Generalization Bounds for the Area Under an ROC Curve

Shivani Agarwal*, Thore Graepel†, Ralf Herbrich†,
Sariel Har-Peled* and Dan Roth*

*Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA

†Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK

May 2004

Technical Report
UIUCDCS-R-2004-2433

Abstract

We study generalization properties of the area under an ROC curve (AUC), a quantity that has been advocated as an evaluation criterion for bipartite ranking problems. The AUC is a different and more complex term than the error rate used for evaluation in classification problems; consequently, existing generalization bounds for the classification error rate cannot be used to draw conclusions about the AUC. In this paper, we define a precise notion of the expected accuracy of a ranking function (analogous to the expected error rate of a classification function), and derive distribution-free probabilistic bounds on the deviation of the empirical AUC of a ranking function (observed on a finite data sequence) from its expected accuracy. We derive both a large deviation bound, which serves to bound the expected accuracy of a ranking function in terms of its empirical AUC on a test sequence, and a uniform convergence bound, which serves to bound the expected accuracy of a learned ranking function in terms of its empirical AUC on a training sequence. Our uniform convergence bound is expressed in terms of a new set of combinatorial parameters that we term the bipartite rank-shatter coefficients; these play the same role in our result as do the standard shatter coefficients (also known variously as the counting numbers or growth function) in uniform convergence results for the classification error rate. We also compare our result with a recent uniform convergence result derived by Freund et al. (2003) for a quantity closely related to the AUC; as we show, the bound provided by our result is considerably tighter.

1 Introduction

In many learning problems, the goal is not simply to classify objects into one of a fixed number of classes; instead, a *ranking* of objects is desired. This is the case, for example, in information retrieval problems, where one is interested in retrieving documents from some database that are 'relevant' to a given query or topic. In such problems, one wants to return to the user a list of documents that contains relevant documents at the top and irrelevant documents at the bottom; in other words, one wants a ranking of the documents such that relevant documents are ranked higher than irrelevant documents.

The problem of ranking has been studied from a learning perspective under a variety of settings (Cohen et al., 1999; Herbrich et al., 2000; Crammer and Singer, 2002; Freund et al., 2003). Here we consider the setting in which objects belong to one of two categories, positive and negative; the learner is given examples of objects labeled as positive or negative, and the goal is to learn a ranking in which positive objects are ranked higher than negative objects. This captures, for example, the information retrieval problem described above; in this case, the training examples given to the learner consist of documents labeled as relevant (positive) or irrelevant (negative). This form of ranking problem corresponds to the 'bipartite feedback' case of Freund et al. (2003); for this reason, we refer to it as the *bipartite* ranking problem.

Formally, the setting of a bipartite ranking problem is similar to that of a binary classification problem. There is an instance space $\mathcal{X}$ from which instances are drawn, and a set of two class labels $\mathcal{Y}$ which we take without loss of generality to be $\mathcal{Y} = \{-1, +1\}$. In both problems, one is given a finite sequence of labeled training examples $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_M, y_M)) \in (\mathcal{X} \times \mathcal{Y})^M$, and the goal is to learn a function based on this training sequence. However, the form of the function to be learned in the two problems is different. In classification, one seeks a binary-valued function $h : \mathcal{X} \to \mathcal{Y}$ that predicts the class of a new instance in $\mathcal{X}$. On the other hand, in ranking, one seeks a *real-valued* function $f : \mathcal{X} \to \mathbb{R}$ that induces a ranking over $\mathcal{X}$; an instance that is assigned a higher value by f is ranked higher than one that is assigned a lower value by f.

What is a good classification or ranking function? Intuitively, a good classification function should classify most instances correctly, while a good ranking function should rank most instances labeled as positive higher than most instances labeled as negative. At first thought, these intuitions might suggest that one problem could be reduced to the other; that a good solution to one could be used to obtain a good solution to the other. Indeed, several approaches to learning ranking functions have involved using a standard classification algorithm that produces a classification function h of the form[1] $h(\mathbf{x}) = \text{sign}(f_h(\mathbf{x}))$ for some real-valued function $f_h : \mathcal{X} \to \mathbb{R}$, and then taking f_h to be the desired ranking function.[2] However, despite the apparently close relation between classification and ranking, on formalizing the above intuitions about evaluation criteria for classification and ranking functions, it turns out that a good classification function may not always translate into a good ranking function.

1.1 Evaluation of (Binary) Classification Functions

In classification, one generally assumes that examples (both training examples and future, unseen examples) are drawn randomly and independently according to some (unknown) underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The mathematical quantity typically used to evaluate a classification function

[1] For $z \in \mathbb{R}$, $\text{sign}(z) = \begin{cases} 1 & \text{if } z > 0 \\ -1 & \text{otherwise.} \end{cases}$

[2] In Herbrich et al. (2000) the problem of learning a ranking function is also reduced to a classification problem, but on *pairs* of instances.

$h : \mathcal{X} \rightarrow \mathcal{Y}$ is then the *expected error rate* (or simply *error rate*) of h, denoted by $L(h)$ and defined as

$$L(h) \quad = \quad \mathbf{E}_{XY \sim \mathcal{D}} \left\{ \mathbf{I}_{\{h(X) \neq Y\}} \right\} , \tag{1}$$

where $\mathbf{I}_{\{.\}}$ denotes the indicator variable whose value is one if its argument is true and zero otherwise. The error rate $L(h)$ is simply the probability that an example drawn randomly from $\mathcal{X} \times \mathcal{Y}$ (according to $\mathcal{D}$) will be misclassified by h; the quantity $(1 - L(h))$ thus measures our intuitive notion of 'how often instances are classified correctly by h'. In practice, since the distribution $\mathcal{D}$ is not known, the true error rate of a classification function cannot be computed exactly. Instead, the error rate must be estimated using a finite data sample. A widely used estimate is the *empirical error rate*: given a finite sequence of labeled examples $T = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$, the empirical error rate of a classification function h with respect to T, which we denote by $\hat{L}(h; T)$, is given by

$$\hat{L}(h; T) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{h(\mathbf{x}_i) \neq y_i\}} . \tag{2}$$

When the examples in T are drawn randomly and independently from $\mathcal{X} \times \mathcal{Y}$ according to $\mathcal{D}$, the sequence T constitutes a random sample. Much work in learning theory research has concentrated on developing bounds on the probability that an error estimate obtained from such a random sample will have a large deviation from the true error rate. While the true error rate of a classification function may not be exactly computable, such generalization bounds allow us to compute confidence intervals within which the true value of the error rate is likely to be contained with high probability.

1.2 Evaluation of (Bipartite) Ranking Functions

Evaluating a ranking function has proved to be somewhat more difficult. One empirical quantity that has been used for this purpose is the average precision, which relates to recall-precision curves. The average precision is often used in applications that contain very few positive examples, such as information retrieval. Another empirical quantity that has recently gained some attention as being well-suited for evaluating ranking functions relates to receiver operating characteristic (ROC) curves. ROC curves were originally developed in signal detection theory for analysis of radar images (Egan, 1975), and have been used extensively in various fields such as medical decision-making. Given a ranking function $f : \mathcal{X} \rightarrow \mathbb{R}$ and a finite data sequence $T = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$, the ROC curve of f with respect to T is obtained as follows. First, a set of $N + 1$ classification functions $h_i : \mathcal{X} \rightarrow \mathcal{Y}$, where $0 \leq i \leq N$, is constructed from f:

$$h_i(\mathbf{x}) \quad = \quad \text{sign}(f(\mathbf{x}) - b_i) , \tag{3}$$

where

$$b_i \quad = \quad \begin{cases} f(\mathbf{x}_i) & \text{if } 1 \leq i \leq N \\ \left(\min\limits_{1 \leq j \leq N} f(\mathbf{x}_j) \right) - 1 & \text{if } i = 0 . \end{cases} \tag{4}$$

The classification function h_0 classifies all instances in T as positive, while for $1 \leq i \leq N$, h_i classifies all instances ranked higher than $\mathbf{x}_i$ as positive, and all others (including $\mathbf{x}_i$) as negative. Next, for

each classification function h_i, one computes the (empirical) true positive and false positive rates on T, denoted by tpr_i and fpr_i respectively:

$$tpr_i \;=\; \frac{\textit{number of positive examples in } T \textit{ classified correctly by } h_i}{\textit{total number of positive examples in } T}\,, \tag{5}$$

$$fpr_i \;=\; \frac{\textit{number of negative examples in } T \textit{ misclassified as positive by } h_i}{\textit{total number of negative examples in } T}\,. \tag{6}$$

Finally, the points (fpr_i, tpr_i) are plotted on a graph with the false positive rate on the x-axis and the true positive rate on the y-axis; the ROC curve is then obtained by connecting these points such that the resulting curve is monotonically increasing. It is the *area under the ROC curve* (AUC) that has been used as an indicator of the quality of the ranking function f (Yan et al., 2003; Cortes and Mohri, 2004). An AUC value of one corresponds to a perfect ranking on the given data sequence (*i.e.*, all positive instances in T are ranked higher than all negative instances); a value of zero corresponds to the opposite scenario (*i.e.*, all negative instances in T are ranked higher than all positive instances).

The AUC can in fact be expressed in a simpler form: if the sample T contains m positive and n negative examples, then it is not difficult to see that the AUC of f with respect to T, which we denote by $\hat{A}(f;T)$, is given simply by the following Wilcoxon-Mann-Whitney statistic (Yan et al., 2003; Cortes and Mohri, 2004):

$$\hat{A}(f;T) \;=\; \frac{1}{mn} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(\mathbf{x}_i)>f(\mathbf{x}_j)\}}\,. \tag{7}$$

In this simplified form, it becomes clear that the AUC of f with respect to T is simply the fraction of positive-negative pairs in T that are ranked correctly by f.

There are two important observations to be made about the AUC defined above. The first is that the error rate of a classification function is not necessarily a good indicator of the AUC of a ranking function derived from it; different classification functions with the same error rate may produce ranking functions with very different AUC values. For example, consider two classification functions h_1, h_2 given by $h_i(\mathbf{x}) = \text{sign}(f_i(\mathbf{x})), i = 1, 2$, where the values assigned by f_1, f_2 to the instances in a sample $T \in (\mathcal{X} \times \mathcal{Y})^8$ are as shown in Table 1. Clearly, $\hat{L}(h_1;T) = \hat{L}(h_2;T) = 2/8$, but $\hat{A}(f_1;T) = 12/16$ while $\hat{A}(f_2;T) = 8/16$. The exact relationship between the (empirical) error rate of a classification function h of the form $h(\mathbf{x}) = \text{sign}(f_h(\mathbf{x}))$ and the AUC value of the corresponding ranking function f_h with respect to a given data sequence was studied in detail in (Yan et al., 2003; Cortes and Mohri, 2004). In particular, it was shown in (Cortes and Mohri, 2004) that when the number of positive examples m in the given data sequence is equal to the number of negative examples n, the average AUC value over all possible rankings corresponding to classification functions with a fixed (empirical) error rate ℓ is given by $(1 - \ell)$, but the standard deviation among the AUC values can be large for large ℓ. As the proportion of positive instances $m/(m+n)$ departs from $1/2$, the average AUC value corresponding to an error rate ℓ departs from $(1 - \ell)$, and the standard deviation increases further. The AUC is thus a different term than the error rate, and therefore requires separate analysis.

The second important observation about the AUC is that, as defined above, it is an empirical quantity that evaluates a ranking function with respect to a particular data sequence. What does the empirical AUC tell us about the expected performance of a ranking function on future examples? This is the question we address in this paper. The question has two parts, both of which

Table 1: Values assigned by two functions f_1, f_2 to eight instances in a hypothetical example. The corresponding classification functions have the same (empirical) error rate, but the AUC values of the ranking functions are different. See text for details.

$\mathbf{x}_i$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$
y_i	-1	-1	-1	-1	+1	+1	+1	+1
$f_1(\mathbf{x}_i)$	-2	-1	3	4	1	2	5	6
$f_2(\mathbf{x}_i)$	-2	-1	5	6	1	2	3	4

are important for machine learning. First, what can be said about the expected performance of a ranking function based on its empirical AUC on an independent test sequence? Second, what can be said about the expected performance of a learned ranking function based on its empirical AUC on the training sequence from which it is learned? The first is a question about the large deviation behaviour of the AUC; the second is a question about its uniform convergence behaviour. Both are addressed in this paper.

We start by defining a precise notion of the expected ranking accuracy of a ranking function (analogous to the expected error rate of a classification function) in Section 2. Section 3 contains our large deviation result, which serves to bound the expected accuracy of a ranking function in terms of its empirical AUC on an independent test sequence. Our conceptual approach in deriving the large deviation result for the AUC is similar to that of (Hill et al., 2002), in which large deviation properties of the average precision were considered. Section 4 contains our uniform convergence result, which serves to bound the expected accuracy of a learned ranking function in terms of its empirical AUC on a training sequence. Our uniform convergence bound is expressed in terms of a new set of combinatorial parameters that we term the bipartite rank-shatter coefficients; these play the same role in our result as do the standard shatter coefficients (also known variously as the counting numbers or growth function) in uniform convergence results for the classification error rate. We also offer in Section 4 a comparison of our result with a recent uniform convergence result derived by Freund et al. (2003) for a quantity closely related to the AUC; as we show, the bound provided by our result is considerably tighter. Finally, we conclude with a discussion in Section 5.

2 Expected Ranking Accuracy

We begin by introducing some additional notation. As in classification, we shall assume that all examples are drawn randomly and independently according to some (unknown) underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The notation $\mathcal{D}_{+1}$ and $\mathcal{D}_{-1}$ will be used to denote the class-conditional distributions $\mathcal{D}_{X|Y=+1}$ and $\mathcal{D}_{X|Y=-1}$, respectively. We shall find it convenient to decompose a data sequence $T = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ into two components, $T_X = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathcal{X}^N$ and $T_Y = (y_1, \ldots, y_N) \in \mathcal{Y}^N$. Several of our results will involve the conditional distribution $\mathcal{D}_{T_X|T_Y=\underline{y}}$ for some label sequence $\underline{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$; this distribution is simply $\mathcal{D}_{y_1} \times \ldots \times \mathcal{D}_{y_N}$.[3] If the distribution is clear from the context it will be dropped in the notation of expectations and

[3]Note that, since the AUC of a ranking function f with respect to a data sequence $T \in (\mathcal{X} \times \mathcal{Y})^N$ is independent of the actual ordering of examples in the sequence, our results involving the conditional distribution $\mathcal{D}_{T_X|T_Y=\underline{y}}$ for some label sequence $\underline{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ depend only on the number m of $+1$ labels in $\underline{y}$ and the number n of -1 labels in $\underline{y}$. We choose to state our results in terms of the distribution $\mathcal{D}_{T_X|T_Y=\underline{y}} \equiv \mathcal{D}_{y_1} \times \ldots \times \mathcal{D}_{y_N}$ only because this is more general than stating them in terms of $\mathcal{D}_{+1}^m \times \mathcal{D}_{-1}^n$.

probabilities, *e.g.*, $\mathbf{E}_{XY} \equiv \mathbf{E}_{XY\sim\mathcal{D}}$.

We define below a quantity that we term the expected ranking accuracy; the purpose of this quantity will be to serve as an evaluation criterion for ranking functions (analogous to the use of the expected error rate as an evaluation criterion for classification functions).

Definition 1 (Expected ranking accuracy). *Let $f : \mathcal{X}\to\mathbb{R}$ be a ranking function on $\mathcal{X}$. Define the* expected ranking accuracy *(or simply* ranking accuracy*) of f, denoted by $A(f)$, as follows:*

$$A(f) \;=\; \mathbf{E}_{X\sim\mathcal{D}_{+1},X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X)>f(X')\}}\right\}. \tag{8}$$

The ranking accuracy $A(f)$ defined above is simply the probability that an instance drawn randomly according to $\mathcal{D}_{+1}$ will be ranked higher by f than an instance drawn randomly according to $\mathcal{D}_{-1}$; $A(f)$ thus measures our intuitive notion of 'how often instances labeled as positive are ranked higher by f than instances labeled as negative'. As in the case of classification, the true ranking accuracy depends on the underlying distribution of the data and cannot be observed directly. Our goal shall be to derive generalization bounds that allow the true ranking accuracy of a ranking function to be estimated from its empirical AUC with respect to a finite data sample. The following simple lemma shows that this makes sense, for given a fixed label sequence, the empirical AUC of a ranking function f is an unbiased estimator of the expected ranking accuracy of f:

Lemma 1. *Let $f : \mathcal{X}\to\mathbb{R}$ be a ranking function on $\mathcal{X}$, and let $\underline{y} = (y_1,\ldots,y_N) \in \mathcal{Y}^N$ be a finite label sequence. Then*

$$\mathbf{E}_{T_X|T_Y=\underline{y}}\left\{\hat{A}(f;T)\right\} \;=\; A(f).$$

Proof. Let m be the number of $+1$ labels in $\underline{y}$, and n the number of -1 labels in $\underline{y}$. Then from the definition of empirical AUC (Eq. (7)) and linearity of expectation, we have

$$\begin{aligned}
\mathbf{E}_{T_X|T_Y=\underline{y}}\left\{\hat{A}(f;T)\right\} &= \frac{1}{mn}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\mathbf{E}_{X_i\sim\mathcal{D}_{+1},X_j\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X_i)>f(X_j)\}}\right\}\\
&= \frac{1}{mn}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}A(f)\\
&= A(f).
\end{aligned}$$

□

We are now ready to present the main results of this paper, namely, a large deviation bound in Section 3 and a uniform convergence bound in Section 4. We note that our results are all distribution-free, in the sense that they hold for any distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$.

3 Large Deviation Bound for the AUC

In this section we are interested in bounding the probability that the empirical AUC of a ranking function f with respect to a random test sequence T will have a large deviation from its expected ranking accuracy. In other words, we are interested in bounding probabilities of the form

$$\mathbf{P}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \epsilon\right\}$$

for given $\epsilon > 0$. Our main tool in deriving such a large deviation bound 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 1 (McDiarmid, 1989). *Let $X_1, \ldots, X_N$ be independent random variables with X_k taking values in a set A_k for each k. Let $\phi : (A_1 \times \cdots \times A_N) \to \mathbb{R}$ be such that*

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

Then for any $\epsilon > 0$,

$$\mathbf{P}\left\{ |\phi(X_1, \ldots, X_N) - \mathbf{E}\{\phi(X_1, \ldots, X_N)\}| \geq \epsilon \right\} \quad \leq \quad 2e^{-2\epsilon^2 / \sum_{k=1}^N c_k^2} \, .$$

Before giving our bound, we define the following quantity which appears in several of the results in this section:

Definition 2 (Positive skew). *Let $\underline{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ be a finite label sequence of length $N \in \mathbb{N}$. Define the* positive skew *of $\underline{y}$, denoted by $\rho(\underline{y})$, as follows:*

$$\rho(\underline{y}) \quad = \quad \frac{1}{N} \sum_{\{i : y_i = +1\}} 1 \, . \tag{9}$$

The following is the main result of this section:

Theorem 2. *Let $f : \mathcal{X} \to \mathbb{R}$ be a fixed ranking function on $\mathcal{X}$ and let $\underline{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ be any label sequence of length $N \in \mathbb{N}$. Then for any $\epsilon > 0$,*

$$\mathbf{P}_{T_X | T_Y = \underline{y}} \left\{ \left| \hat{A}(f; T) - A(f) \right| \geq \epsilon \right\} \quad \leq \quad 2e^{-2\rho(\underline{y})(1-\rho(\underline{y}))N\epsilon^2} \, .$$

Proof. Let m be the number of $+1$ labels in $\underline{y}$, and n the number of -1 labels in $\underline{y}$. We can view $T_X = (X_1, \ldots, X_N) \in \mathcal{X}^N$ as a random vector; given the label sequence $\underline{y}$, the random variables $X_1, \ldots, X_N$ are independent, with each X_k taking values in $\mathcal{X}$. Now, define $\phi : \mathcal{X}^N \to \mathbb{R}$ as follows:

$$\phi\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) \quad = \quad \hat{A}\left(f; ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N))\right) \, .$$

Then, for each k such that $y_k = +1$, we have the following for all $\mathbf{x}_i, \mathbf{x}_k' \in \mathcal{X}$:

$$\begin{aligned}
&\left| \phi(\mathbf{x}_1, \ldots, \mathbf{x}_N) - \phi(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{x}_k', \mathbf{x}_{k+1} \ldots, \mathbf{x}_N) \right| \\
&\quad = \left| \frac{1}{mn} \sum_{\{i : y_i = +1\}} \sum_{\{j : y_j = -1\}} \mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}_j)\}} \right. \\
&\qquad\qquad \left. - \frac{1}{mn} \left(\sum_{\{i : y_i = +1, i \neq k\}} \sum_{\{j : y_j = -1\}} \mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}_j)\}} + \sum_{\{j : y_j = -1\}} \mathbf{I}_{\{f(\mathbf{x}_k') > f(\mathbf{x}_j)\}} \right) \right| \\
&\quad = \frac{1}{mn} \left| \sum_{\{j : y_j = -1\}} \left(\mathbf{I}_{\{f(\mathbf{x}_k) > f(\mathbf{x}_j)\}} - \mathbf{I}_{\{f(\mathbf{x}_k') > f(\mathbf{x}_j)\}} \right) \right| \\
&\quad \leq \frac{1}{mn} n \\
&\quad = \frac{1}{m} \, .
\end{aligned}$$

Similarly, for each k such that $y_k = -1$, one can show for all $\mathbf{x}_i, \mathbf{x}'_k \in \mathcal{X}$:

$$\left|\phi(\mathbf{x}_1, \ldots, \mathbf{x}_N) - \phi(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{x}'_k, \mathbf{x}_{k+1} \ldots, \mathbf{x}_N)\right| \quad \leq \quad \frac{1}{n}\,.$$

Thus, taking $c_k = 1/m$ for k such that $y_k = +1$ and $c_k = 1/n$ for k such that $y_k = -1$, and applying McDiarmid's theorem, we get for any $\epsilon > 0$,

$$\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\left|\hat{A}(f;T) - \mathbf{E}_{T_X|T_Y=\underline{y}}\left\{\hat{A}(f;T)\right\}\right| \geq \epsilon\right\} \quad \leq \quad 2e^{-2\epsilon^2/(m(\frac{1}{m})^2 + n(\frac{1}{n})^2)}\,. \tag{10}$$

Now, from Lemma 1,

$$\mathbf{E}_{T_X|T_Y=\underline{y}}\left\{\hat{A}(f;T)\right\} \quad = \quad A(f)\,.$$

Also, we have

$$\frac{1}{m(\frac{1}{m})^2 + n(\frac{1}{n})^2} = \frac{1}{\frac{1}{m} + \frac{1}{n}} = \frac{mn}{m+n} = \rho(\underline{y})(1-\rho(\underline{y}))N\,.$$

Substituting the above in Eq. (10) gives the desired result. □

From Theorem 2, we can derive a confidence interval interpretation of the bound that gives, for any $0 < \delta \leq 1$, a confidence interval based on the empirical AUC of a ranking function (on a random test sequence) which is likely to contain the true ranking accuracy with probability at least $1 - \delta$. More specifically, we have:

Corollary 1. *Let $f : \mathcal{X} \to \mathbb{R}$ be a fixed ranking function on $\mathcal{X}$ and let $\underline{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ be any label sequence of length $N \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2\rho(\underline{y})(1-\rho(\underline{y}))N}}\right\} \quad \leq \quad \delta\,.$$

Proof. This follows directly from Theorem 2 by setting $2e^{-2\rho(\underline{y})(1-\rho(\underline{y}))N\epsilon^2} = \delta$ and solving for ϵ. □

The above result can in fact be generalized in the following manner:

Corollary 2. *Let $f : \mathcal{X} \to \mathbb{R}$ be a fixed ranking function on $\mathcal{X}$ and let $N \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{T\sim\mathcal{D}^N}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2\rho(T_Y)(1-\rho(T_Y))N}}\right\} \quad \leq \quad \delta\,.$$

Proof. For $T \in (\mathcal{X} \times \mathcal{Y})^N$ and $0 < \delta \leq 1$, define the proposition

$$\Phi(T,\delta) \quad \equiv \quad \left\{\left|\hat{A}(f;T) - A(f)\right| \geq \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2\rho(T_Y)(1-\rho(T_Y))N}}\right\}.$$

Then for any $0 < \delta \leq 1$, we have

$$\begin{aligned}
\mathbf{P}_T\left\{\Phi(T,\delta)\right\} &= \mathbf{E}_T\left\{\mathbf{I}_{\Phi(T,\delta)}\right\} \\
&= \mathbf{E}_{T_Y}\left\{\mathbf{E}_{T_X|T_Y=\underline{y}}\left\{\mathbf{I}_{\Phi(T,\delta)}\right\}\right\} \\
&= \mathbf{E}_{T_Y}\left\{\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\Phi(T,\delta)\right\}\right\} \\
&\leq \mathbf{E}_{T_Y}\left\{\delta\right\} \qquad \text{(by Corollary 1)} \\
&= \delta\,.
\end{aligned}$$

□

Note that the above 'trick' works only once we have gone to a confidence interval; an attempt to generalize the bound of Theorem 2 in a similar way gives an expression in which the final expectation is not easy to evaluate. Interestingly, the above proof does not even require a factorized distribution $\mathcal{D}_{T_Y}$ since it is built on a result for any fixed label sequence $\underline{y}$. We note that the above technique could also be applied to generalize the results of Hill et al. (2002) in a similar manner.

Theorem 2 also allows us to obtain an expression for a test sample size that is sufficient to obtain, for given $0 < \epsilon, \delta \leq 1$, an ϵ-accurate estimate of the ranking accuracy with δ-confidence:

Corollary 3. *Let $f : \mathcal{X} \to \mathbb{R}$ be a fixed ranking function on $\mathcal{X}$ and let $0 < \epsilon, \delta \leq 1$. Let $\underline{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ be any label sequence of length $N \in \mathbb{N}$. If*

$$N \quad \geq \quad \frac{\ln\left(\frac{2}{\delta}\right)}{2\rho(\underline{y})(1-\rho(\underline{y}))\epsilon^2},$$

then

$$\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \epsilon\right\} \quad \leq \quad \delta\,.$$

Proof. This follows directly from Theorem 2 by setting $2e^{-2\rho(\underline{y})(1-\rho(\underline{y}))N\epsilon^2} \leq \delta$ and solving for N. □

Figure 1 illustrates the dependence of the above expression for the sufficient test sample size on the confidence parameter δ, the accuracy parameter ϵ and the positive skew $\rho(\underline{y})$.

3.1 Comparison with Large Deviation Bound for Classification Error Rate

Our use of McDiarmid's inequality in deriving the large deviation bound for the AUC of a ranking function is analogous to the use of Hoeffding's inequality in deriving large deviation bounds for the error rate of a classification function (see, for example, Devroye et al., 1996, chapter 8). The need for the more general inequality of McDiarmid in our derivations arises from the fact that the empirical AUC is a more complex term which, unlike the empirical error rate, cannot be expressed as a sum of independent random variables. In the notation of Section 1, the large deviation bound obtained via Hoeffding's inequality for the classification error rate states that for a fixed classification function $h : \mathcal{X} \to \mathcal{Y}$ and for any $N \in \mathbb{N}$ and any $\epsilon > 0$,

$$\mathbf{P}_{T\sim\mathcal{D}^N}\left\{\left|\hat{L}(h;T) - L(h)\right| \geq \epsilon\right\} \quad \leq \quad 2e^{-2N\epsilon^2}\,. \tag{11}$$

Comparing Eq. (11) to the bound of Theorem 2, we see that the AUC bound differs from the error rate bound by a factor of $\rho(\underline{y})(1-\rho(\underline{y}))$ in the exponent. This difference translates into a $1/(\rho(\underline{y})(1-\rho(\underline{y})))$ factor difference in the resulting sample size bounds; in other words, for given $0 < \epsilon, \delta \leq 1$, the test sample size sufficient to obtain an ϵ-accurate estimate of the expected accuracy of a ranking function with δ-confidence is $1/(\rho(\underline{y})(1-\rho(\underline{y})))$ times larger than the corresponding test sample size sufficient to obtain an ϵ-accurate estimate of the expected error rate of a classification function with the same confidence. For $\rho(\underline{y}) = 1/2$, this means a sample size larger by a factor of 4; as the positive skew $\rho(\underline{y})$ departs from $1/2$, the factor grows larger (see Figure 2).

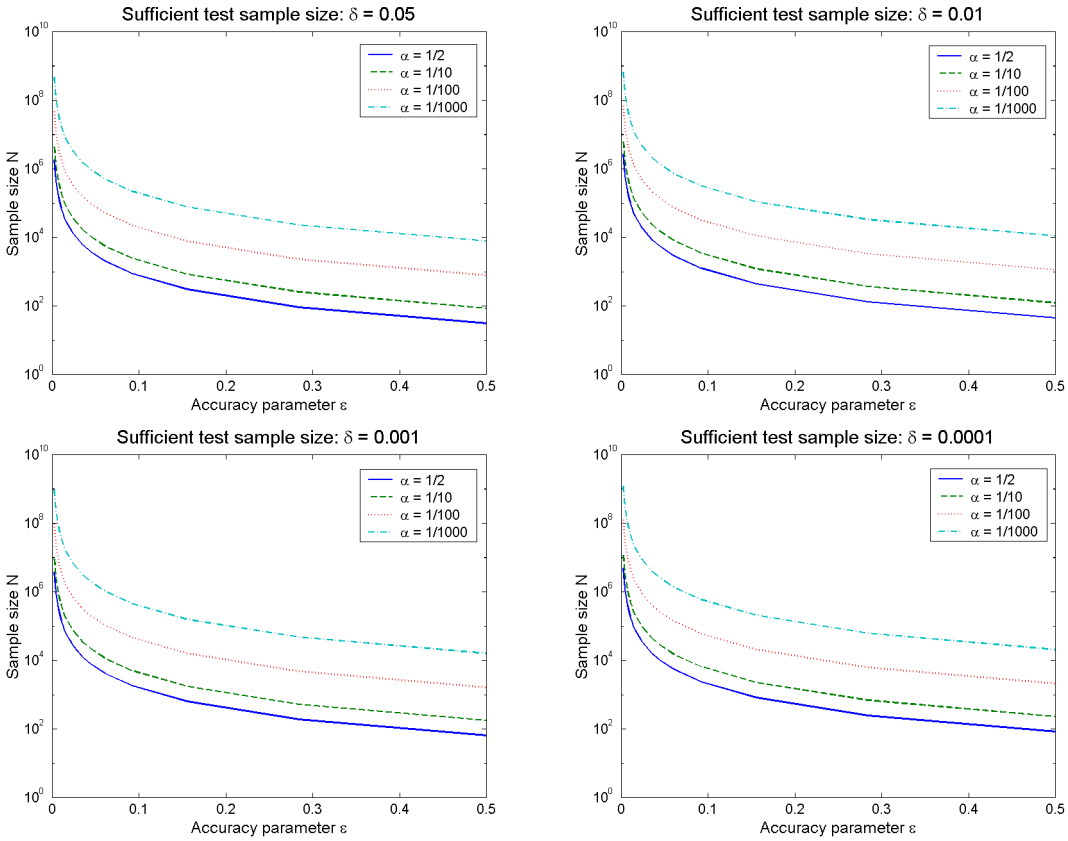

Figure 1: The test sample size N (based on Corollary 3) sufficient to obtain an ϵ-accurate estimate of the ranking accuracy with δ-confidence, for various values of the positive skew $\alpha \equiv \rho(\underline{y})$ for some label sequence $\underline{y}$, when (top-left) $\delta = 0.05$, (top-right) $\delta = 0.01$, (bottom-left) $\delta = 0.001$, and (bottom-right) $\delta = 0.0001$.

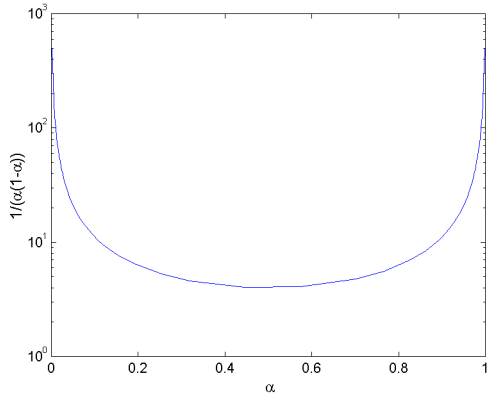

Figure 2: The test sample size bound for the AUC, for positive skew $\alpha \equiv \rho(\underline{y})$ for some label sequence $\underline{y}$, is larger than the corresponding test sample size bound for the error rate by a factor of $1/(\alpha(1-\alpha))$. (See text for discussion.)

3.2 Bound for Learned Ranking Functions Chosen from Finite Function Classes

The large deviation result derived in Theorem 2 bounds the expected accuracy of a ranking function in terms of its empirical AUC on an independent test sequence. A simple application of the union bound allows the result to be extended to bound the expected accuracy of a learned ranking function in terms of its empirical AUC on the training sequence from which it is learned, in the case when the learned ranking function is chosen from a finite function class. More specifically, we have:

Theorem 3. *Let $\mathcal{F}$ be a finite class of real-valued functions on $\mathcal{X}$ and let $f_S \in \mathcal{F}$ denote the ranking function chosen by a learning algorithm based on the training sequence S. Let $\underline{y} = (y_1, \ldots, y_M) \in \mathcal{Y}^M$ be any label sequence of length $M \in \mathbb{N}$. Then for any $\epsilon > 0$,*

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S; S) - A(f_S) \right| \geq \epsilon \right\} \quad \leq \quad 2|\mathcal{F}| e^{-2\rho(\underline{y})(1-\rho(\underline{y}))M\epsilon^2} .$$

Proof. For any $\epsilon > 0$, we have

$$\begin{aligned}
&\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S; S) - A(f_S) \right| \geq \epsilon \right\} \\
&\qquad \leq \mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \max_{f \in \mathcal{F}} \left| \hat{A}(f; S) - A(f) \right| \geq \epsilon \right\} \\
&\qquad \leq \sum_{f \in \mathcal{F}} \mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \left| \hat{A}(f; S) - A(f) \right| \geq \epsilon \right\} \qquad \text{(by the union bound)} \\
&\qquad \leq 2|\mathcal{F}| e^{-2\rho(\underline{y})(1-\rho(\underline{y}))M\epsilon^2} \qquad \text{(by Theorem 2)} .
\end{aligned}$$

□

As before, we can derive from Theorem 3 expressions for confidence intervals and sufficient training sample size; we give these below without proof:

Corollary 4. *Let $\mathcal{F}$ be a finite class of real-valued functions on $\mathcal{X}$ and let $f_S \in \mathcal{F}$ denote the ranking function chosen by a learning algorithm based on the training sequence S. Let $\underline{y} = (y_1, \ldots, y_M) \in \mathcal{Y}^M$ be any label sequence of length $M \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S; S) - A(f_S) \right| \geq \sqrt{\frac{\ln|\mathcal{F}| + \ln\left(\frac{2}{\delta}\right)}{2\rho(\underline{y})(1-\rho(\underline{y}))M}} \right\} \quad \leq \quad \delta .$$

Corollary 5. *Let $\mathcal{F}$ be a finite class of real-valued functions on $\mathcal{X}$ and let $f_S \in \mathcal{F}$ denote the ranking function chosen by a learning algorithm based on the training sequence S. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S \sim \mathcal{D}^M} \left\{ \left| \hat{A}(f_S; S) - A(f_S) \right| \geq \sqrt{\frac{\ln|\mathcal{F}| + \ln\left(\frac{2}{\delta}\right)}{2\rho(S_Y)(1-\rho(S_Y))M}} \right\} \quad \leq \quad \delta .$$

Corollary 6. *Let $\mathcal{F}$ be a finite class of real-valued functions on $\mathcal{X}$ and let $f_S \in \mathcal{F}$ denote the ranking function chosen by a learning algorithm based on the training sequence S. Let $\underline{y} = (y_1, \ldots, y_M) \in \mathcal{Y}^M$ be any label sequence of length $M \in \mathbb{N}$. Then for any $0 < \epsilon, \delta \leq 1$, if*

$$M \quad \geq \quad \frac{1}{2\rho(\underline{y})(1-\rho(\underline{y}))\epsilon^2} \left(\ln|\mathcal{F}| + \ln\left(\frac{2}{\delta}\right) \right),$$

then

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S; S) - A(f_S) \right| \geq \epsilon \right\} \quad \leq \quad \delta .$$

The above results apply only to ranking functions learned from finite function classes. The general case, when the learned ranking function may be chosen from a possibly infinite function class, is the subject of the next section.

4 Uniform Convergence Bound for the AUC

In this section we are interested in bounding the probability that the empirical AUC of a learned ranking function f_S with respect to the (random) training sequence S from which it is learned will have a large deviation from its expected ranking accuracy, when the function f_S is chosen from a possibly infinite function class $\mathcal{F}$. The standard approach for obtaining such bounds is via uniform convergence results. In particular, we have for any $\epsilon > 0$,

$$\mathbf{P}\left\{\left|\hat{A}(f_S;S) - A(f_S)\right| \geq \epsilon\right\} \quad \leq \quad \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \geq \epsilon\right\}. \tag{12}$$

Therefore, to bound probabilities of the form on the left hand side of Eq. (12), it is sufficient to derive a uniform convergence result that bounds probabilities of the form on the right hand side. Our uniform convergence result for the AUC is expressed in terms of a new set of combinatorial parameters, termed the bipartite rank-shatter coefficients, that we define below.

4.1 Bipartite Rank-Shatter Coefficients

We define first the notion of a bipartite rank matrix; this is used in our definition of bipartite rank-shatter coefficients.

Definition 3 (Bipartite rank matrix). *Let $f : \mathcal{X}\to\mathbb{R}$ be a ranking function on $\mathcal{X}$, let $m, n \in \mathbb{N}$, and let $\underline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \in \mathcal{X}^m$, $\underline{\mathbf{x}}' = (\mathbf{x}'_1, \ldots, \mathbf{x}'_n) \in \mathcal{X}^n$. Define the* bipartite rank matrix *of f with respect to $\underline{\mathbf{x}}, \underline{\mathbf{x}}'$, denoted by $\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')$, to be the matrix in $\{0,1\}^{m\times n}$ whose $(i,j)^{\text{th}}$ element is given by*

$$\left[\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')\right]_{ij} \quad = \quad \mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}'_j)\}}$$

for all $1 \leq i \leq m, 1 \leq j \leq n$.

Definition 4 (Bipartite rank-shatter coefficient). *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $m, n \in \mathbb{N}$. Define the $(m,n)^{\text{th}}$* bipartite rank-shatter coefficient *of $\mathcal{F}$, denoted by $r(\mathcal{F}, m, n)$, as follows:*

$$r(\mathcal{F}, m, n) \quad = \quad \max_{\underline{\mathbf{x}}\in\mathcal{X}^m, \underline{\mathbf{x}}'\in\mathcal{X}^n} \left|\left\{\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}') \mid f \in \mathcal{F}\right\}\right|.$$

Clearly, for finite $\mathcal{F}$, we have $r(\mathcal{F}, m, n) \leq |\mathcal{F}|$ for all m, n. In general, $r(\mathcal{F}, m, n) \leq 2^{mn}$ for all m, n. In fact, for $m, n \geq 2$, we have $r(\mathcal{F}, m, n) \leq g(m, n)$, where $g(m, n)$ is the number of matrices in $\{0,1\}^{m\times n}$ that do not contain a sub-matrix of the form

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

To see this, observe that for a bipartite rank matrix $\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')$ to contain a sub-matrix of the form $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ in rows i_1, i_2 and columns j_1, j_2, we must have

$$\begin{array}{ll} f(\mathbf{x}_{i_1}) > f(\mathbf{x}'_{j_1}) & f(\mathbf{x}_{i_1}) \leq f(\mathbf{x}'_{j_2}) \\ f(\mathbf{x}_{i_2}) \leq f(\mathbf{x}'_{j_1}) & f(\mathbf{x}_{i_2}) > f(\mathbf{x}'_{j_2}). \end{array}$$

However, this gives

$$f(\mathbf{x}_{i_1}) > f(\mathbf{x}'_{j_1}) \geq f(\mathbf{x}_{i_2}) > f(\mathbf{x}'_{j_2}) \geq f(\mathbf{x}_{i_1}) \,,$$

which is a contradiction. Therefore, a bipartite rank matrix cannot contain a sub-matrix of the form $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ (and by a similar reasoning, a sub-matrix of the form $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$). We discuss further properties of the bipartite rank-shatter coefficients in Section 4.3; we first derive below our uniform convergence result in terms of these coefficients.

4.2 Uniform Convergence Bound

We first recall some classical inequalities that will be used in deriving our result, namely, Chebyshev's inequality and Hoeffding's inequality (Hoeffding, 1963):

Theorem 4 (Chebyshev's inequality). *Let X be a random variable. Then for any $\epsilon > 0$,*

$$\mathbf{P}\left\{|X - \mathbf{E}\{X\}| \geq \epsilon\right\} \;\leq\; \frac{\mathbf{Var}\{X\}}{\epsilon^2} \,.$$

Theorem 5 (Hoeffding, 1963). *Let $X_1, \ldots, X_N$ be independent bounded random variables such that X_k falls in the interval $[a_k, b_k]$ with probability one. Then for any $\epsilon > 0$,*

$$\mathbf{P}\left\{\left|\sum_{k=1}^N X_k - \mathbf{E}\left\{\sum_{k=1}^N X_k\right\}\right| \geq \epsilon\right\} \;\leq\; 2e^{-2\epsilon^2 / \sum_{k=1}^N (b_k - a_k)^2} \,.$$

We shall also need the following result of Devroye (1991), which bounds the variance of any fuction of a sample for which a single change in the sample has limited effect:

Theorem 6 (Devroye, 1991; Devroye et al., 1996, Chapter 9). *Let $X_1, \ldots, X_N$ be independent random variables with X_k taking values in a set A_k for each k. Let $\phi : (A_1 \times \cdots \times A_N) \to \mathbb{R}$ be such that*

$$\sup_{x_i \in A_i, x'_k \in A_k} \left|\phi(x_1, \ldots, x_N) - \phi(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_N)\right| \;\leq\; c_k \,.$$

Then

$$\mathbf{Var}\left\{\phi(X_1, \ldots, X_N)\right\} \;\leq\; \frac{1}{4}\sum_{k=1}^N c_k^2 \,.$$

We are now ready to give the main result of this section:

Theorem 7. *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\underline{y} = (y_1, \ldots, y_M) \in \mathcal{Y}^M$ be any label sequence of length $M \in \mathbb{N}$. Let m be the number of $+1$ labels in $\underline{y}$, and $n = M - m$ the number of -1 labels in $\underline{y}$. Then for any $\epsilon > 0$,*

$$\mathbf{P}_{S_X | S_Y = \underline{y}}\left\{\sup_{f \in \mathcal{F}} \left|\hat{A}(f; S) - A(f)\right| \geq \epsilon\right\} \;\leq\; 8r(\mathcal{F}, m, n)e^{-\max(m,n)\epsilon^2/32} \,.$$

Proof. We assume that $\max(m, n)\epsilon^2 \geq 2$, since otherwise the bound is trivial. We prove the result for the case $m \geq n$, *i.e.*, $\max(m, n) = m$; the case $m < n$ can be proved similarly. The proof consists of four steps and follows closely the proof of uniform convergence of relative frequencies to probabilities given in (Devroye et al., 1996, Chapter 12). To keep notation concise, we drop the

subscripts specifying random variables from all probabilities and expectations below; in each case, the probability/expectation is over all unconditioned random variables involved in the associated event. Note that all probabilities/expectations below are conditional given the label sequence $\underline{y}$.

Step 1. First symmetrization by a ghost sample.

For $i : y_i = +1$, define the random variables $\tilde{X}_i$ such that $X_i, \tilde{X}_i$ are all independent and identically distributed (according to $\mathcal{D}_{+1}$). Denote by $\tilde{S}_X$ the random sequence obtained from S_X by replacing X_i, for all $i : y_i = +1$, with $\tilde{X}_i$, and denote by $\tilde{S}$ the joint sequence $(\tilde{S}_X, \underline{y})$. Then for any $\epsilon > 0$ satisfying $m\epsilon^2 \geq 2$, we have

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \geq \epsilon\right\} \;\leq\; 2\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - \hat{A}(f;\tilde{S})\right| \geq \frac{\epsilon}{2}\right\}. \tag{13}$$

To see this, let $f^* \in \mathcal{F}$ be a function for which $|\hat{A}(f^*;S) - A(f^*)| \geq \epsilon$ if such a function exists, and let f^* be a fixed function in $\mathcal{F}$ otherwise. Then

$$\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - \hat{A}(f;\tilde{S})\right| \geq \frac{\epsilon}{2}\right\} \\
&\geq\; \mathbf{P}\left\{\left|\hat{A}(f^*;S) - \hat{A}(f^*;\tilde{S})\right| \geq \frac{\epsilon}{2}\right\} \\
&\geq\; \mathbf{P}\left\{\left|\hat{A}(f^*;S) - A(f^*)\right| \geq \epsilon, \left|\hat{A}(f^*;\tilde{S}) - A(f^*)\right| \leq \frac{\epsilon}{2}\right\} \\
&=\; \mathbf{E}\left\{\mathbf{I}_{\{|\hat{A}(f^*;S)-A(f^*)|\geq\epsilon\}}\mathbf{P}\left\{\left|\hat{A}(f^*;\tilde{S}) - A(f^*)\right| \leq \frac{\epsilon}{2} \;\middle|\; S_X\right\}\right\}.
\end{aligned} \tag{14}$$

The conditional probability inside can be bounded using Chebyshev's inequality (and Lemma 1):

$$\mathbf{P}\left\{\left|\hat{A}(f^*;\tilde{S}) - A(f^*)\right| \leq \frac{\epsilon}{2} \;\middle|\; S_X\right\} \;\geq\; 1 - \frac{\mathbf{Var}\left\{\hat{A}(f^*;\tilde{S}) \mid S_X\right\}}{\epsilon^2/4}.$$

Now, by the same reasoning as in the proof of Theorem 2, a change in the value of a single random variable $\tilde{X}_i$ can cause a change of at most $1/m$ in $\hat{A}(f^*;\tilde{S})$. Thus, by Theorem 6, we have

$$\begin{aligned}
\mathbf{Var}\left\{\hat{A}(f^*;\tilde{S}) \mid S_X\right\} &\leq \frac{1}{4}\sum_{\{i:y_i=+1\}}\left(\frac{1}{m}\right)^2 \\
&= \frac{1}{4m}.
\end{aligned}$$

This gives

$$\begin{aligned}
\mathbf{P}\left\{\left|\hat{A}(f^*;\tilde{S}) - A(f^*)\right| \leq \frac{\epsilon}{2} \;\middle|\; S_X\right\} &\geq 1 - \frac{\frac{1}{4m}}{\epsilon^2/4} \\
&= 1 - \frac{1}{m\epsilon^2} \\
&\geq \frac{1}{2},
\end{aligned}$$

whenever $m\epsilon^2 \geq 2$. Thus, from Eq. (14), we have

$$\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - \hat{A}(f;\tilde{S})\right| \geq \frac{\epsilon}{2}\right\} \;\geq\; \frac{1}{2}\mathbf{E}\left\{\mathbf{I}_{\{|\hat{A}(f^*;S)-A(f^*)|\geq\epsilon\}}\right\}$$

$$
\begin{aligned}
&= \frac{1}{2}\mathbf{P}\left\{\left|\hat{A}(f^*;S) - A(f^*)\right| \geq \epsilon\right\} \\
&\geq \frac{1}{2}\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \geq \epsilon\right\}.
\end{aligned}
$$

Step 2. Second symmetrization by random signs.

For $i : y_i = +1$, let σ_i be i.i.d. sign variables, independent of $S_X, \tilde{S}_X$, with $\mathbf{P}(\sigma_i = -1) = \mathbf{P}(\sigma_i = +1) = 1/2$. Clearly, since $X_i, \tilde{X}_i$ are all independent and identically distributed, the distribution of

$$
\sup_{f\in\mathcal{F}}\left|\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}} - \mathbf{I}_{\{f(\tilde{X}_i)>f(X_j)\}}\right)\right|
$$

is the same as the distribution of

$$
\sup_{f\in\mathcal{F}}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}} - \mathbf{I}_{\{f(\tilde{X}_i)>f(X_j)\}}\right)\right|
$$

Thus, from Step 1, we have

$$
\begin{aligned}
&\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \geq \epsilon\right\} \\
&\leq 2\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}} - \mathbf{I}_{\{f(\tilde{X}_i)>f(X_j)\}}\right)\right| \geq \frac{\epsilon}{2}\right\} \\
&= 2\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}} - \mathbf{I}_{\{f(\tilde{X}_i)>f(X_j)\}}\right)\right| \geq \frac{\epsilon}{2}\right\}. \qquad (15)
\end{aligned}
$$

Applying the union bound, we can remove the auxiliary random variables $\tilde{X}_i$:

$$
\begin{aligned}
&2\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}} - \mathbf{I}_{\{f(\tilde{X}_i)>f(X_j)\}}\right)\right| \geq \frac{\epsilon}{2}\right\} \\
&\leq 2\left[\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)>f(X_j)\}}\right| \geq \frac{\epsilon}{4}\right\}\right. \\
&\qquad\left. + \mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(\tilde{X}_i)>f(X_j)\}}\right| \geq \frac{\epsilon}{4}\right\}\right] \\
&= 4\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)>f(X_j)\}}\right| \geq \frac{\epsilon}{4}\right\}. \qquad (16)
\end{aligned}
$$

Step 3. Conditioning.

To bound the probability

$$
\mathbf{P}\left\{\sup_{f\in\mathcal{F}}\frac{1}{mn}\left|\sum_{\{i:y_i=+1\}}\sigma_i\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)>f(X_j)\}}\right| \geq \frac{\epsilon}{4}\right\},
$$

we condition on $S_X = (X_1, \ldots, X_M)$. Fix $\mathbf{x}_1, \ldots, \mathbf{x}_M \in \mathcal{X}$, and note that as f ranges over $\mathcal{F}$, the number of different random variables

$$\sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(\mathbf{x}_i)>f(\mathbf{x}_j)\}}$$

is at most the number of different bipartite rank matrices $\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')$ that can be realized by functions in $\mathcal{F}$, where $\underline{\mathbf{x}} \in \mathcal{X}^m$ contains $\mathbf{x}_i : y_i = +1$ and $\underline{\mathbf{x}}' \in \mathcal{X}^n$ contains $\mathbf{x}_j : y_j = -1$. This number, by definition, cannot exceed $r(\mathcal{F}, m, n)$. Therefore, conditional on $S_X = (X_1, \ldots, X_M)$, the supremum in the above probability is a maximum of at most $r(\mathcal{F}, m, n)$ random variables. Thus, by the union bound, we get

$$\begin{aligned}
&\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{1}{mn} \left| \sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(X_i)>f(X_j)\}} \right| \geq \frac{\epsilon}{4} \;\middle|\; S_X \right\} \\
&\quad \leq \; r(\mathcal{F}, m, n) \sup_{f\in\mathcal{F}} \mathbf{P}\left\{ \frac{1}{mn} \left| \sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(X_i)>f(X_j)\}} \right| \geq \frac{\epsilon}{4} \;\middle|\; S_X \right\} . \qquad (17)
\end{aligned}$$

Step 4. Hoeffding's inequality.

With $\mathbf{x}_1, \ldots, \mathbf{x}_M$ fixed, the quantity

$$\sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(\mathbf{x}_i)>f(\mathbf{x}_j)\}}$$

is the sum of m independent zero-mean random variables bounded between $-n$ and n. Therefore, by Hoeffding's inequality, we have

$$\begin{aligned}
\mathbf{P}\left\{ \frac{1}{mn} \left| \sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(X_i)>f(X_j)\}} \right| \geq \frac{\epsilon}{4} \;\middle|\; S_X \right\} &\leq 2e^{-2m^2n^2\epsilon^2/16m4n^2} \\
&= 2e^{-m\epsilon^2/32} .
\end{aligned}$$

Thus, by Step 3,

$$\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{1}{mn} \left| \sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(X_i)>f(X_j)\}} \right| \geq \frac{\epsilon}{4} \;\middle|\; S_X \right\} \leq 2r(\mathcal{F}, m, n)e^{-m\epsilon^2/32} .$$

Taking the expected value on both sides, we have

$$\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \frac{1}{mn} \left| \sum_{\{i:y_i=+1\}} \sigma_i \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(X_i)>f(X_j)\}} \right| \geq \frac{\epsilon}{4} \right\} \leq 2r(\mathcal{F}, m, n)e^{-m\epsilon^2/32} . \qquad (18)$$

Thus, putting everything together, we get

$$\mathbf{P}\left\{ \sup_{f\in\mathcal{F}} \left| \hat{A}(f; S) - A(f) \right| \geq \epsilon \right\} \leq 8r(\mathcal{F}, m, n)e^{-m\epsilon^2/32} .$$

□

From Theorem 7, we can derive a confidence interval interpretation of the bound as follows:

Corollary 7. *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\underline{y} = (y_1, \ldots, y_M) \in \mathcal{Y}^M$ be any label sequence of length $M \in \mathbb{N}$. Let m be the number of $+1$ labels in $\underline{y}$, and $n = M - m$ the number of -1 labels in $\underline{y}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f;S) - A(f) \right| \geq \sqrt{\frac{32 \left(\ln r(\mathcal{F}, m, n) + \ln \left(\frac{8}{\delta} \right) \right)}{\max(m,n)}} \right\} \quad \leq \quad \delta \, .$$

Proof. This follows directly from Theorem 7 by setting $8r(\mathcal{F}, m, n)e^{-\max(m,n)\epsilon^2/32} = \delta$ and solving for ϵ. □

4.3 Properties of Bipartite Rank-Shatter Coefficients

We mentioned in Section 4.1 that $r(\mathcal{F}, m, n) \leq g(m, n)$ for all $m, n \geq 2$, where $g(m, n)$ is the number of matrices in $\{0,1\}^{m \times n}$ that do not contain a sub-matrix of the form

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} .$$

The number $g(m, n)$ is strictly smaller than 2^{mn}, but is still exponential in m, n. For the bound of Theorem 7 to be meaningful, one needs to find an upper bound on $r(\mathcal{F}, m, n)$ that is polynomial in m, n. Below we provide one method for deriving such an upper bound by relating the bipartite rank-shatter coefficients $r(\mathcal{F}, m, n)$ of a class of ranking functions $\mathcal{F}$ to the standard shatter coefficients and VC dimension of a class of classification functions derived from $\mathcal{F}$.

We first recall the definitions of standard shatter coefficients and VC dimension, quantities that play a central role in uniform convergence results for the classification error rate. Let $\mathcal{H}$ be a class of binary-valued functions on $\mathcal{X}$, and let $N \in \mathbb{N}$. Then the N-th *shatter coefficient* of $\mathcal{H}$, denoted by $s(\mathcal{H}, N)$, is defined as follows:

$$s(\mathcal{H}, N) \quad = \quad \max_{\underline{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in \mathcal{X}^N} \left| \{ (h(\mathbf{x}_1), \ldots, h(\mathbf{x}_N)) \mid h \in \mathcal{H} \} \right| . \tag{19}$$

Clearly, $s(\mathcal{H}, N) \leq 2^N$ for all N. The largest integer k for which $s(\mathcal{H}, k) = 2^k$ is called the *Vapnik-Chervonenkis dimension* (or *VC dimension*) of the class $\mathcal{H}$, denoted by $V_{\mathcal{H}}$.

We also recall the following standard result due to Vapnik and Chervonenkis (1971) and Sauer (1972) which provides an upper bound on the shatter coefficients in terms of the VC dimension:

Theorem 8 (Vapnik and Chervonenkis, 1971; Sauer, 1972). *Let $\mathcal{H}$ be a class of binary-valued functions on $\mathcal{X}$, with VC dimension $V_{\mathcal{H}}$. Then for all $N \geq 2V_{\mathcal{H}}$,*

$$s(\mathcal{H}, N) \;\leq\; \sum_{i=0}^{V_{\mathcal{H}}} \binom{N}{i} \;\leq\; \left(\frac{eN}{V_{\mathcal{H}}} \right)^{V_{\mathcal{H}}} .$$

Next, we define a series of classification function classes derived from a given ranking function class. Only the first two function classes are used in this section; the third is needed in Section 4.4.

Definition 5 (Function classes). *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Define the following classes of binary-valued functions derived from $\mathcal{F}$:*

$$\begin{aligned}
1.\quad \bar{\mathcal{F}} &= \{\bar{f} : \mathcal{X}\to\mathcal{Y} \mid \bar{f}(\mathbf{x}) = \operatorname{sign}(f(\mathbf{x})) \text{ for some } f \in \mathcal{F}\} && (20)\\
2.\quad \tilde{\mathcal{F}} &= \{\tilde{f} : \mathcal{X}\times\mathcal{X}\to\mathcal{Y} \mid \tilde{f}(\mathbf{x},\mathbf{x}') = \operatorname{sign}(f(\mathbf{x}) - f(\mathbf{x}')) \text{ for some } f \in \mathcal{F}\} && (21)\\
3.\quad \check{\mathcal{F}} &= \{\check{f}_{\mathbf{z}} : \mathcal{X}\to\mathcal{Y} \mid \check{f}_{\mathbf{z}}(\mathbf{x}) = \operatorname{sign}(f(\mathbf{x}) - f(\mathbf{z})) \text{ for some } f \in \mathcal{F}, \mathbf{z} \in \mathcal{X}\} && (22)
\end{aligned}$$

The following result relates the bipartite rank-shatter coefficients of a given class of ranking functions $\mathcal{F}$ to the standard shatter coefficients of $\tilde{\mathcal{F}}$:

Theorem 9. *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\tilde{\mathcal{F}}$ be the class of binary-valued functions on $\mathcal{X}\times\mathcal{X}$ defined by Eq. (21). Then for all $m, n \in \mathbb{N}$,*

$$r(\mathcal{F}, m, n) \;\leq\; s(\tilde{\mathcal{F}}, mn)\,.$$

Proof. For any $m, n \in \mathbb{N}$, we have

$$\begin{aligned}
r(\mathcal{F}, m, n) &= \max_{\underline{\mathbf{x}}\in\mathcal{X}^m, \underline{\mathbf{x}}'\in\mathcal{X}^n} \left|\left\{ \left[\mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}'_j)\}}\right] \;\middle|\; f \in \mathcal{F}\right\}\right| \\
&= \max_{\underline{\mathbf{x}}\in\mathcal{X}^m, \underline{\mathbf{x}}'\in\mathcal{X}^n} \left|\left\{ \left[\mathbf{I}_{\{\tilde{f}(\mathbf{x}_i,\mathbf{x}'_j) = +1\}}\right] \;\middle|\; \tilde{f} \in \tilde{\mathcal{F}}\right\}\right| \\
&= \max_{\underline{\mathbf{x}}\in\mathcal{X}^m, \underline{\mathbf{x}}'\in\mathcal{X}^n} \left|\left\{ \left[\tilde{f}(\mathbf{x}_i,\mathbf{x}'_j)\right] \;\middle|\; \tilde{f} \in \tilde{\mathcal{F}}\right\}\right| \\
&\leq \max_{\mathbf{X},\mathbf{X}'\in\mathcal{X}^{m\times n}} \left|\left\{ \left[\tilde{f}(\mathbf{x}_{ij},\mathbf{x}'_{ij})\right] \;\middle|\; \tilde{f} \in \tilde{\mathcal{F}}\right\}\right| \\
&= \max_{\underline{\mathbf{x}},\underline{\mathbf{x}'}\in\mathcal{X}^{mn}} \left|\left\{ \left(\tilde{f}(\mathbf{x}_1,\mathbf{x}'_1),\ldots,\tilde{f}(\mathbf{x}_{mn},\mathbf{x}'_{mn})\right) \;\middle|\; \tilde{f} \in \tilde{\mathcal{F}}\right\}\right| \\
&= s(\tilde{\mathcal{F}}, mn)\,.
\end{aligned}$$

□

From Theorem 8, we have the following immediate corollary to Theorem 9:

Corollary 8. *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\tilde{\mathcal{F}}$ be the class of binary-valued functions on $\mathcal{X}\times\mathcal{X}$ defined by Eq. (21). Then for all $m, n \in \mathbb{N}$ satisfying $mn \geq 2V_{\tilde{\mathcal{F}}}$,*

$$r(\mathcal{F}, m, n) \;\leq\; \left(\frac{emn}{V_{\tilde{\mathcal{F}}}}\right)^{V_{\tilde{\mathcal{F}}}}.$$

Characterizing the shatter coefficients $s(\tilde{\mathcal{F}}, N)$ or VC dimension $V_{\tilde{\mathcal{F}}}$ of the class $\tilde{\mathcal{F}}$ may not be straightforward for all classes of ranking functions $\mathcal{F}$. However, in cases where $s(\tilde{\mathcal{F}}, N)$ can be shown to grow polynomially in N or $V_{\tilde{\mathcal{F}}}$ can be shown to be finite, the above results provide an immediate polynomial upper bound on $r(\mathcal{F}, m, n)$. Below we derive such a polynomial bound for the case of linear ranking functions.

Theorem 10. *For $d \in \mathbb{N}$, let $\mathcal{F}_{\mathrm{lin}(d)}$ denote the class of linear ranking functions on $\mathbb{R}^d$, given by*

$$\mathcal{F}_{\mathrm{lin}(d)} \;=\; \{f : \mathbb{R}^d\to\mathbb{R} \mid f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b \text{ for some } \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}\,.$$

Then for all $m, n \in \mathbb{N}$ satisfying $mn \geq 2(2d+1)$,

$$r(\mathcal{F}_{\mathrm{lin}(d)}, m, n) \;\leq\; \left(\frac{emn}{2d+1}\right)^{2d+1}.$$

Proof. We have,

$$\begin{aligned}\tilde{F}_{\mathrm{lin}(d)} &= \{\tilde{f} : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{Y} \mid \tilde{f}(\mathbf{x}, \mathbf{x}') = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}\cdot\mathbf{x}') \text{ for some } \mathbf{w} \in \mathbb{R}^d\} \\ &\subset \{g : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{Y} \mid g(\mathbf{x}, \mathbf{x}') = \mathrm{sign}(\mathbf{w}\cdot\mathbf{x} + \mathbf{w}'\cdot\mathbf{x}' + b) \text{ for some } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d, b \in \mathbb{R}\} \\ &= \bar{F}_{\mathrm{lin}(2d)}\,.\end{aligned}$$

Thus, for all $N \in \mathbb{N}$,

$$s(\tilde{\mathcal{F}}_{\mathrm{lin}(d)}, N) \;\leq\; s(\bar{\mathcal{F}}_{\mathrm{lin}(2d)}, N)\,. \tag{23}$$

This gives

$$\begin{aligned}r(\mathcal{F}_{\mathrm{lin}(d)}, m, n) &\leq s(\tilde{\mathcal{F}}_{\mathrm{lin}(d)}, mn) \qquad \text{(by Theorem 9)} \\ &\leq s(\bar{\mathcal{F}}_{\mathrm{lin}(2d)}, mn) \qquad \text{(from Eq. (23))} \\ &\leq \left(\frac{emn}{2d+1}\right)^{2d+1}\end{aligned}$$

whenever $mn \geq 2(2d+1)$, by Theorem 8 and the fact that the VC dimension of $\bar{F}_{\mathrm{lin}(2d)}$ (*i.e.*, the VC dimension of the class of linear classification functions on $\mathbb{R}^{2d}$) is $2d+1$. □

We note that the method used in the above proof can be used to establish a similar result for higher-order polynomial ranking functions.

4.4 Comparison with Uniform Convergence Bound of Freund et al.

Freund et al. (2003) recently derived a uniform convergence bound for a quantity closely related to the AUC. They define a ranking loss which, as pointed out by Cortes and Mohri (2004), reduces to one minus the AUC in the case of bipartite ranking problems. More specifically, given a data sequence $T = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ in a bipartite ranking problem, containing m positive and n negative examples, the *ranking loss* of a ranking function $f : \mathcal{X} \to \mathbb{R}$ with respect to T, which we denote by $\hat{R}(f;T)$, is given by

$$\hat{R}(f;T) \;=\; \frac{1}{mn} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(\mathbf{x}_i) \leq f(\mathbf{x}_j)\}}\,. \tag{24}$$

Comparing this to Eq. (7), it is clear that

$$\hat{R}(f;T) \;=\; 1 - \hat{A}(f;T)\,. \tag{25}$$

For the case of bipartite ranking problems, Freund et al. (2003) also define a notion of the expected ranking loss. In our notation, the *expected ranking loss* of a ranking function f, which we denote by $R(f)$, can be expressed as

$$R(f) \;=\; \mathbf{E}_{X \sim \mathcal{D}_{+1}, X' \sim \mathcal{D}_{-1}} \left\{\mathbf{I}_{\{f(X) \leq f(X')\}}\right\}\,. \tag{26}$$

Again, comparing to Eq. (8), it is clear that

$$R(f) \;=\; 1 - A(f)\,. \tag{27}$$

Freund et al. (2003) derived a uniform convergence bound for the ranking loss in bipartite ranking problems. Since from Eqs. (25) and (27) we have $|\hat{R}(f;T) - R(f)| = |\hat{A}(f;T) - A(f)|$, this implies

a uniform convergence bound for the AUC. Although the result in (Freund et al., 2003) is given only for function classes considered by their RankBoost algorithm, their technique is generally applicable. We state and prove their result below, using our notation, for the general case (*i.e.*, function classes not restricted to those considered by RankBoost), and then offer a comparison of our bound with theirs. As in (Freund et al., 2003), the result is given in the form of a confidence interval. The result in (Freund et al., 2003) was stated in terms of the VC dimension; we state the result directly in terms of shatter coefficients since this provides a tighter bound.[4] The proof makes use of the following uniform convergence result of Vapnik (1982) for the classification error rate:

Theorem 11 (Vapnik, 1982). *Let $\mathcal{H}$ be a class of binary-valued functions on $\mathcal{X}$, and let $M \in \mathbb{N}$. Then for any $\epsilon > 0$,*

$$\mathbf{P}_{S\sim\mathcal{D}^M}\left\{\sup_{h\in\mathcal{H}}\left|\hat{L}(h;S) - L(h)\right| \geq \epsilon\right\} \;\leq\; 6s(\mathcal{H},2M)e^{-M\epsilon^2/4}\,.$$

Theorem 12 (Generalization of Freund et al., 2003, Theorem 3). *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\underline{y} = (y_1,\ldots,y_M) \in \mathcal{Y}^M$ be any label sequence of length $M \in \mathbb{N}$. Let m be the number of $+1$ labels in $\underline{y}$, and $n = M - m$ the number of -1 labels in $\underline{y}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \geq 2\sqrt{\frac{\ln s(\check{\mathcal{F}},2m) + \ln\left(\frac{12}{\delta}\right)}{m}} + 2\sqrt{\frac{\ln s(\check{\mathcal{F}},2n) + \ln\left(\frac{12}{\delta}\right)}{n}}\right\} \;\leq\; \delta\,,$$

where $\check{\mathcal{F}}$ is the class of binary-valued functions on $\mathcal{X}$ defined by Eq. (22).

Proof. Given $S_Y = \underline{y}$, we have for all $f \in \mathcal{F}$

$$\begin{aligned}
\left|\hat{A}(f;S) - A(f)\right| &= \left|\hat{R}(f;S) - R(f)\right| \\
&= \left|\frac{1}{mn}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)\leq f(X_j)\}} - \mathbf{E}_{X\sim\mathcal{D}_{+1},X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X)\leq f(X')\}}\right\}\right| \\
&= \left|\frac{1}{mn}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)\leq f(X_j)\}} - \frac{1}{m}\sum_{\{i:y_i=+1\}}\mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X_i)\leq f(X')\}}\right\}\right. \\
&\qquad \left. + \frac{1}{m}\sum_{\{i:y_i=+1\}}\mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X_i)\leq f(X')\}}\right\} - \mathbf{E}_{X\sim\mathcal{D}_{+1},X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X)\leq f(X')\}}\right\}\right| \\
&= \left|\frac{1}{m}\sum_{\{i:y_i=+1\}}\left(\frac{1}{n}\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)\leq f(X_j)\}} - \mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X_i)\leq f(X')\}}\right\}\right)\right. \\
&\qquad \left. + \mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\frac{1}{m}\sum_{\{i:y_i=+1\}}\mathbf{I}_{\{f(X_i)\leq f(X')\}} - \mathbf{E}_{X\sim\mathcal{D}_{+1}}\left\{\mathbf{I}_{\{f(X)\leq f(X')\}}\right\}\right\}\right|
\end{aligned}$$

[4] In fact, due to the use of a looser upper bound on the shatter coefficients than that given by Theorem 8, even the VC dimension statement of Freund et al. (2003) is slightly looser than it could be; in particular, the $\ln(18/\delta)$ terms in their bound can be replaced by $\ln(12/\delta)$.

$$
\begin{aligned}
&\leq \frac{1}{m}\sum_{\{i:y_i=+1\}}\left|\frac{1}{n}\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f(X_i)\leq f(X_j)\}}-\mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f(X_i)\leq f(X')\}}\right\}\right| \\
&\quad +\mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\left|\frac{1}{m}\sum_{\{i:y_i=+1\}}\mathbf{I}_{\{f(X_i)\leq f(X')\}}-\mathbf{E}_{X\sim\mathcal{D}_{+1}}\left\{\mathbf{I}_{\{f(X)\leq f(X')\}}\right\}\right|\right\} \\
&\leq \sup_{f'\in\mathcal{F},\mathbf{z}\in\mathcal{X}}\left|\frac{1}{n}\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{f'(\mathbf{z})\leq f'(X_j)\}}-\mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{f'(\mathbf{z})\leq f'(X')\}}\right\}\right| \\
&\quad +\sup_{f'\in\mathcal{F},\mathbf{z}\in\mathcal{X}}\left|\frac{1}{m}\sum_{\{i:y_i=+1\}}\mathbf{I}_{\{f'(X_i)\leq f'(\mathbf{z})\}}-\mathbf{E}_{X\sim\mathcal{D}_{+1}}\left\{\mathbf{I}_{\{f'(X)\leq f'(\mathbf{z})\}}\right\}\right| \\
&= \sup_{\check{f}_{\mathbf{z}}\in\check{\mathcal{F}}}\left|\frac{1}{n}\sum_{\{j:y_j=-1\}}\mathbf{I}_{\{\check{f}_{\mathbf{z}}(X_j)=+1\}}-\mathbf{E}_{X'\sim\mathcal{D}_{-1}}\left\{\mathbf{I}_{\{\check{f}_{\mathbf{z}}(X')=+1\}}\right\}\right| \\
&\quad +\sup_{\check{f}_{\mathbf{z}}\in\check{\mathcal{F}}}\left|\frac{1}{m}\sum_{\{i:y_i=+1\}}\mathbf{I}_{\{\check{f}_{\mathbf{z}}(X_i)=-1\}}-\mathbf{E}_{X\sim\mathcal{D}_{+1}}\left\{\mathbf{I}_{\{\check{f}_{\mathbf{z}}(X')=-1\}}\right\}\right| .
\end{aligned}
$$

If we augment the notation $L(h)$ used to denote the expected error rate with the distribution, *e.g.*, $L_{\mathcal{D}}(h)$, we thus get

$$
\begin{aligned}
\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S)-A(f)\right| &\leq \sup_{\check{f}_{\mathbf{z}}\in\check{\mathcal{F}}}\left|\hat{L}(\check{f}_{\mathbf{z}};S_{-1}^{(n)})-L_{\mathcal{D}_{-1}}(\check{f}_{\mathbf{z}})\right| \\
&\quad +\sup_{\check{f}_{\mathbf{z}}\in\check{\mathcal{F}}}\left|\hat{L}(\check{f}_{\mathbf{z}};S_{+1}^{(m)})-L_{\mathcal{D}_{+1}}(\check{f}_{\mathbf{z}})\right| , \qquad (28)
\end{aligned}
$$

where $S_{+1}^{(m)}$ and $S_{-1}^{(n)}$ denote the subsequences of S containing the m positive and n negative examples, respectively. Now, from the confidence interval interpretation of Theorem 11, we have

$$
\mathbf{P}_{S_{+1}^{(m)}\sim\mathcal{D}_{+1}^{m}}\left\{\sup_{\check{f}_{\mathbf{z}}\in\check{\mathcal{F}}}\left|\hat{L}(\check{f}_{\mathbf{z}};S_{+1}^{(m)})-L_{\mathcal{D}_{+1}}(\check{f}_{\mathbf{z}})\right|\geq 2\sqrt{\frac{\ln s(\check{\mathcal{F}},2m)+\ln\left(\frac{12}{\delta}\right)}{m}}\right\} \leq \frac{\delta}{2}, \qquad (29)
$$

$$
\mathbf{P}_{S_{-1}^{(n)}\sim\mathcal{D}_{-1}^{n}}\left\{\sup_{\check{f}_{\mathbf{z}}\in\check{\mathcal{F}}}\left|\hat{L}(\check{f}_{\mathbf{z}};S_{-1}^{(n)})-L_{\mathcal{D}_{-1}}(\check{f}_{\mathbf{z}})\right|\geq 2\sqrt{\frac{\ln s(\check{\mathcal{F}},2n)+\ln\left(\frac{12}{\delta}\right)}{n}}\right\} \leq \frac{\delta}{2}. \qquad (30)
$$

Combining Eqs. (28-30) gives the desired result. □

We now compare the uniform convergence bound derived in Section 4.2 with that of Freund et al. Since we do not have a means to compare the quantities involved in the two bounds (namely, $r(\mathcal{F},m,n)$ and $s(\check{\mathcal{F}},2m), s(\check{\mathcal{F}},2n)$) for general classes of ranking functions $\mathcal{F}$, we compare the two bounds for the case of linear ranking functions for which both quantities can be characterized. We first note that the constants and exponent in our bound have not been optimized, whereas the bound of Freund et al. above, through its use of the relatively optimized result of Theorem 11, contains tight constants. To remove artefacts due to this difference and make a fair comparison, we use a looser version of Theorem 12 obtained by replacing the optimized result of Theorem 11 with a result whose proof technique and constants correspond more closely to ours. (As we note below,

this does not affect the nature of our conclusions regarding the relative quality of the two bounds.) In particular, on replacing Theorem 11 with a result given in (Devroye et al., 1996, Theorem 12.6) which states that for any $\epsilon > 0$,

$$\mathbf{P}_{S\sim\mathcal{D}^M}\left\{\sup_{h\in\mathcal{H}}\left|\hat{L}(h;S) - L(h)\right| \geq \epsilon\right\} \quad \leq \quad 8s(\mathcal{H}, M)e^{-M\epsilon^2/32}\,, \tag{31}$$

the bound of Theorem 12 becomes

$$\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \;\geq\; \sqrt{\frac{32\left(\ln s(\check{\mathcal{F}}, m) + \ln\left(\frac{16}{\delta}\right)\right)}{m}} + \sqrt{\frac{32\left(\ln s(\check{\mathcal{F}}, n) + \ln\left(\frac{16}{\delta}\right)\right)}{n}}\right\} \leq \delta\,. \tag{32}$$

For the case of linear ranking functions on $\mathbb{R}^d$, it is easily seen that $\check{\mathcal{F}}_{\mathrm{lin}(d)} = \bar{\mathcal{F}}_{\mathrm{lin}(d)}$. Therefore for all $N \geq 2(d+1)$, we have

$$s(\check{\mathcal{F}}_{\mathrm{lin}(d)}, N) = s(\bar{\mathcal{F}}_{\mathrm{lin}(d)}, N) \leq \left(\frac{eN}{d+1}\right)^{d+1}\,,$$

by Theorem 8 and the fact that the VC dimension of $\bar{\mathcal{F}}_{\mathrm{lin}(d)}$ is $d+1$. Using this in Eq. (32), we thus get from the bound of Freund et al. (for $m \geq 2(d+1)$, $n \geq 2(d+1)$):

$$\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}_{\mathrm{lin}(d)}}\left|\hat{A}(f;S) - A(f)\right| \;\geq\; \sqrt{\frac{32\left((d+1)\left(\ln\left(\frac{m}{d+1}\right)+1\right) + \ln\left(\frac{16}{\delta}\right)\right)}{m}} + \sqrt{\frac{32\left((d+1)\left(\ln\left(\frac{n}{d+1}\right)+1\right) + \ln\left(\frac{16}{\delta}\right)\right)}{n}}\right\} \leq \delta\,. \tag{33}$$

On the other hand, from Corollary 7 and Theorem 10, we get from our bound (for $mn \geq 2(2d+1)$):

$$\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}_{\mathrm{lin}(d)}}\left|\hat{A}(f;S) - A(f)\right| \;\geq\; \sqrt{\frac{32\left((2d+1)\left(\ln\left(\frac{mn}{2d+1}\right)+1\right) + \ln\left(\frac{8}{\delta}\right)\right)}{\max(m,n)}}\right\} \quad \leq \quad \delta\,. \tag{34}$$

We plot the two bounds in Figure 3 for $d = 10$, $\delta = 0.01$ and various values of $m/(m+n)$. As can be seen, the bounds are comparable for $m = n$, but as soon as the proportion of positive examples $m/(m+n)$ departs from $1/2$, our bound is tighter. The difference is considerable when $m/(m+n)$ is far from $1/2$; when $m/(m+n) = 1/100$, for example, our bound is already meaningful ($\epsilon < 0.5$) at a sample size of $50,000$, whereas the bound of Freund et al. remains larger than 0.5 even at a sample size of $1,000,000$. We note that the absolute sample sizes here are rather large due to the unoptimized constants in the above results; it should be possible to improve these by optimizing the constants. We also note that the qualitative conclusions regarding the relative tightness of the two bounds are unaffected by our decision to use loose constants in the bound of Freund et al.; in particular, when their bound is allowed to retain the optimized constants given in Theorem 12, it outperforms our unoptimized bound over a small range of values of $m/(m+n)$ close to $1/2$ (this range is roughly between $1/21$ and $20/21$ for $d = 10$, $\delta = 0.01$), but again, for values of $m/(m+n)$ far from $1/2$, our bound is considerably tighter.

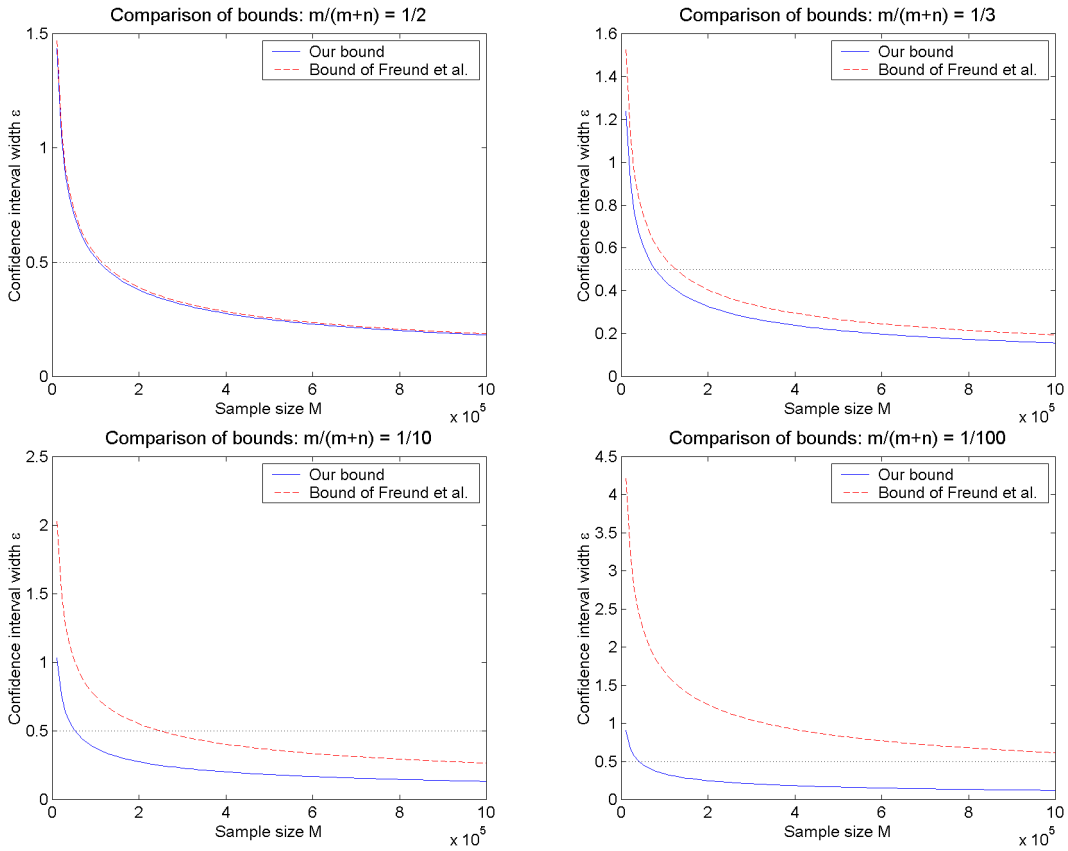

Figure 3: A comparison of our uniform convergence bound with that of Freund et al. (2003) for the case of linear ranking functions on $\mathbb{R}^d$. The plots are for $d = 10, \delta = 0.01$, and show how the confidence interval width ϵ given by the two bounds varies with the sample size M, for various values of $m/(m+n)$. The bounds are comparable for $m = n$, but as the proportion of positive examples $m/(m+n)$ departs from $1/2$, our bound is increasingly tighter. In particular, for $m/(m+n) = 1/100$, our bound is already meaningful ($\epsilon < 0.5$) at a sample size of $50,000$, whereas the bound of Freund et al. remains above 0.5 even at a sample size of $1,000,000$. (See text for details.)

5 Discussion

We have developed generalization bounds for the area under an ROC curve (AUC), a quantity used as an evaluation criterion for bipartite ranking problems. We have derived both a large deviation bound for the AUC and a uniform convergence bound. The large deviation bound serves to bound the expected accuracy of a ranking function in terms of its empirical AUC on a test sequence; the uniform convergence bound serves to bound the expected accuracy of a learned ranking function in terms of its empirical AUC on a training sequence. Both our bounds are distribution-free.

Our large deviation result was derived using a powerful concentration inequality of McDiarmid. A comparison with corresponding large deviation results for the error rate in classification suggests that, for given $0 < \epsilon, \delta \leq 1$, the test sample size required to obtain an ϵ-accurate estimate of the expected accuracy of a ranking function with δ-confidence is larger than the test sample size required to obtain a similar estimate of the expected error rate of a classification function. A simple application of the union bound allows the large deviation bound to be extended to learned ranking

functions chosen from finite function classes.

Our uniform convergence bound applies to learned ranking functions chosen from general (possibly infinite) function classes. The bound is expressed in terms of a new set of combinatorial parameters that we have termed the bipartite rank-shatter coefficients. These coefficients define a new measure of complexity for classes of real-valued functions. We have derived an upper bound on the bipartite rank-shatter coefficients in terms of the standard shatter coefficients studied in classification. This upper bound allows a characterization of the bipartite rank-shatter coefficients in the case of linear (and more generally, polynomial) ranking functions. A comparison of our bound in the case of linear ranking functions with a recent uniform convergence bound of Freund et al. (2003), which is expressed directly in terms of standard shatter coefficients from results for classification, shows that our bound is considerably tighter, thus attesting to the appropriateness of the new coefficients as a suitable complexity measure for the study of bipartite ranking problems.

Our study raises several interesting questions. First, what other function classes can be shown to have small complexity for ranking, *i.e.*, for what other function classes $\mathcal{F}$ can we show that the bipartite rank-shatter coefficients $r(\mathcal{F}, m, n)$ grow polynomially in m, n? Is it possible to find tighter bounds on the bipartite rank-shatter coefficients of $\mathcal{F}$ than those provided by the relation to standard shatter coefficients of $\tilde{\mathcal{F}}$? Second, can we establish generalization bounds for other forms of ranking problems (*i.e.*, other than bipartite)? Third, do there exist data-dependent bounds for ranking, analogous to existing margin bounds for classification?

Finally, we point out a curious complementarity in the behaviours of the large deviation bound of Section 3 and the uniform convergence bound of Section 4. In particular, for fixed $0 < \delta \leq 1$, the confidence interval provided by the large deviation bound is smallest when the number of positive examples m is equal to the number of negative examples n, and grows larger as the proportion of positive examples $m/(m+n)$ departs from $1/2$. On the contrary, the confidence interval provided by the uniform convergence bound is smallest when $m/(m+n)$ is far from $1/2$, and grows larger as $m/(m+n)$ approaches $1/2$. This contrast in the behaviour of the two bounds can be seen most clearly in the case of learned ranking functions chosen from finite function classes, to which both bounds apply. In this case, for large values of the sample size $M = m + n$, the bipartite rank-shatter coefficients $r(\mathcal{F}, m, n)$ can be taken to be equal to $|\mathcal{F}|$, and the confidence interval provided by Corollary 7 reduces to

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - A(f) \right| \geq \sqrt{\frac{32 \left(\ln |\mathcal{F}| + \ln \left(\frac{8}{\delta} \right) \right)}{\max(m, n)}} \right\} \quad \leq \quad \delta \,. \tag{35}$$

Comparing this to Corollary 4 (and noting that $\rho(\underline{y})(1 - \rho(\underline{y})) = mn/(m+n)$ here), we see that the large deviation bounds gives a tighter confidence interval for

$$\frac{\min(m, n)}{m + n} \quad \geq \quad \frac{1}{64} - \frac{\ln 4}{64(\ln |\mathcal{F}| + \ln \left(\frac{8}{\delta} \right))} \,,$$

while the uniform convergence bound gives a tighter confidence interval for

$$\frac{\min(m, n)}{m + n} \quad < \quad \frac{1}{64} - \frac{\ln 4}{64(\ln |\mathcal{F}| + \ln \left(\frac{8}{\delta} \right))} \,.$$

The two bounds are plotted for $|\mathcal{F}| = 1000$, $\delta = 0.01$, $M = 100,000$ in Figure 4. Whether the constants in one bound can be improved to the extent that it will be universally better than the other (for finite function classes) remains an open question. It also remains an open question

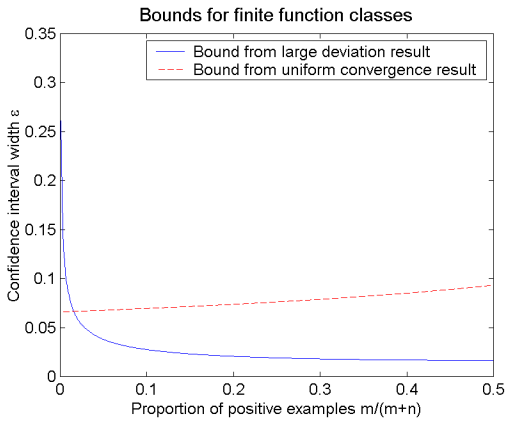

Figure 4: A comparison of the two bounds for learned ranking functions chosen from finite function classes provided by the large deviation result of Section 3 and the uniform convergence result of Section 4. The plot is for $|\mathcal{F}| = 1000$, $\delta = 0.01$, $M = 100,000$. The bounds are complementary in behaviour with respect to the proportion of positive examples $m/(m+n)$; the large deviation bound is tighter for values of $m/(m+n)$ closer to $1/2$, while the uniform convergence bound is tighter for values of $m/(m+n)$ far from $1/2$. (Note that the bounds are shown only for $0 < m/(m+n) \leq 1/2$ as they are symmetric about $m/(m+n) = 1/2$.)

whether large deviation and uniform convergence results for the AUC will necessarily have this complementary behaviour, or whether one of the two bounds derived here is fundamentally loose and can be replaced by a better bound derived using a different proof technique. A possible route for deriving an alternative large deviation bound for the AUC could be via the theory of U-statistics (de la Pẽna and Giné, 1999); possible routes for an alternative uniform convergence bound could include the theory of compression bounds (Littlestone and Warmuth, 1986; Graepel et al., 2004).

Acknowledgements

This research was supported in part by NSF ITR grants IIS 00-85980 and IIS 00-85836.

References

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

Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

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

Víctor H. de la Pẽna and Evarist Giné. *Decoupling: From Dependence to Independence*. Springer-Verlag, New York, 1999.

Luc Devroye. Exponential inequalities in nonparametric estimation. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, NATO ASI Series, pages 31–44. Kluwer Academic Publishers, 1991.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

James P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

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.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 2004. To appear.

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

Simon I. Hill, Hugo Zaragoza, Ralf Herbrich, and Peter J. W. Rayner. Average precision and the problem of generalisation. In *Proceedings of the ACM SIGIR Workshop on Mathematical and Formal Methods in Information Retrieval*, 2002.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.

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

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

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

Lian Yan, Robert Dodier, Michael C. Mozer, and Richard Wolniewicz. Optimizing classifier performance via an approximation to the Wilcoxon-Mann-Whitney statistic. In *Proceedings of the 20th International Conference on Machine Learning*, pages 848–855, 2003.