamma}(f_S;S) + \frac{4\nu(m)}{\gamma} + \left(\frac{2m\nu(m)}{\gamma} + M\right)\sqrt{\frac{2\ln(1/\delta)}{m}}.
$$

Proof The result follows by applying Theorem 6 to $\mathcal{A}$ with the ranking loss ℓ_γ (using Lemma 7), which satisfies $0 \leq \ell_\gamma \leq M$, and from the fact that $R \leq R_{\ell_\gamma}$. ■

5. Stable Ranking Algorithms

In this section we show (uniform) 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 two such ranking algorithms, and apply the results of Section 4 to obtain generalization bounds for these algorithms. Again, the methods we use to show stability of these algorithms are based on those of Bousquet and Elisseeff (2002), who showed similar results for classification and regression algorithms. We also use these stability results to give a consistency theorem for kernel-based ranking algorithms (Section 5.2.3).

5.1 General Regularizers

Given a ranking loss function ℓ, 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 $S \in (\mathcal{X}\times\mathcal{Y})^m$), with regularization parameter $\lambda > 0$:

$$
\widehat{R}^\lambda_\ell(f;S) \quad = \quad \widehat{R}_\ell(f;S) + \lambda N(f).
$$

We consider ranking algorithms that minimize such a regularized objective function, that is, ranking algorithms that, given a training sample S, output a ranking function $f_S \in \mathcal{F}$ that satisfies

$$f_S \;=\; \arg\min_{f\in\mathcal{F}} \widehat{R}^{\lambda}_{\ell}(f;S)\,, \tag{4}$$

for some fixed choice of ranking loss ℓ, function class $\mathcal{F}$, regularizer N, and regularization parameter λ. It is worth noting that the parameter λ often depends on the sample size m; we use λ here rather than λ_m for notational simplicity. We give below a general result that will be useful for showing stability of such regularization-based algorithms.

Definition 9 (σ-admissibility) *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Let ℓ be a ranking loss and 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 $(x,y),(x',y') \in (\mathcal{X}\times\mathcal{Y})$, we have*

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

Lemma 10 *Let $\mathcal{F}$ be a convex class of real-valued functions on $\mathcal{X}$. Let ℓ be a ranking loss such that $\ell(f,(x,y),(x',y'))$ is convex in f, 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 $S \in (\mathcal{X}\times\mathcal{Y})^m$, the regularized empirical ℓ-error $\widehat{R}^{\lambda}_{\ell}(f;S)$ has a minimum (not necessarily unique) in $\mathcal{F}$. Let $\mathcal{A}$ be a ranking algorithm defined by Eq. (4), and let $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X}\times\mathcal{Y})^m$, $1 \leq i \leq m$, and $(x_i',y_i') \in (\mathcal{X}\times\mathcal{Y})$. For brevity, denote $f \equiv f_S$, $f^i \equiv f_{S^i}$, and let $\Delta f = f^i - f$. Then we have that for any $t \in [0,1]$,*

$$\begin{aligned} N(f) - N(f+t\Delta f) + N(f^i) - N(f^i - t\Delta f) & \\ \leq\; \frac{t\sigma}{\lambda\binom{m}{2}} \sum_{j\neq i} \Big(|\Delta f(x_i)| + 2|\Delta f(x_j)| + |\Delta f(x_i')|\Big)\,. & \end{aligned}$$

The proof follows that of a similar result for regularization-based algorithms for classification and regression (Bousquet and Elisseeff, 2002); it is based crucially on the convexity of the ranking loss ℓ. Details are provided in Appendix B.

As we shall see below, the above result can be used to establish stability of certain regularization-based ranking algorithms.

5.2 Regularization in Hilbert Spaces

Let $\mathcal{F}$ be a reproducing kernel Hilbert space (RKHS) of real-valued functions on $\mathcal{X}$, with kernel $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Then the reproducing property of $\mathcal{F}$ gives that for all $f \in \mathcal{F}$ and all $x \in \mathcal{X}$,

$$|f(x)| \;=\; |\langle f, K_x\rangle_K|\,, \tag{5}$$

where $K_x : \mathcal{X} \to \mathbb{R}$ is defined as

$$K_x(x') \;=\; K(x,x')\,,$$

and $\langle\cdot,\cdot\rangle_K$ denotes the RKHS inner product in $\mathcal{F}$. In particular, applying the Cauchy-Schwartz inequality to Eq. (5) then gives that for all $f \in \mathcal{F}$ and all $x \in X$,

$$\begin{aligned} |f(x)| &\leq \|f\|_K\, \|K_x\|_K \\ &= \|f\|_K \sqrt{K(x,x)}\,, \end{aligned} \tag{6}$$

where $\|\cdot\|_K$ denotes the RKHS norm in $\mathcal{F}$.

Further details about RKHSs can be found, for example, in expositions by Haussler (1999) and Evgeniou et al. (2000). We shall consider ranking algorithms that perform regularization in the RKHS $\mathcal{F}$ using the squared norm in $\mathcal{F}$ as a regularizer. Specifically, let $N : \mathcal{F} \to \mathbb{R}^+ \cup \{0\}$ be the regularizer defined by

$$N(f) \quad = \quad \|f\|_K^2 .$$

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 11 *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 $\ell(f,(x,y),(x',y'))$ is convex in f and ℓ is σ-admissible with respect to $\mathcal{F}$. Let $\lambda > 0$, and let $\mathcal{A}$ be a ranking algorithm that, given a training sample S, outputs a ranking function $f_S \in \mathcal{F}$ that satisfies $f_S = \arg\min_{f \in \mathcal{F}} \left\{ \widehat{R}_\ell(f;S) + \lambda \|f\|_K^2 \right\}$. Then $\mathcal{A}$ has uniform score stability ν, where for all $m \in \mathbb{N}$,*

$$\nu(m) = \frac{8\sigma\kappa^2}{\lambda m} .$$

Proof Let $m \in \mathbb{N}$, $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, $1 \leq i \leq m$, and $(x_i',y_i') \in (\mathcal{X} \times \mathcal{Y})$. Applying Lemma 10 with $t = 1/2$, we get (using the notation of Lemma 10) that

$$\begin{aligned} &\|f\|_K^2 - \|f + \tfrac{1}{2}\Delta f\|_K^2 + \|f^i\|_K^2 - \|f^i - \tfrac{1}{2}\Delta f\|_K^2 \\ &\qquad \leq \quad \frac{\sigma}{\lambda m(m-1)} \sum_{j \neq i} \Big(|\Delta f(x_i)| + 2|\Delta f(x_j)| + |\Delta f(x_i')| \Big) . \end{aligned} \tag{7}$$

Note that since $\mathcal{F}$ is a vector space, $\Delta f \in \mathcal{F}$, $(f + \frac{1}{2}\Delta f) \in \mathcal{F}$, and $(f^i - \frac{1}{2}\Delta f) \in \mathcal{F}$, so that $\|f + \frac{1}{2}\Delta f\|_K$ and $\|f^i - \frac{1}{2}\Delta f\|_K$ are well-defined. Now, we have

$$\begin{aligned} &\|f\|_K^2 - \|f + \tfrac{1}{2}\Delta f\|_K^2 + \|f^i\|_K^2 - \|f^i - \tfrac{1}{2}\Delta f\|_K^2 \\ &\qquad = \quad \|f\|_K^2 + \|f^i\|_K^2 - \frac{1}{2}\|f + f^i\|_K^2 \\ &\qquad = \quad \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\|f^i\|_K^2 - \langle f, f^i \rangle_K \\ &\qquad = \quad \frac{1}{2}\|\Delta f\|_K^2 . \end{aligned}$$

Combined with Eq. (7), this gives

$$\frac{1}{2}\|\Delta f\|_K^2 \quad \leq \quad \frac{\sigma}{\lambda m(m-1)} \sum_{j \neq i} \Big(|\Delta f(x_i)| + 2|\Delta f(x_j)| + |\Delta f(x_i')| \Big) .$$

Since (as noted above) $\Delta f \in \mathcal{F}$, by Eq. (6), we thus get that

$$\begin{aligned} \frac{1}{2}\|\Delta f\|_K^2 \quad &\leq \quad \frac{\sigma}{\lambda m(m-1)} \|\Delta f\|_K \sum_{j \neq i} \Big(\sqrt{K(x_i,x_i)} + 2\sqrt{K(x_j,x_j)} + \sqrt{K(x_i',x_i')} \Big) \\ &\leq \quad \frac{4\sigma\kappa}{\lambda m} \|\Delta f\|_K , \end{aligned}$$

which gives

$$\|\Delta f\|_K \;\leq\; \frac{8\sigma\kappa}{\lambda m}. \tag{8}$$

Thus, by Eqs. (6) and (8), we have for all $x \in \mathcal{X}$,

$$|f_S(x) - f_{S^i}(x)| = |\Delta f(x)| \;\leq\; \frac{8\sigma\kappa^2}{\lambda m}.$$

The result follows. ■

This gives the following generalization bound for kernel-based ranking algorithms:

Corollary 12 *Under the conditions of Theorem 11, we have that for any $0 < \delta < 1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$), the expected ranking error of the ranking function f_S learned by $\mathcal{A}$ is bounded by*

$$R(f_S) \;<\; \widehat{R}_{\ell_1}(f_S;S) + \frac{32\sigma\kappa^2}{\lambda m} + \left(\frac{16\sigma\kappa^2}{\lambda} + M\right)\sqrt{\frac{2\ln(1/\delta)}{m}}.$$

Proof The result follows from Theorem 11, by applying Theorem 8 with $\gamma = 1$. ■

Under the conditions of the above results, a kernel-based ranking algorithm minimizing a regularized empirical ℓ-error also has good uniform loss stability with respect to ℓ; this follows from the following simple lemma:[3]

Lemma 13 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\mathcal{A}$ be a ranking algorithm that, given a training sample S, outputs a ranking function $f_S \in \mathcal{F}$. If $\mathcal{A}$ has uniform score stability ν and ℓ is a ranking loss that is σ-admissible with respect to $\mathcal{F}$, then $\mathcal{A}$ has uniform loss stability β with respect to ℓ, where for all $m \in \mathbb{N}$,*

$$\beta(m) \;=\; 2\sigma\nu(m).$$

Proof Let $m \in \mathbb{N}$, $S \in (\mathcal{X}\times\mathcal{Y})^m$, $1 \leq i \leq m$ and $(x_i', y_i') \in (\mathcal{X}\times\mathcal{Y})$. Let $(x,y),(x',y') \in (\mathcal{X}\times\mathcal{Y})$. Then we have

$$\begin{aligned}\left|\ell(f_S,(x,y),(x',y')) - \ell(f_{S^i},(x,y),(x',y'))\right| &\leq \sigma\Big(|f_S(x) - f_{S^i}(x)| + \left|f_S(x') - f_{S^i}(x')\right|\Big)\\ &\leq 2\sigma\nu(m),\end{aligned}$$

where the first inequality follows from σ-admissibility and the second from uniform score stability. ■

3. We note that the proof of Lemma 7 in Section 4 really amounts to showing that ℓ_γ is $\frac{1}{\gamma}$-admissible with respect to the set of all ranking functions on $\mathcal{X}$; the result then follows by the observation in Lemma 13.

Corollary 14 *Under the conditions of Theorem 11, $\mathcal{A}$ has uniform loss stability β with respect to ℓ, where for all $m \in \mathbb{N}$,*

$$\beta(m) = \frac{16\sigma^2\kappa^2}{\lambda m}.$$

Proof Immediate from Theorem 11 and Lemma 13. ■

This leads to the following additional bound for kernel-based ranking algorithms that minimize a regularized empirical ℓ-error where ℓ is bounded and satisfies $\ell_{\text{disc}} \leq \ell$:

Corollary 15 *Under the conditions of Theorem 11, if in addition the ranking loss ℓ satisfies $\ell_{\text{disc}}(f,(x,y),(x',y')) \leq \ell(f,(x,y),(x',y')) \leq B$ for all $f : \mathcal{X} \to \mathbb{R}$ and $(x,y),(x',y') \in (\mathcal{X} \times \mathcal{Y})$, then we have that for any $0 < \delta < 1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$), the expected ranking error of the ranking function f_S learned by $\mathcal{A}$ is bounded by*

$$R(f_S) \quad < \quad \widehat{R}_\ell(f_S;S) + \frac{32\sigma^2\kappa^2}{\lambda m} + \left(\frac{16\sigma^2\kappa^2}{\lambda} + B\right)\sqrt{\frac{2\ln(1/\delta)}{m}}.$$

Proof The result follows from Corollary 14, by applying Theorem 6 and the fact that $R \leq R_\ell$. ■

The results of both Corollary 12 and Corollary 15 show that a larger regularization parameter λ leads to better stability and, therefore, a tighter confidence interval in the resulting generalization bound. In particular, when σ is independent of λ, one must have $\lambda \gg \frac{1}{\sqrt{m}}$ in order for either of these bounds to be meaningful. In practice, the ranking losses ℓ minimized by kernel-based ranking algorithms tend to be larger than the loss ℓ_1, and therefore Corollary 12 tends to provide tighter bounds on the expected ranking error than Corollary 15. Below we look at two specific algorithms that minimize two different (regularized) ranking losses.

5.2.1 HINGE RANKING LOSS

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

$$\ell_{\text{h}}(f,(x,y),(x',y')) \quad = \quad \Big(|y-y'| - (f(x)-f(x'))\cdot \text{sgn}(y-y')\Big)_+,$$

where for $a \in \mathbb{R}$,

$$a_+ \quad = \quad \begin{cases} a & \text{if } a > 0 \\ 0 & \text{otherwise.} \end{cases}$$

We consider a ranking algorithm $\mathcal{A}_{\text{h}}$ that minimizes the regularized ℓ_h-error in an RKHS $\mathcal{F}$. Specifically, let $\mathcal{A}_{\text{h}}$ be a ranking algorithm which, given a training sample $S \in (\mathcal{X} \times \mathcal{Y})^m$, outputs a ranking function $f_S \in \mathcal{F}$ that satisfies (for some fixed $\lambda > 0$)

$$f_S \quad = \quad \arg\min_{f \in \mathcal{F}} \left\{\widehat{R}_{\ell_{\text{h}}}(f;S) + \lambda \|f\|_K^2\right\}.$$

We note that this algorithm has an equivalent quadratic programming formulation similar to SVMs in the case of classification. In particular, the problem of minimizing $\widehat{R}^{\lambda}_{\ell_h}(f;S)$ is equivalent to that of minimizing

$$\frac{1}{2}\|f\|_K^2 + C\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\xi_{ij}$$

subject to

$$\begin{aligned}\xi_{ij} &\geq |y_i - y_j| - (f(x_i) - f(x_j)\cdot \mathrm{sgn}(y_i - y_j))\\ \xi_{ij} &\geq 0\end{aligned}$$

for all $1 \leq i < j \leq m$, where $C = 1/(\lambda m(m-1))$. The dual formulation of this problem obtained by introducing Lagrange multipliers leads to a quadratic program similar to that obtained for SVMs; for example, see the related algorithms described by Herbrich et al. (1998, 2000), Joachims (2002), Rakotomamonjy (2004), and Agarwal (2006).

It can be verified that $\ell_h(f,(x,y),(x',y'))$ is convex in f, and that ℓ_h is 1-admissible with respect to $\mathcal{F}$ (the proof of 1-admissibility is similar to the proof of Lemma 7 which, as noted earlier in Footnote 3, effectively shows $\frac{1}{\gamma}$-admissibility of the γ ranking loss ℓ_γ). Thus, if $K(x,x) \leq \kappa^2$ for all $x \in \mathcal{X}$, then from Corollary 12 we get that for any $0 < \delta < 1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$), the expected ranking error of the ranking function f_S learned by the above algorithm $\mathcal{A}_h$ is bounded by

$$R(f_S) \quad < \quad \widehat{R}_{\ell_1}(f_S;S) + \frac{32\kappa^2}{\lambda m} + \left(\frac{16\kappa^2}{\lambda} + M\right)\sqrt{\frac{2\ln(1/\delta)}{m}}.$$

5.2.2 LEAST SQUARES RANKING LOSS

Consider now the following *least squares ranking loss*:

$$\ell_{sq}(f,(x,y),(x',y')) \quad = \quad \Big(|y-y'| - \mathrm{sgn}(y-y')\cdot(f(x)-f(x'))\Big)^2. \tag{9}$$

Let $\mathcal{A}_{sq}$ be a ranking algorithm that minimizes the regularized ℓ_{sq}-error in an RKHS $\mathcal{F}$, that is, given a training sample $S \in (\mathcal{X}\times\mathcal{Y})^m$, the algorithm $\mathcal{A}_{sq}$ outputs a ranking function $f_S \in \mathcal{F}$ that satisfies (for some fixed $\lambda > 0$)

$$f_S \quad = \quad \arg\min_{f\in\mathcal{F}} \left\{\widehat{R}_{\ell_{sq}}(f;S) + \lambda\|f\|_K^2\right\}.$$

It can be verified that $\ell_{sq}(f,(x,y),(x',y'))$ is convex in f. Now, we claim that the effective search space of $\mathcal{A}_{sq}$ on training samples of size m is actually $\mathcal{F}_\lambda = \left\{ f \in \mathcal{F} \;\middle|\; \|f\|_K^2 \leq \frac{M^2}{\lambda}\right\}$, that is, that for any training sample $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X}\times\mathcal{Y})^m$, the ranking function f_S returned by $\mathcal{A}_{sq}$ satisfies $\|f_S\|_K^2 \leq \frac{M^2}{\lambda}$. To see this, note that for the zero function $f_0 \in \mathcal{F}$ which assigns $f_0(x) = 0$ for all $x \in \mathcal{X}$, we have

$$\begin{aligned}\widehat{R}_{\ell_{sq}}(f_0;S) + \lambda\|f_0\|_K^2 &= \frac{1}{\binom{m}{2}}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}(|y_i-y_j|-0)^2 + \lambda\cdot 0\\ &\leq M^2.\end{aligned}$$

Therefore, by definition of f_S,

$$\widehat{R}_{\ell_{\mathrm{sq}}}(f_S;S)+\lambda\|f_S\|_K^2 \;\leq\; M^2\,.$$

Since $\widehat{R}_{\ell_{\mathrm{sq}}}(f_S;S)\geq 0$, this implies

$$\|f_S\|_K^2 \;\leq\; \frac{M^2}{\lambda}\,.$$

It is easily verified that the effective search space $\mathcal{F}_\lambda$ is a convex set. The following lemma shows that if $K(x,x)\leq\kappa^2$ for all $x\in\mathcal{X}$, then ℓ_{sq} is $(2M+\frac{4\kappa M}{\sqrt{\lambda}})$-admissible with respect to $\mathcal{F}_\lambda$.

Lemma 16 *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 $\lambda>0$, and let $\mathcal{F}_\lambda=\left\{f\in\mathcal{F}\;\middle|\;\|f\|_K^2\leq\frac{M^2}{\lambda}\right\}$. Then the least squares ranking loss ℓ_{sq}, defined in Eq. (9) above, is $(2M+\frac{4\kappa M}{\sqrt{\lambda}})$-admissible with respect to $\mathcal{F}_\lambda$.*

Proof First, note that for any $f\in\mathcal{F}_\lambda$ and any $x\in\mathcal{X}$, we have from Eq. (6) that

$$|f(x)| \;\leq\; \kappa\|f\|_K \;\leq\; \frac{\kappa M}{\sqrt{\lambda}}\,.$$

Now, let $f_1,f_2\in\mathcal{F}_\lambda$, and let $(x,y),(x',y')\in(\mathcal{X}\times\mathcal{Y})$. Then we have,

$$\begin{aligned}
&\Big|\ell_{\mathrm{sq}}(f_1,(x,y),(x',y'))-\ell_{\mathrm{sq}}(f_1,(x,y),(x',y'))\Big|\\
&= \Big|\Big((y-y')-(f_1(x)-f_1(x'))\Big)^2-\Big((y-y')-(f_2(x)-f_2(x'))\Big)^2\Big|\\
&= \Big|2(y-y')-(f_1(x)-f_1(x'))-(f_2(x)-f_2(x'))\Big|\cdot\Big|(f_2(x)-f_2(x'))-(f_1(x)-f_1(x'))\Big|\\
&\leq \Big(2|y-y'|+|f_1(x)|+|f_1(x')|+|f_2(x)|+|f_2(x')|\Big)\cdot\Big(|f_1(x)-f_2(x)|+|f_1(x')-f_2(x')|\Big)\\
&\leq \Big(2M+\frac{4\kappa M}{\sqrt{\lambda}}\Big)\cdot\Big(|f_1(x)-f_2(x)|+|f_1(x')-f_2(x')|\Big)\,,
\end{aligned}$$

where the second equality follows from the identity $|a^2-b^2|=|a+b|\cdot|a-b|$. ■

An analysis of the proof of Lemma 10 shows that if a regularization-based ranking algorithm minimizing a regularized ℓ-error in some convex function class $\mathcal{F}$ is such that the ranking function returned by it for training samples of a given size m always lies in some effective search space $\mathcal{F}_m\subseteq\mathcal{F}$ that is also convex, then the result of the lemma holds even when for each m, the loss function ℓ is σ_m-admissible, for some $\sigma_m>0$, only with respect to the smaller function class $\mathcal{F}_m$. This in turn implies more general versions of Theorem 11 and Corollary 12, in which the same results hold even if a ranking algorithm minimizing a regularized ℓ-error in an RKHS $\mathcal{F}$ is such that for each m, ℓ is σ_m-admissible with respect to a convex subset $\mathcal{F}_m\subseteq\mathcal{F}$ that serves as an effective search space for training samples of size m. Thus, from Lemma 16 and the discussion preceding it, it follows that if $K(x,x)\leq\kappa^2$ for all $x\in\mathcal{X}$, then we can apply (the more general version of) Corollary 12 with $\sigma_m=(2M+\frac{4\kappa M}{\sqrt{\lambda}})$ (recall that $\lambda\equiv\lambda_m$ may depend on the sample size m) to get

that for any $0 < \delta < 1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$), the expected ranking error of the ranking function f_S learned by the algorithm $\mathcal{A}_{\text{sq}}$ is bounded by

$$R(f_S) \quad < \quad \widehat{R}_{\ell_1}(f_S;S) + \frac{64\kappa^2 M}{\lambda m}\left(1+\frac{2\kappa}{\sqrt{\lambda}}\right) + \left(\frac{32\kappa^2 M}{\lambda}\left(1+\frac{2\kappa}{\sqrt{\lambda}}\right)+M\right)\sqrt{\frac{2\ln(1/\delta)}{m}}.$$

5.2.3 CONSISTENCY

We can also use the above results to show consistency of kernel-based ranking algorithms. In particular, let $R^*_\ell(\mathcal{F})$ denote the optimal expected ℓ-error in an RKHS $\mathcal{F}$ (for a given distribution $\mathcal{D}$):

$$R^*_\ell(\mathcal{F}) \quad = \quad \inf_{f\in\mathcal{F}} R_\ell(f).$$

Then for bounded loss functions ℓ, we can show that with an appropriate choice of the regularization parameter λ, the expected ℓ-error $R_\ell(f_S)$ of the ranking function f_S learned by a kernel-based ranking algorithm that minimizes a regularized empirical ℓ-error in $\mathcal{F}$ converges (in probability) to this optimal value. To show this, we shall need the following simple lemma:

Lemma 17 *Let $f : \mathcal{X}\to\mathbb{R}$ be a fixed ranking function, and let ℓ be a bounded ranking loss function such that $0 \le \ell(f,(x,y),(x',y')) \le B$ for all $f : \mathcal{X}\to\mathbb{R}$ and $(x,y),(x',y') \in (\mathcal{X}\times\mathcal{Y})$. Then for any $0<\delta<1$, with probability at least $1-\delta$ over the draw of $S \in (\mathcal{X}\times\mathcal{Y})^m$ (according to $\mathcal{D}^m$),*

$$\widehat{R}_\ell(f;S) \quad < \quad R_\ell(f) + B\sqrt{\frac{2\ln(1/\delta)}{m}}.$$

The proof involves a simple application of McDiarmid's inequality to the random variable $\widehat{R}_\ell(f;S)$; details are provided in Appendix C. We then have the following consistency result:

Theorem 18 *Let $\mathcal{F}$ be an RKHS with kernel K such that for all $x \in \mathcal{X}$, $K(x,x) \le \kappa^2 < \infty$. Let ℓ be a ranking loss such that $\ell(f,(x,y),(x',y'))$ is convex in f and ℓ is σ-admissible with respect to $\mathcal{F}$. Furthermore, let ℓ be bounded such that $0 \le \ell(f,(x,y),(x',y')) \le B$ for all $f : \mathcal{X}\to\mathbb{R}$ and $(x,y),(x',y') \in (\mathcal{X}\times\mathcal{Y})$, and let f^*_ℓ be any fixed function in $\mathcal{F}$ that satisfies*

$$R_\ell(f^*_\ell) \quad \le \quad R^*_\ell(\mathcal{F}) + \frac{1}{m}. \tag{10}$$

(Note that such a function f^*_ℓ exists by definition of $R^*_\ell(\mathcal{F})$.) Let $\lambda > 0$, and let $\mathcal{A}$ be a ranking algorithm that, given a training sample S, outputs a ranking function $f_S \in \mathcal{F}$ that satisfies $f_S = \arg\min_{f\in\mathcal{F}}\left\{\widehat{R}_\ell(f;S) + \lambda\|f\|_K^2\right\}$. Then we have that for any $0<\delta<1$, with probability at least $1-\delta$ over the draw of S (according to $\mathcal{D}^m$),

$$R_\ell(f_S) \quad < \quad R^*_\ell(\mathcal{F}) + \lambda\|f^*_\ell\|_K^2 + \frac{1}{m} + \frac{32\sigma^2\kappa^2}{\lambda m} + \left(\frac{16\sigma^2\kappa^2}{\lambda}+2B\right)\sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Thus, if $\lambda = o(1)$ and $\lambda = \omega(\frac{1}{\sqrt{m}})$ (for example, if $\lambda = m^{-1/4}$), then $R_\ell(f_S)$ converges in probability to the optimal value $R^*_\ell(\mathcal{F})$ (as $m\to\infty$).

Proof As in Corollary 15, we can use Corollary 14 and apply Theorem 6 with $\frac{\delta}{2}$ to get that with probability at least $1-\frac{\delta}{2}$,

$$R_\ell(f_S) \quad < \quad \widehat{R}_\ell(f_S;S) + \frac{32\sigma^2\kappa^2}{\lambda m} + \left(\frac{16\sigma^2\kappa^2}{\lambda} + B\right)\sqrt{\frac{2\ln(2/\delta)}{m}}. \tag{11}$$

Now,

$$\begin{aligned}\widehat{R}_\ell(f_S;S) &\leq \widehat{R}_\ell(f_S;S) + \lambda\|f_S\|_K^2 \\ &\leq \widehat{R}_\ell(f_\ell^*;S) + \lambda\|f_\ell^*\|_K^2,\end{aligned}$$

where the first inequality is due to non-negativity of $\|f_S\|_K^2$ and the second inequality follows from the definition of f_S. Applying Lemma 17 to f_ℓ^* with $\frac{\delta}{2}$, we thus get that with probability at least $1-\frac{\delta}{2}$,

$$\widehat{R}_\ell(f_S;S) \quad < \quad \left(R_\ell(f_\ell^*) + B\sqrt{\frac{2\ln(2/\delta)}{m}}\right) + \lambda\|f_\ell^*\|_K^2. \tag{12}$$

Combining the inequalities in Eqs. (11-12), each of which holds with probability at least $1-\frac{\delta}{2}$, together with the condition in Eq. (10), gives the desired result. ■

6. Comparisons With Related Work

In the above sections we have derived stability-based generalization bounds for ranking algorithms in a setting that is more general than what has been considered previously, and have shown that kernel-based ranking algorithms that perform regularization in an RKHS in this setting satisfy the required stability conditions. In this section we discuss how our results relate to other recent studies.

6.1 Comparison with Stability Bounds for Classification/Regression

As pointed out earlier, our stability analysis of ranking algorithms is based on a similar analysis for classification and regression algorithms by Bousquet and Elisseeff (2002); therefore it is instructive to compare the generalization bounds we obtain here with those obtained in that work. Such a comparison shows that the bounds we obtain here for ranking are very similar to those obtained for classification and regression, differing only in constants (and, of course, in the precise definition of stability, which is problem-dependent). The difference in constants in the two bounds is due in part to the difference in loss functions in ranking and classification/regression, 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). As discussed in Section 4, the reason that we are able to obtain bounds for ranking algorithms using the same methods as those of Bousquet and Elisseeff lies in the power of McDiarmid's inequality, which was used by Bousquet and Elisseeff to capture the effect of stability but is also general enough to capture the structure of the ranking problem.

It is also instructive to compare our bounds with those obtained for bipartite ranking algorithms (Agarwal and Niyogi, 2005). In particular, the bounds for kernel-based ranking algorithms in the bipartite setting differ from our bounds in that the sample size m in our case is replaced with the term $mn/(m+n)$, where m,n denote the numbers of positive and negative examples, respectively, in the bipartite setting. This suggests that in general, the effective sample size in ranking is the ratio between the number of pairs in the training sample (mn in the bipartite setting) and the total number of examples ($m+n$); indeed, in our setting, this ratio is $\binom{m}{2}/m = m/2$, which fits our bounds.

6.2 Comparison with Uniform Convergence Bounds

While uniform convergence bounds for ranking have not been derived explicitly in the setting we consider, it is not hard to see that the techniques used to derive such bounds in the settings of Agarwal et al. (2005) and Clemencon et al. (2008) can be extended to obtain similar bounds in our setting. The crucial difference between such bounds and those derived in this paper is that uniform convergence bounds depend on the complexity of the function class searched by an algorithm. As discussed in the context of bipartite ranking (Agarwal and Niyogi, 2005), this means that such bounds can quickly become uninformative in high dimensions; in the case of function classes whose complexity cannot be bounded (such as the RKHS corresponding to a Gaussian kernel), uniform convergence bounds cannot be applied at all. In both these cases, the stability analysis provides a more useful viewpoint.

6.3 Comparison with Cortes et al. (2007)

The work that is perhaps most closely related to ours is that of Cortes et al. (2007), who study "magnitude-preserving" ranking algorithms—most of which perform regularization in an RKHS—and also use algorithmic stability to derive generalization bounds for such algorithms. The setting considered by Cortes et al. is similar to ours: the learner is given a finite sequence of labeled training examples $((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X}\times\mathbb{R})^m$, and the goal is to learn a real-valued function $f : \mathcal{X}\to\mathbb{R}$ that ranks future instances with larger labels higher than those with smaller labels. The term "magnitude-preserving" comes from their view of the magnitude of the difference $(y-y')$ in the labels of two examples $(x,y),(x',y')$ as not just a penalty for mis-ranking the pair (x,x'), but as a quantity to be explicitly preserved by f, in the sense that $(f(x)-f(x'))$ be close to $(y-y')$.

There are three important differences between our work and that of Cortes et al. (2007). First, Cortes et al. implicitly assume the existence of a "truth" function: their results are derived under that assumption that only instances $x \in \mathcal{X}$ are drawn randomly, and that labels $y \in \mathcal{Y}$ are then determined according to a "truth" function $f^* : \mathcal{X}\to\mathcal{Y}$. In our work, we make the more general distributional assumption that all examples (x,y) are drawn randomly according to a joint distribution on $\mathcal{X}\times\mathcal{Y}$.

Second, Cortes et al. consider only the notion of uniform loss stability; they do not consider uniform score stability. As a consequence, the bounds they obtain for ranking algorithms performing regularization in an RKHS are similar to our bound in Corollary 15, which bounds the expected ranking error $R(f_S)$ of a learned ranking function f_S in terms of the empirical ℓ-error $\widehat{R}_\ell(f_S;S)$, where ℓ is the loss minimized by the algorithm.[4] In contrast, our bound in Corollary 12, which

4. Cortes et al. do not actually bound the expected ranking error $R(f_S)$ (which in our view is the primary performance measure of a ranking function in our setting); their results simply give bounds on the expected ℓ-error $R_\ell(f_S)$, where ℓ again is the loss minimized by the algorithm. However, for $\ell_{\text{disc}} \le \ell$, one can easily deduce bounds on $R(f_S)$ from these results.

makes use of the notion of score stability, bounds the expected error $R(f_S)$ in terms of the empirical ℓ_1-error $\widehat{R}_{\ell_1}(f_S;S)$. As mentioned in Section 5.2, the ℓ_1 loss tends to be smaller than many of the loss functions minimized by kernel-based ranking algorithms in practice (indeed, it is never greater than the hinge ranking loss ℓ_h), thus leading to tighter bounds for the same algorithms.

The third and most important difference between our work and that of Cortes et al. is that the results of Cortes et al. involve certain strong assumptions on the function class searched by an algorithm. In particular, they assume that the function class $\mathcal{F}$ searched by a ranking algorithm is bounded in the sense that $\exists\, M' > 0$ such that for all $f \in \mathcal{F}$ and all $x \in \mathcal{X}$, $|f(x) - f^*(x)| \leq M'$ (where, as mentioned above, $f^* : \mathcal{X} \to \mathbb{R}$ is the truth function in their setting). No RKHS $\mathcal{F}$ satisfies this condition, and so their results as stated do not apply to the kernel-based ranking algorithms they consider. On closer inspection, one notices that for the loss functions they show to be 1-admissible, they do not actually need this assumption. For the loss functions they show to be $4M'$-admissible, however, a non-trivial fix is required. We expect that an approach similar to that of separating out the effective search space, as we have done in the case of the least squares ranking loss in Section 5.2.2, should work; indeed, in the case of the least squares ranking loss, assuming a bounded label space Y, it is easy to see from the observations in Section 5.2.2 that $|f(x) - y|$ is bounded for all f in the *effective* search space (and all $(x,y) \in (\mathcal{X} \times \mathcal{Y})$).

6.4 Comparison with Rudin et al. (2005)

Rudin et al. (2005) studied a different setting of the ranking problem, in which one assumes the existence of a pair-wise "truth" function $\pi : \mathcal{X} \times \mathcal{X} \to \{0,1\}$ that assigns a binary ranking preference $\pi(x,x')$ to each pair of instances (x,x'), such that $\pi(x,x') = 1 \Rightarrow \pi(x',x) = 0$: if $\pi(x,x') = 1$, then x is to be ranked higher than x'; if $\pi(x',x) = 1$, then x' is to be ranked higher than x; and if $\pi(x,x') = \pi(x',x) = 0$, then there is no ranking preference between x and x'. The training sample given to a learner in this setting consists of a finite number of instances $x_1, \ldots, x_m$, each drawn randomly and independently according to some (unknown) distribution on $\mathcal{X}$, together with the corresponding ranking preferences $\pi(x_i, x_j)$ for $i, j \in \{1, \ldots, m\}, i \neq j$. Rudin et al. derived a generalization bound for this setting using techniques inspired by the works of Cucker and Smale (2002), Koltchinskii and Panchenko (2002), and Bousquet (2003). Their bound, expressed in terms of covering numbers, is a margin-based bound, which is of interest when the learned ranking function has zero empirical error on the training sample.

Noting that the labels y_i, y_j corresponding to a pair of instances x_i, x_j in our setting are used in our results only in the form of ranking preferences $(y_i - y_j)$, we can clearly view the setting of Rudin et al. as a special case of ours: the ranking preferences $(y_i - y_j)$, which are random variables in our setting, get replaced by the deterministic preferences $\pi(x_i, x_j)$. All quantities can be adapted accordingly; for example, the expected ranking error of a ranking function $f : \mathcal{X} \to \mathbb{R}$ in this setting (with respect to a distribution $\mathcal{D}$ on $\mathcal{X}$) becomes

$$R(f) \;=\; \mathbf{E}_{(X,X') \sim \mathcal{D} \times \mathcal{D}} \left[|\pi(X,X')| \left(\mathbf{I}_{\{\pi(X,X')(f(X) - f(X')) < 0\}} + \frac{1}{2} \mathbf{I}_{\{f(X) = f(X')\}} \right) \right].$$

In fact, rather than restrict ourselves to binary preferences, we can extend the setting of Rudin et al. to allow real-valued preferences, requiring that $\pi(x,x') = -\pi(x',x)$ for all $x, x' \in \mathcal{X}$; if π takes values in $[-M, M]$, we get the same results as we have derived in our setting (with the probabilities now being over the random draw of instances only).

We can also consider another extension of interest in the pair-wise truth function setting, where we may not have access to the preferences for all pairs of instances in the training sample, but rather are given preferences for only a limited number of pairs; in this case, the learner receives a sequence of instances $S = (x_1, \ldots, x_m)$, and preferences $\pi(x_i, x_j)$ for only a subset of pairs $E \subseteq \{(i,j) \mid 1 \leq i < j \leq m\}$. It is interesting then to consider a model in which not only the instances in S but also the pairs in E for which the ranking preferences are provided may be chosen randomly. If the instances in S are drawn randomly and independently according to some distribution $\mathcal{D}$ on $\mathcal{X}$, and the set E is drawn randomly (independently of S) from some distribution $\mathcal{E}_m$ on the set of possible (undirected) edge sets for a graph on m vertices $\{1, \ldots, m\}$ (for example, E could be obtained by including each edge with some fixed probability $0 < p_m < 1$, or by selecting at random a subset of α_m edges for some $\alpha_m \leq \binom{m}{2}$), then under some natural conditions on $\mathcal{E}_m$, we can extend our techniques to obtain generalization bounds in this setting too. In the remainder of this section we discuss some details of this extension.

In particular, the empirical ranking error of a ranking function $f : \mathcal{X} \to \mathbb{R}$ in the above setting, with respect to a training sample $T = (S, E, \pi|_{(S,E)})$ where S and E are as above and $\pi|_{(S,E)}$ is the restriction of π to $\{(x_i, x_j) \mid (i,j) \in E\}$, is given by

$$\widehat{R}(f;T) \;=\; \frac{1}{|E|} \sum_{(i,j)\in E} |\pi(x_i,x_j)| \left(\mathbf{I}_{\{\pi(x_i,x_j)(f(x_i)-f(x_j))<0\}} + \frac{1}{2}\mathbf{I}_{\{f(x_i)=f(x_j)\}} \right).$$

Given a loss function ℓ that assigns for any ranking function f, any pair of instances (x, x') and any $r \in [-M, M]$ a non-negative loss $\ell(f, x, x', r)$, the expected and empirical ℓ-errors can be defined similarly:

$$\begin{aligned} R_\ell(f) &= \mathbf{E}_{(X,X')\sim\mathcal{D}\times\mathcal{D}} \left[\ell(f, X, X', \pi(X, X')) \right]; \\ \widehat{R}_\ell(f;T) &= \frac{1}{|E|} \sum_{(i,j)\in E} \ell(f, x_i, x_j, \pi(x_i, x_j)). \end{aligned}$$

A ranking algorithm $\mathcal{A}$ in this setting, which given a training sample T outputs a ranking function f_T, has uniform loss stability β with respect to ℓ if for all $m \in \mathbb{N}$, $S \in \mathcal{X}^m$, $E \subseteq \{(i,j) \mid 1 \leq i < j \leq m\}$, $\pi : \mathcal{X} \times \mathcal{X} \to [-M, M]$, $1 \leq i \leq m$ and $x_i' \in \mathcal{X}$, we have for all $x, x' \in \mathcal{X}$ and all $r \in [-M, M]$,

$$\left| \ell(f_T, x, x', r) - \ell(f_{T^i}, x, x', r) \right| \;\leq\; \beta(m),$$

where $T = (S, E, \pi|_{(S,E)})$ as above and $T^i = (S^i, E, \pi|_{(S^i,E)})$, with S^i being the sequence obtained from S by replacing x_i with x_i'. Then we can show the following bound for ranking algorithms with good loss stability in this setting:

Theorem 19 *Let $\pi : \mathcal{X} \times \mathcal{X} \to [-M, M]$ be a pair-wise truth function such that $\pi(x, x') = -\pi(x', x)$ for all $x, x' \in \mathcal{X}$. Let $\mathcal{A}$ be a symmetric ranking algorithm in the pair-wise truth function setting whose output on a training sample $T = (S, E, \pi|_{(S,E)})$ we denote by f_T, and let ℓ be a bounded ranking loss function such that $0 \leq \ell(f, x, x', r) \leq B$ for all $f : \mathcal{X} \to \mathbb{R}$, $x, x' \in \mathcal{X}$ and $r \in [-M, M]$. Let $\beta : \mathbb{N} \to \mathbb{R}$ be such that $\mathcal{A}$ has uniform loss stability β with respect to ℓ as defined above. Let $\mathcal{D}$ be any distribution on $\mathcal{X}$, and let $\{\mathcal{E}_m\}$ be any family of distributions on the sets of possible (undirected) edge sets for a graph on m vertices $\{1, \ldots, m\}$ that satisfies the following condition: $\exists\, s > 1$ and a sequence (δ_m) satisfying $\delta_m \geq 0$, $\lim_{m\to\infty} \delta_m = 0$, such that for all m, $\mathbf{P}_{E\sim\mathcal{E}_m}(|E| < m^s) \leq \delta_m$. Then*

for any $0 < \delta < 1$*, for all* m *large enough such that* $\delta_m \leq \frac{\delta}{2}$*, we have with probability at least* $1-\delta$ *over the draw of* (S,E) *according to* $\mathcal{D}^m \times \mathcal{E}_m$*,*

$$R_\ell(f_T) \quad < \quad \widehat{R}_\ell(f_T;T) + 2\beta(m) + \sqrt{\frac{1}{2}\left(4m(\beta(m))^2 + 8B\beta(m) + \frac{2B^2}{m-1} + \frac{B^2}{m^{s-1}}\right)\ln(2/\delta)}\,.$$

The proof makes use of McDiarmid's inequality and is similar to that of Theorem 6, with two main differences. First, McDiarmid's inequality is first applied to obtain a bound conditioned on an edge set E with $|E| \geq m^s$; the bound on the probability that $|E| < m^s$ then allows us to obtain the unconditional bound. Second, the constants c_k in the application of McDiarmid's inequality are different for each k, and depend on the degrees of the corresponding vertices in the graph with edge set E; the sum $\sum_k c_k^2$ is then bounded using a bound on the sum of squares of degrees in a graph due to de Caen (1998). Details of the proof are provided in Appendix D.

The condition on $\mathcal{E}_m$ in the above theorem states that with high probability (probability at least $1-\delta_m$), an edge set E drawn according to $\mathcal{E}_m$ has at least m^s edges for some $s > 1$, that is, $|E|$ is super-linear in m.[5] Thus, in the pair-wise truth function setting, we need not have access to the ranking preferences $\pi(x_i, x_j)$ for all $\binom{m}{2}$ pairs of instances in S; having the preferences for any super-linear number of pairs (where the pairs are selected independently of the instances S themselves) suffices to give a generalization bound (albeit with a correspondingly slower convergence rate, as quantified by the $\frac{1}{m^{s-1}}$ term in the bound above). Below we give examples of two families of distributions $\{\mathcal{E}_m\}$ that satisfy the above condition.

Example 1 *Fix any* $1 < s < 2$*, and let* $\mathcal{E}_m$ *be the distribution that corresponds to selecting a random subset of* α_m *edges, with* $\alpha_m = \min\left(\binom{m}{2}, \lceil m^s \rceil\right)$*. Then for large enough* m*,* $\alpha_m = \lceil m^s \rceil$*, and* $\mathbf{P}_{E\sim\mathcal{E}_m}(|E| < m^s) = 0$*.*

Example 2 *Fix any* $1 < s < 2$*, and let* $\mathcal{E}_m$ *be the distribution that corresponds to including each edge with a fixed probability* p_m*, with* $p_m = \min\left(1, \frac{8}{m^{2-s}}\right)$*. Then for large enough* m*,* $p_m = \frac{8}{m^{2-s}}$*, and assuming without loss of generality that* $m \geq 2$*, it is easy to show using a Chernoff bound that in this case,* $\mathbf{P}_{E\sim\mathcal{E}_m}(|E| < m^s) \leq e^{-m^s/4}$*. Indeed, for any such* m*, let* Z_m *be the random variable equal to the number of edges* $|E|$*. Then* Z_m *is a binomial random variable with parameters* $\binom{m}{2}$ *and* $p_m = \frac{8}{m^{2-s}}$*, and*

$$\begin{aligned}
\mathbf{E}[Z_m] &= \tbinom{m}{2} p_m \\
&= 4(m-1)m^{s-1} \\
&= 2(m^s + (m-2)m^{s-1}) \\
&\geq 2m^s\,,
\end{aligned}$$

where the inequality follows from our assumption that $m \geq 2$*. The Chernoff bound for deviations below the mean of a binomial random variable tells us that for any* $0 < \delta < 1$*,* $\mathbf{P}(Z_m < (1-\delta)\mathbf{E}[Z_m]) \leq e^{-\mathbf{E}[Z_m]\delta^2/2}$*. Thus we have*

$$\mathbf{P}(Z_m < m^s) \quad = \quad \mathbf{P}\left(Z_m < \left(1 - \frac{1}{2}\right)2m^s\right)$$

5. Note that in the statement of Theorem 19 we require $|E|$ to be super-linear in m by a polynomial factor; a smaller super-linear growth also leads to a generalization bound, but with a slower convergence rate.

$$\begin{aligned} &\leq \mathbf{P}\Big(Z_m < \Big(1-\frac{1}{2}\Big)\mathbf{E}[Z_m]\Big) \\ &\leq e^{-\mathbf{E}[Z_m](\frac{1}{2})^2/2} \\ &\leq e^{-m^s/4}. \end{aligned}$$

Comparing the bound in Theorem 19 with that of Theorem 6, it is worth noting that the effective sample size in the above setting becomes m^{s-1}. This is consistent with the discussion in Section 6.1, in that the ratio between the number of pairs in the training sample and the total number of examples in this case is (with high probability) at least $m^s/m = m^{s-1}$.

7. Discussion

Our goal in this paper has been to study generalization properties of ranking algorithms in a setting where ranking preferences among instances are indicated by real-valued labels on the instances. This setting of the ranking problem arises frequently in practice and is more general than those considered previously. We have derived generalization bounds for ranking algorithms in this setting using the notion of algorithmic stability; in particular, we have shown that ranking algorithms that exhibit good stability properties also have good generalization properties, and have applied our results to obtain generalization bounds for kernel-based ranking algorithms that perform regularization in a reproducing kernel Hilbert space. Such algorithms often cannot be analyzed using uniform convergence results.

The main difference in the mathematical formulation of ranking problems as compared to classification or regression problems is that the loss function in ranking is 'pair-wise' rather than 'point-wise'. Indeed, ranking often resembles weighted 'classification on pairs', with the weights being given by the corresponding ranking preferences (although learning a real-valued function that induces a total ordering on the instance space is not quite the same as learning a binary-valued function on instance pairs that simply decides which of two instances should be ranked higher, Cohen et al. 1999; Balcan et al. 2007; Ailon and Mohri 2008). However, generalization bounds from classification cannot be applied directly to ranking, due to dependences among the instance pairs. As discussed earlier, the reason that we are able to obtain bounds for ranking algorithms using the same methods as those used by Bousquet and Elisseeff (2002) for classification and regression algorithms lies in the power of McDiarmid's inequality, which was used by Bousquet and Elisseeff to capture the effect of stability but is also general enough to capture the structure of the ranking problems we consider. A comparison of our results with those of Bousquet and Elisseeff (2002) and Agarwal and Niyogi (2005) suggests that the effective sample size in ranking is proportional to the ratio between the number of pairs in the training sample and the total number of examples; this can be smaller than the number of examples m if ranking preferences are provided for less than m^2 pairs.

The notions of uniform stability studied in this paper correspond most closely to those studied by Bousquet and Elisseeff (2002). These notions are strict in that they require changes in a sample to have bounded effect uniformly over all samples and replacements. One can define weaker notions of stability, analogous to the hypothesis stability considered by Devroye and Wagner (1979), Kearns and Ron (1999), and Bousquet and Elisseeff (2002), or the almost-everywhere stability considered by Kutin and Niyogi (2002), which would require the bounds to hold only in expectation or with high probability. Such notions would lead 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.

For the sake of simplicity and to keep the focus on the main ideas involved in applying stability techniques to ranking, we have focused in this paper on bounding the expected ranking error in terms of the empirical ranking error. In classification and regression, stability analysis has also been used to provide generalization bounds in terms of the leave-one-out error (Bousquet and Elisseeff, 2002; Kearns and Ron, 1999), and with a slight change in definitions of stability, similar results can be obtained in the case of ranking as well. In particular, in this case we need to define (for a ranking algorithm which given a training sample S returns a ranking function f_S) a 'leave-two-out' ranking error as follows:

$$\widetilde{R}_\ell(f_S;S) \quad = \quad \frac{1}{\binom{m}{2}} \sum_{1\le i<j\le m} \ell(f_{S^{\setminus ij}},(x_i,y_i),(x_j,y_j))\,,$$

where $S^{\setminus ij}$ denotes the sequence obtained by removing the ith and jth examples from a sample $S=((x_1,y_1),\ldots,(x_m,y_m))\in(\mathcal{X}\times\mathcal{Y})^m$. Then, defining notions of stability in terms of changes to a sample that consist of removing two examples rather than replacing one example, one can obtain very similar generalization bounds for ranking algorithms in terms of the above leave-two-out error as those we have derived in terms of the empirical error.

An open question concerns the analysis of other ranking algorithms using the algorithmic stability framework. For example, it has been shown that the AdaBoost algorithm for classification is stability-preserving, in the sense that stability of base classifiers implies stability of the final learned classifier (Kutin and Niyogi, 2001). It would be interesting if a similar result could be shown for the RankBoost ranking algorithm (Freund et al., 2003), 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 even more general settings of the ranking problem. For example, a more general setting would be one in which a finite number of instances $x_1,\ldots,x_m\in\mathcal{X}$ are drawn randomly and independently according to some distribution $\mathcal{D}$ on $\mathcal{X}$, and then pair-wise ranking preferences $r_{ij}\in[-M,M]$ for $i<j$ are drawn from a conditional distribution, conditioned on the corresponding pair of instances (x_i,x_j). We are not aware of any generalization bounds for such a setting.

Acknowledgments

We would like to thank the anonymous referees for constructive comments that helped improve the paper. SA is supported in part by NSF award DMS-0732334. PN thanks the NSF for financial support.

Appendix A. Proof of Theorem 6

Our main tool will be the following powerful concentration inequality of McDiarmid (1989), which bounds the deviation of any function of a sample for which a single change in the sample has limited effect.

Theorem 20 (McDiarmid 1989) *Let $X_1,\ldots,X_m$ be independent random variables, each taking values in a set A. Let $\phi : A^m \to \mathbb{R}$ be such that for each $1 \leq k \leq m$, there exists $c_k > 0$ such that*

$$\sup_{x_1,\ldots,x_m \in A, x_k' \in A} \left| \phi(x_1,\ldots,x_m) - \phi(x_1,\ldots,x_{k-1},x_k',x_{k+1},\ldots,x_m) \right| \quad \leq \quad c_k \,.$$

Then for any $\varepsilon > 0$,

$$\mathbf{P}\Big(\phi(X_1,\ldots,X_m) - \mathbf{E}\left[\phi(X_1,\ldots,X_m)\right] \geq \varepsilon \Big) \quad \leq \quad e^{-2\varepsilon^2 / \sum_{k=1}^m c_k^2} \,.$$

Before we apply McDiarmid's inequality to prove Theorem 6, we shall need the following technical lemma. In the following, we shall drop explicit references to distributions where clear from context, replacing, for example, $\mathbf{E}_{(X,Y)\sim\mathcal{D}}[\ldots]$ with simply $\mathbf{E}_{(X,Y)}[\ldots]$.

Lemma 21 *Let $\mathcal{A}$ be a symmetric ranking algorithm whose output on a training sample $S \in (\mathcal{X} \times \mathcal{Y})^m$ we denote by f_S, and let ℓ be a ranking loss function. Then for all $1 \leq i < j \leq m$, we have*

$$\begin{aligned} &\mathbf{E}_{S\sim\mathcal{D}^m}\left[R_\ell(f_S) - \widehat{R}_\ell(f_S;S) \right] \\ &\quad = \quad \mathbf{E}_{S\sim\mathcal{D}^m,((X_i',Y_i'),(X_j',Y_j'))\sim\mathcal{D}\times\mathcal{D}}\Big[\ell(f_S,(X_i',Y_i'),(X_j',Y_j')) - \ell(f_{S^{i,j}},(X_i',Y_i'),(X_j',Y_j')) \Big] \,. \end{aligned}$$

Proof Denoting $S = ((X_1,Y_1),\ldots,(X_m,Y_m))$, we have by linearity of expectation,

$$\mathbf{E}_S\Big[\widehat{R}_\ell(f_S;S) \Big] \quad = \quad \frac{1}{\binom{m}{2}} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathbf{E}_S\Big[\ell(f_S,(X_i,Y_i),(X_j,Y_j)) \Big] \,.$$

By symmetry, the term in the summation is the same for all i, j. Therefore, for all $1 \leq i < j \leq m$, we get

$$\begin{aligned} \mathbf{E}_S\Big[\widehat{R}_\ell(f_S;S) \Big] \quad &= \quad \mathbf{E}_S\Big[\ell(f_S,(X_i,Y_i),(X_j,Y_j)) \Big] \\ &= \quad \mathbf{E}_{S,((X_i',Y_i'),(X_j',Y_j'))}\Big[\ell(f_S,(X_i,Y_i),(X_j,Y_j)) \Big] \,. \end{aligned}$$

Interchanging the roles of (X_i,Y_i) with (X_i',Y_i') and (X_j,Y_j) with (X_j',Y_j'), we get

$$\mathbf{E}_S\Big[\widehat{R}_\ell(f_S;S) \Big] \quad = \quad \mathbf{E}_{S,((X_i',Y_i'),(X_j',Y_j'))}\Big[\ell(f_{S^{i,j}},(X_i',Y_i'),(X_j',Y_j')) \Big] \,.$$

Since by definition

$$\mathbf{E}_S\Big[R_\ell(f_S) \Big] \quad = \quad \mathbf{E}_{S,((X_i',Y_i'),(X_j',Y_j'))}\Big[\ell(f_S,(X_i',Y_i'),(X_j',Y_j')) \Big] \,,$$

the result follows. ■

Proof *(of Theorem 6)*
Let $\phi : (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$ be defined as follows:

$$\phi(S) \quad = \quad R_\ell(f_S) - \widehat{R}_\ell(f_S;S) \,.$$

We shall show that ϕ satisfies the conditions of McDiarmid's inequality (Theorem 20). Let $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X}\times\mathcal{Y})^m$. Then for each $1 \le k \le m$, we have for any $(x_k',y_k') \in (\mathcal{X}\times\mathcal{Y})$:

$$\begin{aligned}\left|\phi(S)-\phi(S^k)\right| &= \left|\left(R_\ell(f_S)-\widehat{R}_\ell(f_S;S)\right)-\left(R_\ell(f_{S^k})-\widehat{R}_\ell(f_{S^k};S^k)\right)\right| \\ &\le \left|R_\ell(f_S)-R_\ell(f_{S^k})\right|+\left|\widehat{R}_\ell(f_S;S)-\widehat{R}_\ell(f_{S^k};S^k)\right|.\end{aligned}$$

Now,

$$\begin{aligned}\left|R_\ell(f_S)-R_\ell(f_{S^k})\right| &= \left|\mathbf{E}_{((X,Y),(X',Y'))}\left[\ell(f_S,(X,Y),(X',Y'))-\ell(f_{S^k},(X,Y),(X',Y'))\right]\right| \\ &\le \mathbf{E}_{((X,Y),(X',Y'))}\left[\left|\ell(f_S,(X,Y),(X',Y'))-\ell(f_{S_k},(X,Y),(X',Y'))\right|\right] \\ &\le \beta(m),\end{aligned}$$

and

$$\begin{aligned}&\left|\widehat{R}_\ell(f_S;S)-\widehat{R}_\ell(f_{S^k};S^k)\right| \\ &\le \frac{1}{\binom{m}{2}}\sum_{1\le i<j\le m, i\ne k, j\ne k}\left|\ell(f_S,(x_i,y_i),(x_j,y_j))-\ell(f_{S^k},(x_i,y_i),(x_j,y_j))\right| \\ &\quad+\frac{1}{\binom{m}{2}}\sum_{i\ne k}\left|\ell(f_S,(x_i,y_i),(x_k,y_k))-\ell(f_{S^k},(x_i,y_i),(x_k',y_k'))\right| \\ &\le \frac{1}{\binom{m}{2}}\sum_{1\le i<j\le m, i\ne k, j\ne k}\beta(m)\;+\;\frac{1}{\binom{m}{2}}\sum_{i\ne k}B \\ &= \frac{1}{\binom{m}{2}}\left(\left(\binom{m}{2}-(m-1)\right)\beta(m)+(m-1)B\right) \\ &< \beta(m)+\frac{2B}{m}.\end{aligned}$$

Thus we have

$$\left|\phi(S)-\phi(S^k)\right| \le 2\left(\beta(m)+\frac{B}{m}\right).$$

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

$$\begin{aligned}&\mathbf{P}_S\left(\left(R_\ell(f_S)-\widehat{R}_\ell(f_S;S)\right)-\mathbf{E}_S\left[R_\ell(f_S)-\widehat{R}_\ell(f_S;S)\right]\ge\varepsilon\right) \\ &\qquad\le e^{-2\varepsilon^2/m(2(\beta(m)+\frac{B}{m}))^2} \\ &\qquad= e^{-m\varepsilon^2/2(m\beta(m)+B)^2}.\end{aligned}$$

Now, by Lemma 21, we have (for any $1\le i<j\le m$),

$$\begin{aligned}&\mathbf{E}_S\left[R_\ell(f_S)-\widehat{R}_\ell(f_S;S)\right] \\ &\quad= \mathbf{E}_{S,((X_i',Y_i'),(X_j',Y_j'))}\left[\ell(f_S,(X_i',Y_i'),(X_j',Y_j'))-\ell(f_{S^{i,j}},(X_i',Y_i'),(X_j',Y_j'))\right] \\ &\quad\le \mathbf{E}_{S,((X_i',Y_i'),(X_j',Y_j'))}\left[\left|\ell(f_S,(X_i',Y_i'),(X_j',Y_j'))-\ell(f_{S^i},(X_i',Y_i'),(X_j',Y_j'))\right|\right] \\ &\qquad+\mathbf{E}_{S,((X_i',Y_i'),(X_j',Y_j'))}\left[\left|\ell(f_{S^i},(X_i',Y_i'),(X_j',Y_j'))-\ell(f_{S^{i,j}},(X_i',Y_i'),(X_j',Y_j'))\right|\right] \\ &\quad\le 2\beta(m).\end{aligned}$$

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

$$\mathbf{P}_S\left(R_\ell(f_S) - \widehat{R}_\ell(f_S;S) - 2\beta(m) \geq \varepsilon\right) \quad \leq \quad e^{-m\varepsilon^2/2(m\beta(m)+B)^2}.$$

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

Appendix B. Proof of Lemma 10

Proof *(of Lemma 10)*
Recall that a convex function $\phi : \mathcal{U} \to \mathbb{R}$ satisfies for all $u, v \in \mathcal{U}$ and for all $t \in [0,1]$,

$$\phi(u + t(v-u)) - \phi(u) \;\leq\; t(\phi(v) - \phi(u)).$$

Since $\ell(f,(x,y),(x',y'))$ is convex in f, we have that $\widehat{R}_\ell(f;S)$ is convex in f. Therefore for any $t \in [0,1]$, we have

$$\widehat{R}_\ell(f + t\Delta f;S) - \widehat{R}_\ell(f;S) \quad \leq \quad t\left(\widehat{R}_\ell(f^i;S) - \widehat{R}_\ell(f;S)\right), \tag{13}$$

and also (interchanging the roles of f and f^i),

$$\widehat{R}_\ell(f^i - t\Delta f;S) - \widehat{R}_\ell(f^i;S) \quad \leq \quad t\left(\widehat{R}_\ell(f;S) - \widehat{R}_\ell(f^i;S)\right). \tag{14}$$

Adding Eqs. (13) and (14), we get

$$\widehat{R}_\ell(f + t\Delta f;S) - \widehat{R}_\ell(f;S) + \widehat{R}_\ell(f^i - t\Delta f;S) - \widehat{R}_\ell(f^i;S) \quad \leq \quad 0. \tag{15}$$

Now, since $\mathcal{F}$ is convex, we have that $(f + t\Delta f) \in \mathcal{F}$ and $(f^i - t\Delta f) \in \mathcal{F}$. Since f minimizes $\widehat{R}^{\lambda}_\ell(f;S)$ in $\mathcal{F}$ and f^i minimizes $\widehat{R}^{\lambda}_\ell(f;S^i)$ in $\mathcal{F}$, we thus have

$$\widehat{R}^{\lambda}_\ell(f;S) - \widehat{R}^{\lambda}_\ell(f + t\Delta f;S) \quad \leq \quad 0, \tag{16}$$

$$\widehat{R}^{\lambda}_\ell(f^i;S^i) - \widehat{R}^{\lambda}_\ell(f^i - t\Delta f;S^i) \quad \leq \quad 0. \tag{17}$$

Adding Eqs. (15), (16) and (17), we get

$$\begin{aligned}
&\lambda\Big(N(f) - N(f + t\Delta f) + N(f^i) - N(f^i - t\Delta f)\Big) \\
&\leq \quad \widehat{R}_\ell(f^i;S) - \widehat{R}_\ell(f^i;S^i) + \widehat{R}_\ell(f^i - t\Delta f;S^i) - \widehat{R}_\ell(f^i - t\Delta f;S) \\
&= \quad \frac{1}{\binom{m}{2}} \sum_{j \neq i} \Big(\ell(f^i,(x_i,y_i),(x_j,y_j)) - \ell(f^i,(x'_i,y'_i),(x_j,y_j)) \\
&\qquad\qquad + \ell(f^i - t\Delta f,(x'_i,y'_i),(x_j,y_j)) - \ell(f^i - t\Delta f,(x_i,y_i),(x_j,y_j))\Big) \\
&= \quad \frac{1}{\binom{m}{2}} \sum_{j \neq i} \Big(\Big(\ell(f^i,(x_i,y_i),(x_j,y_j)) - \ell(f^i - t\Delta f,(x_i,y_i),(x_j,y_j))\Big) \\
&\qquad\qquad + \Big(\ell(f^i - t\Delta f,(x'_i,y'_i),(x_j,y_j)) - \ell(f^i,(x'_i,y'_i),(x_j,y_j))\Big)\Big) \\
&\leq \quad \frac{t\sigma}{\binom{m}{2}} \sum_{j \neq i} \Big(|\Delta f(x_i)| + 2|\Delta f(x_j)| + |\Delta f(x'_i)|\Big),
\end{aligned}$$

where the last inequality follows by σ-admissibility. The result follows. ■

Appendix C. Proof of Lemma 17

The proof is a simple application of McDiarmid's inequality.
Proof *(of Lemma 17)*
Let $\phi : (\mathcal{X} \times \mathcal{Y})^m \to \mathbb{R}$ be defined as follows:

$$\phi(S) \quad = \quad \widehat{R}_\ell(f;S)\,.$$

Then by linearity of expectation,

$$\begin{aligned}\mathbf{E}_{S\sim\mathcal{D}^m}\Big[\phi(S)\Big] &= \frac{1}{\binom{m}{2}}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\mathbf{E}_{((X_i,Y_i),(X_j,Y_j))\sim\mathcal{D}\times\mathcal{D}}\Big[\ell(f,(X_i,Y_i),(X_j,Y_j))\Big]\\ &= R_\ell(f)\,.\end{aligned}$$

We shall show that ϕ satisfies the conditions of McDiarmid's inequality (Theorem 20). Let $S = ((x_1,y_1),\ldots,(x_m,y_m)) \in (\mathcal{X}\times\mathcal{Y})^m$. Then for each $1 \le k \le m$, we have for any $(x'_k,y'_k) \in (\mathcal{X}\times\mathcal{Y})$:

$$\begin{aligned}\left|\phi(S)-\phi(S^k)\right| &= \left|\widehat{R}_\ell(f;S)-\widehat{R}_\ell(f;S^k)\right|\\ &\le \frac{1}{\binom{m}{2}}\sum_{i\neq k}\left|\ell(f,(x_i,y_i),(x_k,y_k))-\ell(f,(x_i,y_i),(x'_k,y'_k))\right|\\ &\le \frac{1}{\binom{m}{2}}(m-1)B\\ &= \frac{2B}{m}\,.\end{aligned}$$

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

$$\begin{aligned}\mathbf{P}_{S\sim\mathcal{D}^m}\left(\widehat{R}_\ell(f;S)-R_\ell(f)\ge\varepsilon\right) &\le e^{-2\varepsilon^2/m(\frac{2B}{m})^2}\\ &= e^{-m\varepsilon^2/2B^2}\,.\end{aligned}$$

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

Appendix D. Proof of Theorem 19

The proof of this result also makes use of McDiarmid's inequality and is similar to the proof of Theorem 6 in Appendix A, with two main differences. First, McDiarmid's inequality is first applied to obtain a bound conditioned on an edge set E with $|E| \ge m^s$; the bound on the probability that $|E| < m^s$ then allows us to obtain the unconditional bound. Second, the constants c_k in the application of McDiarmid's inequality are different for each k, and depend on the degrees of the corresponding vertices in the graph with edge set E; the sum $\sum_k c_k^2$ is then bounded using a bound on the sum of squares of degrees in a graph due to de Caen (1998).
Proof *(of Theorem 19)*
Let $m \in \mathbb{N}$, and fix any edge set $E_0 \subseteq \{(i,j) \mid 1 \le i < j \le m\}$ with $|E_0| \ge m^s$. Let $\phi_{E_0} : \mathcal{X}^m \to \mathbb{R}$ be defined as follows:

$$\phi_{E_0}(S) \quad = \quad R_\ell(f_T) - \widehat{R}_\ell(f_T;T)\,,$$

where $T = (S, E_0, \pi|_{(S,E_0)})$. We shall show that ϕ_{E_0} satisfies the conditions of McDiarmid's inequality (Theorem 20). Let $S = (x_1, \ldots, x_m) \in \mathcal{X}^m$. Then for each $1 \leq k \leq m$, we have for any $x'_k \in \mathcal{X}$:

$$\begin{aligned} \left|\phi_{E_0}(S) - \phi_{E_0}(S^k)\right| &= \left| \left(R_\ell(f_T) - \widehat{R}_\ell(f_T;T)\right) - \left(R_\ell(f_{T^k}) - \widehat{R}_\ell(f_{T^k};T^k)\right) \right| \\ &\leq \left|R_\ell(f_T) - R_\ell(f_{T^k})\right| + \left|\widehat{R}_\ell(f_T;T) - \widehat{R}_\ell(f_{T^k};T^k)\right|. \end{aligned}$$

Now,

$$\begin{aligned} \left|R_\ell(f_T) - R_\ell(f_{T^k})\right| &= \left|\mathbf{E}_{(X,X')}\left[\ell(f_T, X, X', \pi(X,X')) - \ell(f_{T^k}, X, X', \pi(X,X'))\right]\right| \\ &\leq \mathbf{E}_{(X,X')}\left[\left|\ell(f_T, X, X', \pi(X,X')) - \ell(f_{T^k}, X, X', \pi(X,X'))\right|\right] \\ &\leq \beta(m), \end{aligned}$$

and

$$\begin{aligned} &\left|\widehat{R}_\ell(f_T;T) - \widehat{R}_\ell(f_{T^k};T^k)\right| \\ &\leq \frac{1}{|E_0|} \sum_{(i,j)\in E_0, i\neq k, j\neq k} \left|\ell(f_T, x_i, x_j, \pi(x_i,x_j)) - \ell(f_{T^k}, x_i, x_j, \pi(x_i,x_j))\right| \\ &\quad + \frac{1}{|E_0|} \sum_{(i,k)\in E_0} \left|\ell(f_T, x_i, x_k, \pi(x_i,x_k)) - \ell(f_{T^k}, x_i, x'_k, \pi(x_i,x'_k))\right| \\ &\quad + \frac{1}{|E_0|} \sum_{(k,j)\in E_0} \left|\ell(f_T, x_k, x_j, \pi(x_k,x_j)) - \ell(f_{T^k}, x'_k, x_j, \pi(x'_k,x_j))\right| \\ &\leq \frac{1}{|E_0|} \sum_{(i,j)\in E_0, i\neq k, j\neq k} \beta(m) + \frac{1}{|E_0|} \sum_{(i,k)\in E_0} B + \frac{1}{|E_0|} \sum_{(k,j)\in E_0} B \\ &\leq \beta(m) + \frac{d_k}{|E_0|} B, \end{aligned}$$

where d_k is the degree of vertex k in the graph with edge set E_0. Thus we have

$$\left|\phi_{E_0}(S) - \phi_{E_0}(S^k)\right| \leq c_k,$$

where

$$c_k = 2\beta(m) + \frac{d_k}{|E_0|} B.$$

Now,

$$\sum_{k=1}^m d_k = 2|E_0|,$$

and using a bound on the sum of squares of degrees in a graph due to de Caen (1998), we have

$$\sum_{k=1}^m d_k^2 \leq |E_0| \left(\frac{2|E_0|}{m-1} + m - 2\right).$$

Therefore,

$$\begin{aligned}
\sum_{k=1}^{m} c_k^2 &= \sum_{k=1}^{m} \left(2\beta(m) + \frac{d_k}{|E_0|}B\right)^2 \\
&= 4m(\beta(m))^2 + \frac{4B\beta(m)}{|E_0|}\sum_{k=1}^{m} d_k + \frac{B^2}{|E_0|^2}\sum_{k=1}^{m} d_k^2 \\
&\leq 4m(\beta(m))^2 + 8B\beta(m) + \frac{2B^2}{m-1} + \frac{B^2(m-2)}{|E_0|} \\
&\leq \underbrace{4m(\beta(m))^2 + 8B\beta(m) + \frac{2B^2}{m-1} + \frac{B^2}{m^{s-1}}}_{\gamma(m)},
\end{aligned}$$

where the last inequality follows since $|E_0| \geq m^s$. Thus, applying McDiarmid's inequality to ϕ_{E_0}, we get for any $\varepsilon > 0$,

$$\mathbf{P}_S\left(\left(R_\ell(f_T) - \widehat{R}_\ell(f_T;T)\right) - \mathbf{E}_S\left[R_\ell(f_T) - \widehat{R}_\ell(f_T;T) \;\middle|\; E = E_0\right] \geq \varepsilon \;\middle|\; E = E_0\right) \leq e^{-2\varepsilon^2/\gamma(m)}.$$

Now, as in Theorem 6, we can show that

$$\mathbf{E}_S\left[R_\ell(f_T) - \widehat{R}_\ell(f_T;T) \;\middle|\; E = E_0\right] \leq 2\beta(m).$$

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

$$\mathbf{P}_S\left(R_\ell(f_T) - \widehat{R}_\ell(f_T;T) - 2\beta(m) \geq \varepsilon \;\middle|\; E = E_0\right) \leq e^{-2\varepsilon^2/\gamma(m)}.$$

Since the above bound holds for all E_0 with $|E_0| \geq m^s$, we have

$$\begin{aligned}
\mathbf{P}_{(S,E)}\left(R_\ell(f_T) - \widehat{R}_\ell(f_T;T) - 2\beta(m) \geq \varepsilon\right) &\leq \mathbf{P}_E(|E| < m^s) + e^{-2\varepsilon^2/\gamma(m)} \\
&\leq \delta_m + e^{-2\varepsilon^2/\gamma(m)}.
\end{aligned}$$

Thus for m large enough such that $\delta_m \leq \frac{\delta}{2}$, we get

$$\mathbf{P}_{(S,E)}\left(R_\ell(f_T) - \widehat{R}_\ell(f_T;T) - 2\beta(m) \geq \varepsilon\right) \leq \frac{\delta}{2} + e^{-2\varepsilon^2/\gamma(m)}.$$

Setting the second term on the right hand side equal to $\frac{\delta}{2}$ and solving for ε gives the desired result. ■

References

Shivani Agarwal. Ranking on graph data. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Shivani Agarwal and Partha Niyogi. Stability and generalization of bipartite ranking algorithms. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.

Shivani Agarwal, Thore Graepel, Ralf Herbrich, Sariel Har-Peled, and Dan Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

Nir Ailon and Mehryar Mohri. An efficient reduction of ranking to classification. In *Proceedings of the 21st Annual Conference on Learning Theory*, 2008.

Kenneth J. Arrow. *Social Choice and Individual Values*. Yale University Press, Second edition, 1970.

Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory B. Sorkin. Robust reductions from ranking to classification. In *Proceedings of the 20th Annual Conference on Learning Theory*, 2007.

Olivier Bousquet. New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55(2):371–389, 2003.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

Alpha C. Chiang and Kevin Wainwright. *Fundamental Methods of Mathematical Economics*. McGraw-Hill Irwin, Fourth edition, 2005.

Stephan Clemencon and Nicolas Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

Stephan Clemencon, Gabor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *Annals of Statistics*, 36:844–874, 2008.

William W. Cohen, Robert E. Schapire, and Yoram Singer. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270, 1999.

Corinna Cortes, Mehryar Mohri, and Ashish Rastogi. Magnitude-preserving ranking algorithms. In *Proceedings of 24th International Conference on Machine Learning*, 2007.

David Cossock and Tong Zhang. Subset ranking using regression. In *Proceedings of the 19th Annual Conference on Learning Theory*, 2006.

Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in Neural Information Processing Systems 14*, pages 641–647. MIT Press, 2002.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2002.

Dominique de Caen. An upper bound on the sum of squares of degrees in a graph. *Discrete Mathematics*, 185:245–248, 1998.

Luc Devroye and Terry J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

David Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, University of California at Santa Cruz, 1999.

Ralf Herbrich, Thore Graepel, Peter Bollmann-Sdorra, and Klaus Obermayer. Learning preference relations for information retrieval. In *Proceedings of the ICML-1998 Workshop on Text Categorization and Machine Learning*, 1998.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2002.

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

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Samuel Kutin and Partha Niyogi. The interaction of stability and weakness in AdaBoost. Technical Report TR-2001-30, Computer Science Department, University of Chicago, 2001.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 2002.

Erich L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, 1975.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics 1989*, pages 148–188. Cambridge University Press, 1989.

Tomaso Poggio, Ryan Rifkin, Sayan Mukherjee, and Partha Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.

Filip Radlinski and Thorsten Joachims. Query chains: Learning to rank from implicit feedback. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, 2005.

Alain Rakotomamonjy. Optimizing area under ROC curves with SVMs. In *Proceedings of the ECAI-2004 Workshop on ROC Analysis in AI*, 2004.

William H. Rogers and Terry J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *Annals of Statistics*, 6(3):506–514, 1978.

Cynthia Rudin. Ranking with a *p*-norm push. In *Proceedings of the 19th Annual Conference on Learning Theory*, 2006.

Cynthia Rudin, Corinna Cortes, Mehryar Mohri, and Robert E. Schapire. Margin-based ranking meets boosting in the middle. In *Proceedings of the 18th Annual Conference on Learning Theory*, 2005.

Vladimir N. Vapnik and Alexey Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.