Generalization Bounds for the Area Under the ROC Curve*

Shivani Agarwal SAGARWAL@CS.UIUC.EDU
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue
Urbana, IL 61801, USA

Thore Graepel THOREG@MICROSOFT.COM
Ralf Herbrich RHERB@MICROSOFT.COM
Microsoft Research
7 JJ Thomson Avenue
Cambridge CB3 0FB, UK

Sariel Har-Peled SARIEL@CS.UIUC.EDU
Dan Roth DANR@CS.UIUC.EDU
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue
Urbana, IL 61801, USA

Editor: Michael I. Jordan

Abstract

We study generalization properties of the area under the ROC curve (AUC), a quantity that has been advocated as an evaluation criterion for the bipartite ranking problem. The AUC is a different 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 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 VC-dimension related shatter coefficients (also known as the growth function) in uniform convergence results for the classification error rate. 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 shows that the bound provided by our result can be considerably tighter.

Keywords: Generalization Bounds, Area Under the ROC Curve, Ranking, Large Deviations, Uniform Convergence.

*. Parts of the results contained in this paper were presented at the *18th Annual Conference on Neural Information Processing Systems* in December 2004 (Agarwal et al., 2005a) and at the *10th International Workshop on Artificial Intelligence and Statistics* in January 2005 (Agarwal et al., 2005b).

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 come from 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 ones. 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 the bipartite ranking problem is similar to that of the binary classification problem. In both problems, 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 $\mathcal{Y} = \{-1, +1\}$. 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 $h(\mathbf{x}) = \theta(f_h(\mathbf{x}))$ for some real-valued function $f_h : \mathcal{X} \to \mathbb{R}$, where

$$\theta(u) = \begin{cases} 1 & \text{if } u > 0 \\ -1 & \text{otherwise} \end{cases}, \tag{1}$$

and then taking f_h to be the desired ranking function.[1] 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. In Herbrich et al. (2000) the problem of learning a ranking function is also reduced to a classification problem, but on *pairs* of instances.

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

$$L(h) \quad = \quad \mathbf{E}_{XY \sim \mathcal{D}} \left\{ \mathbf{I}_{\{h(X) \neq Y\}} \right\} , \tag{2}$$

where $\mathbf{I}_{\{\cdot\}}$ 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

$$\hat{L}(h; T) \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{h(\mathbf{x}_i) \neq y_i\}} . \tag{3}$$

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 $f : \mathcal{X} \to \mathbb{R}$ 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 $h_i : \mathcal{X} \to \mathcal{Y}$, where $0 \leq i \leq N$, is constructed from f:

$$h_i(\mathbf{x}) \quad = \quad \theta(f(\mathbf{x}) - b_i) ,$$

where $\theta(\cdot)$ is as defined by Eq. (1) and

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

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:

$$tpr_i = \frac{\text{number of positive examples in } T \text{ classified correctly by } h_i}{\text{total number of positive examples in } T} \,,$$

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

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 (Cortes and Mohri, 2004; Rosset, 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 (Cortes and Mohri, 2004):

$$\hat{A}(f;T) = \frac{1}{mn} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}_j)\}} + \frac{1}{2} \mathbf{I}_{\{f(\mathbf{x}_i) = f(\mathbf{x}_j)\}} \,. \tag{4}$$

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, assuming that ties are broken uniformly at random.[2]

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 $h_i(\mathbf{x}) = \theta(f_i(\mathbf{x})), i = 1, 2$, 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

2. In (Cortes and Mohri, 2004), a slightly simpler form of the Wilcoxon-Mann-Whitney statistic is used, which does not account for ties.

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

form $h(\mathbf{x}) = \theta(f_h(\mathbf{x}))$ and the AUC value of the corresponding ranking function f_h with respect to a given data sequence was studied in detail by Cortes and Mohri (2004). In particular, they showed 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 are important for machine learning practice. 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 part of the question concerns the large deviation behaviour of the AUC; the second part concerns its uniform convergence behaviour. Both are addressed in this paper.

We start by defining 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 as the growth function) in uniform convergence results for the classification error rate. 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

shows that the bound provided by our result can be considerably tighter. We conclude with a summary and some open questions 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 use an underline to denote a sequence, *e.g.*, $\underline{y} \in \mathcal{Y}^N$ to denote a sequence of elements in $\mathcal{Y}$. 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 probabilities, *e.g.*, $\mathbf{E}_{XY} \equiv \mathbf{E}_{XY \sim \mathcal{D}}$. As a final note of convention, we use $T \in (\mathcal{X} \times \mathcal{Y})^N$ to denote a general data sequence (*e.g.*, an independent test sequence), and $S \in (\mathcal{X} \times \mathcal{Y})^M$ to denote a training sequence.

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:*

$$A(f) \;=\; \mathbf{E}_{X \sim \mathcal{D}_{+1}, X' \sim \mathcal{D}_{-1}} \left\{ \mathbf{I}_{\{f(X) > f(X')\}} + \frac{1}{2} \mathbf{I}_{\{f(X) = f(X')\}} \right\} . \tag{5}$$

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 $\mathcal{D}_{-1}$, assuming that ties are broken uniformly at random; the quantity $A(f)$ 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 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:

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 positive labels in $\underline{y}$ and the number n of negative 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 $\mathcal{D}_{+1}^m \times \mathcal{D}_{-1}^n$.

Lemma 2 *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*

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

Proof Let m be the number of positive labels in $\underline{y}$, and n the number of negative labels in $\underline{y}$. Then from the definition of empirical AUC (Eq. (4)) and linearity of expectation, we have

$$\begin{aligned}
\mathbf{E}_{T_X|T_Y=\underline{y}} \left\{ \hat{A}(f;T) \right\} &= \frac{1}{mn} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \mathbf{E}_{X_i \sim \mathcal{D}_{+1}, X_j \sim \mathcal{D}_{-1}} \left\{ \mathbf{I}_{\{f(X_i)>f(X_j)\}} + \frac{1}{2} \mathbf{I}_{\{f(X_i)=f(X_j)\}} \right\} \\
&= \frac{1}{mn} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} A(f) \\
&= A(f) \,.
\end{aligned}$$

■

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 3 (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 $\epsilon > 0$,

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

Note that when $X_1, \ldots, X_N$ are independent bounded random variables with $X_k \in [a_k, b_k]$ with probability one, and $\phi(X_1, \ldots, X_N) = \sum_{k=1}^N X_k$, McDiarmid's inequality (with $c_k = b_k - a_k$) reduces to Hoeffding's inequality. Next we define the following quantity which appears in several of our results:

Definition 4 (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:*

$$\rho(\underline{y}) = \frac{1}{N} \sum_{\{i:y_i=+1\}} 1 \,. \tag{6}$$

The following is the main result of this section:

Theorem 5 *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 $N \in \mathbb{N}$. Let m be the number of positive labels in $\underline{y}$, and $n = N - m$ the number of negative labels in $\underline{y}$. Then for any $\epsilon > 0$,*

$$\begin{aligned} \mathbf{P}_{T_X|T_Y=\underline{y}} \left\{ \left| \hat{A}(f;T) - A(f) \right| \geq \epsilon \right\} &\leq 2e^{-2mn\epsilon^2/(m+n)} \\ &= 2e^{-2\rho(\underline{y})(1-\rho(\underline{y}))N\epsilon^2} \,. \end{aligned}$$

Proof 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:

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

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

$$\begin{aligned} &\left|\phi(\mathbf{x}_1, \ldots, \mathbf{x}_N) - \phi(\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{x}'_k, \mathbf{x}_{k+1} \ldots, \mathbf{x}_N)\right| \\ &= \frac{1}{mn} \left| \sum_{\{j:y_j=-1\}} \left(\left(\mathbf{I}_{\{f(\mathbf{x}_k) > f(\mathbf{x}_j)\}} + \frac{1}{2} \mathbf{I}_{\{f(\mathbf{x}_k) = f(\mathbf{x}_j)\}} \right) - \right. \right. \\ &\qquad \left. \left. \left(\mathbf{I}_{\{f(\mathbf{x}'_k) > f(\mathbf{x}_j)\}} + \frac{1}{2} \mathbf{I}_{\{f(\mathbf{x}'_k) = f(\mathbf{x}_j)\}} \right) \right) \right| \\ &\leq \frac{1}{mn} n \\ &= \frac{1}{m} \,. \end{aligned}$$

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

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

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 $\epsilon > 0$,

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

The result follows from Lemma 2. ∎

We note that the result of Theorem 5 can be strengthened so that the conditioning is only on the numbers m and n of positive and negative labels, and not on the specific label vector $\underline{y}$. From Theorem 5, 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 6 *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 5 by setting $2e^{-2\rho(\underline{y})(1-\rho(\underline{y}))N\epsilon^2} = \delta$ and solving for ϵ. ■

We note that a different approach for deriving confidence intervals for the AUC has recently been taken by Cortes and Mohri (2005); in particular, their confidence intervals for the AUC are constructed from confidence intervals for the classification error rate.

Theorem 5 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 7 *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 5 by setting $2e^{-2\rho(\underline{y})(1-\rho(\underline{y}))N\epsilon^2} \leq \delta$ and solving for N. ■

The confidence interval of Corollary 6 can in fact be generalized to remove the conditioning on the label vector completely:

Theorem 8 *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

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

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

$$\begin{aligned}
\mathbf{P}_T\{\Phi(T,\delta)\} &= \mathbf{E}_T\left\{\mathbf{I}_{\Phi(T,\delta)}\right\} \\
&= \mathbf{E}_{T_Y}\left\{\mathbf{E}_{T_X|T_Y=\underline{y}}\left\{\mathbf{I}_{\Phi(T,\delta)}\right\}\right\} \\
&= \mathbf{E}_{T_Y}\left\{\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\Phi(T,\delta)\right\}\right\} \\
&\leq \mathbf{E}_{T_Y}\{\delta\} \qquad \text{(by Corollary 6)} \\
&= \delta\,.
\end{aligned}$$

■

Note that the above 'trick' works only once we have gone to a confidence interval; an attempt to generalize the bound of Theorem 5 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.

3.1 Comparison with Bounds from Statistical Literature

The AUC, in the form of the Wilcoxon-Mann-Whitney statistic, has been studied extensively in the statistical literature. In particular, Lehmann (1975) derives an exact expression for the variance of the Wilcoxon-Mann-Whitney statistic which can be used to obtain large deviation bounds for the AUC. Below we compare the large deviation bound we have derived above with these bounds obtainable from the statistical literature. We note that the expression derived by Lehmann (1975) is for a simpler form of the Wilcoxon-Mann-Whitney statistic that does not account for ties; therefore, in this section we assume the AUC and the expected ranking accuracy are defined without the terms that account for ties (the large deviation result we have derived above applies also in this setting).

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 $N \in \mathbb{N}$. Let m be the number of positive labels in $\underline{y}$, and $n = N - m$ the number of negative labels in $\underline{y}$. Then the variance of the AUC of f is given by the following expression (Lehmann, 1975):

$$\begin{aligned}
\sigma_A^2 &= \mathbf{Var}_{T_X|T_Y=\underline{y}}\left\{\hat{A}(f;T)\right\} \\
&= \frac{A(f)(1-A(f)) + (m-1)(p_1 - A(f)^2) + (n-1)(p_2 - A(f)^2)}{mn}\,, \qquad (7)
\end{aligned}$$

where

$$p_1 = \mathbf{P}_{X_1^+, X_2^+ \sim \mathcal{D}_{+1}, X_1^- \sim \mathcal{D}_{-1}}\left\{\left\{f(X_1^+) > f(X_1^-)\right\} \cap \left\{f(X_2^+) > f(X_1^-)\right\}\right\} \qquad (8)$$

$$p_2 = \mathbf{P}_{X_1^+ \sim \mathcal{D}_{+1}, X_1^-, X_2^- \sim \mathcal{D}_{-1}}\left\{\left\{f(X_1^+) > f(X_1^-)\right\} \cap \left\{f(X_1^+) > f(X_2^-)\right\}\right\}. \qquad (9)$$

Next we recall the following classical inequality:

Theorem 9 (Chebyshev's inequality) *Let X be a random variable. Then for any $\epsilon > 0$,*

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

The expression for the variance σ_A^2 of the AUC can be used with Chebyshev's inequality to give the following bound: for any $\epsilon > 0$,

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

This leads to the following confidence interval: for any $0 < \delta \leq 1$,

$$\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \frac{\sigma_A}{\sqrt{\delta}}\right\} \quad \leq \quad \delta\,. \tag{11}$$

It has been established that the AUC follows an asymptotically normal distribution. Therefore, for large N, one can use a normal approximation to obtain a tighter bound:

$$\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \epsilon\right\} \quad \leq \quad 2(1 - \Phi(\epsilon/\sigma_A))\,, \tag{12}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function given by $\Phi(u) = \int_0^u e^{-z^2/2}\,dz/\sqrt{2\pi}$. The resulting confidence interval is given by

$$\mathbf{P}_{T_X|T_Y=\underline{y}}\left\{\left|\hat{A}(f;T) - A(f)\right| \geq \sigma_A \Phi^{-1}(1 - \delta/2)\right\} \quad \leq \quad \delta\,. \tag{13}$$

The quantities p_1 and p_2 that appear in the expression for σ_A^2 in Eq. (7) depend on the underlying distributions $\mathcal{D}_{+1}$ and $\mathcal{D}_{-1}$; for example, Hanley and McNeil (1982) derive expressions for p_1 and p_2 in the case when the scores $f(X^+)$ assigned to positive instances X^+ and the scores $f(X^-)$ assigned to negative instances X^- both follow negative exponential distributions. Distribution-independent bounds can be obtained by using the fact that the variance σ_A^2 is at most (Cortes and Mohri, 2005; Dantzig, 1915; Birnbaum and Klose, 1957)

$$\sigma_{\max}^2 \quad = \quad \frac{A(f)(1 - A(f))}{\min(m,n)} \quad \leq \quad \frac{1}{4\min(m,n)}\,. \tag{14}$$

A comparison of the resulting bounds with the large deviation bound we have derived above using McDiarmid's inequality is shown in Figure 1. The McDiarmid bound is tighter than the bound obtained using Chebyshev's inequality. It is looser than the bound obtained using the normal approximation; however, since the normal approximation is valid only for large N, for smaller values of N the McDiarmid bound is safer.

Of course, it should be noted that this comparison holds only in the distribution-free setting. In practice, depending on the underlying distribution, the actual variance of the AUC may be much smaller than $\sigma_{\max}^2$; indeed, in the best case, the variance could be as small as

$$\sigma_{\min}^2 \quad = \quad \frac{A(f)(1 - A(f))}{mn} \quad \leq \quad \frac{1}{4mn}\,. \tag{15}$$

Therefore, if one can estimate the variance of the AUC reliably, it may be possible to obtain tighter bounds using Eqs. (10) and (12).

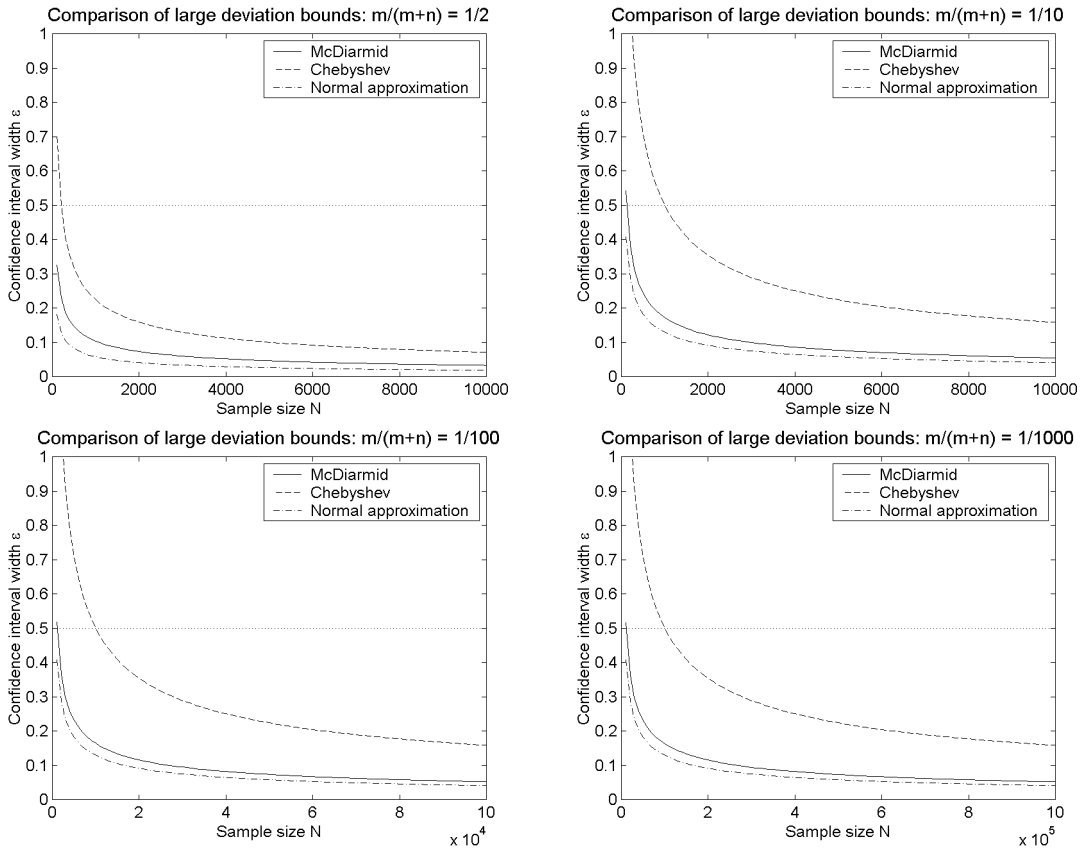

Figure 1: A comparison of our large deviation bound, derived using McDiarmid's inequality, with large deviation bounds obtainable from the statistical literature (see Section 3.1). The plots are for $\delta = 0.01$ and show how the confidence interval size ϵ given by the different bounds varies with the sample size $N = m + n$, for various values of $m/(m + n)$.

3.2 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 a similar large deviation bound 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 derivation arises from the fact that the empirical AUC, 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 for the classification error rate obtained via Hoeffding's inequality states that for a fixed classification function $h : \mathcal{X} \to \mathcal{Y}$ and for any $N \in \mathbb{N}$ and any $\epsilon > 0$,

$$\mathbf{P}_{T \sim \mathcal{D}^N} \left\{ \left| \hat{L}(h;T) - L(h) \right| \geq \epsilon \right\} \quad \leq \quad 2e^{-2N\epsilon^2} . \tag{16}$$

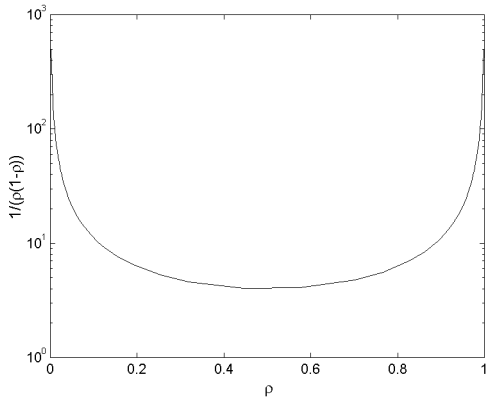

Figure 2: The test sample size bound for the AUC, for positive skew $\rho \equiv \rho(\underline{y})$ for some label sequence $\underline{y}$, is larger than the corresponding test sample size bound for the error rate by a factor of $1/(\rho(1-\rho))$ (see text for discussion).

Comparing Eq. (16) to the bound of Theorem 5, 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).

Again, it should be noted that the above conclusion holds only in the distribution-free setting. Indeed, the variance σ_L^2 of the error rate (which follows a binomial distribution) is given by

$$\sigma_L^2 \;=\; \mathbf{Var}_{T\sim\mathcal{D}^N}\left\{\hat{L}(h;T)\right\} \;=\; \frac{L(h)(1-L(h))}{N} \;\leq\; \frac{1}{4N}\,. \tag{17}$$

Comparing to Eqs. (14) and (15), we see that although this is smaller than the worst-case variance of the AUC, in the best case, the variance of the AUC can be considerably smaller, leading to a tighter bound for the AUC and therefore a smaller sufficient test sample size.

3.3 Bound for Learned Ranking Functions Chosen from Finite Function Classes

The large deviation result of Theorem 5 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 10 *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

$$\begin{aligned}
&\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\left|\hat{A}(f_S;S) - A(f_S)\right| \geq \epsilon\right\} \\
&\leq \quad \mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\max_{f\in\mathcal{F}}\left|\hat{A}(f;S) - A(f)\right| \geq \epsilon\right\} \\
&\leq \quad \sum_{f\in\mathcal{F}} \mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\left|\hat{A}(f;S) - A(f)\right| \geq \epsilon\right\} \qquad \text{(by the union bound)} \\
&\leq \quad 2|\mathcal{F}|e^{-2\rho(\underline{y})(1-\rho(\underline{y}))M\epsilon^2} \qquad \text{(by Theorem 5)}\,.
\end{aligned}$$

■

As before, we can derive from Theorem 10 expressions for confidence intervals and sufficient training sample size; we give these below without proof:

Corollary 11 *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 12 *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

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

Theorem 13 *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 $M \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

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

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 $\epsilon > 0$,

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

Therefore, to bound probabilities of the form on the left hand side above, 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 14 (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 $\{0, \frac{1}{2}, 1\}^{m \times n}$ whose (i, j)-th element is given by*

$$\left[\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')\right]_{ij} \quad = \quad \mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}'_j)\}} + \frac{1}{2}\mathbf{I}_{\{f(\mathbf{x}_i) = f(\mathbf{x}'_j)\}} \tag{18}$$

for all $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$.

Definition 15 (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)-th bipartite rank-shatter coefficient *of $\mathcal{F}$, denoted by $r(\mathcal{F}, m, n)$, as follows:*

$$r(\mathcal{F}, m, n) \quad = \quad \max_{\underline{\mathbf{x}} \in \mathcal{X}^m, \underline{\mathbf{x}}' \in \mathcal{X}^n} \left|\left\{\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}') \mid f \in \mathcal{F}\right\}\right|. \tag{19}$$

Clearly, for finite $\mathcal{F}$, we have $r(\mathcal{F}, m, n) \leq |\mathcal{F}|$ for all m, n. In general, $r(\mathcal{F}, m, n) \leq 3^{mn}$ for all m, n. In fact, not all 3^{mn} matrices in $\{0, \frac{1}{2}, 1\}^{m \times n}$ can be realized as bipartite rank matrices. Therefore, we have

$$r(\mathcal{F}, m, n) \leq \psi(m, n),$$

where $\psi(m, n)$ is the number of matrices in $\{0, \frac{1}{2}, 1\}^{m \times n}$ that can be realized as a bipartite rank matrix. The number $\psi(m, n)$ can be characterized in the following ways:

Theorem 16 *Let $\psi(m, n)$ be the number of matrices in $\{0, \frac{1}{2}, 1\}^{m \times n}$ that can be realized as a bipartite rank matrix $\mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')$ for some $f : \mathcal{X} \to \mathbb{R}$, $\underline{\mathbf{x}} \in \mathcal{X}^m$, $\underline{\mathbf{x}}' \in \mathcal{X}^n$. Then*

Table 2: Sub-matrices that cannot appear in a bipartite rank matrix.

$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} \end{bmatrix}$
$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ \frac{1}{2} & 0 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 1 \\ 1 & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & \frac{1}{2} \\ 1 & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ 1 & \frac{1}{2} \end{bmatrix}$

1. *$\psi(m,n)$ is equal to the number of complete mixed acyclic (m,n)-bipartite graphs (where a mixed graph is one which may contain both directed and undirected edges, and where we define a cycle in such a graph as a cycle that contains at least one directed edge and in which all directed edges have the same directionality along the cycle).*
2. *$\psi(m,n)$ is equal to the number of matrices in $\{0,\frac{1}{2},1\}^{m\times n}$ that do not contain a sub-matrix of any of the forms shown in Table 2.*

Proof
Part 1. Let $\mathcal{G}(m,n)$ denote the set of all complete mixed (m,n)-bipartite graphs. Clearly, $|\mathcal{G}(m,n)| = 3^{mn}$, since there are mn edges and three possibilities for each edge. Let $V = \{v_1,\ldots,v_m\}$, $V' = \{v'_1,\ldots,v'_n\}$ be sets of m and n vertices respectively, and for any matrix $\mathbf{B} = [b_{ij}] \in \{0,\frac{1}{2},1\}^{m\times n}$, let $E(\mathbf{B})$ denote the set of edges between V and V' given by $E(\mathbf{B}) = \{(v_i \leftarrow v'_j) \mid b_{ij} = 1\} \cup \{(v_i \rightarrow v'_j) \mid b_{ij} = 0\} \cup \{(v_i - v'_j) \mid b_{ij} = \frac{1}{2}\}$. Define the mapping $G : \{0,\frac{1}{2},1\}^{m\times n} \rightarrow \mathcal{G}(m,n)$ as follows:

$$G(\mathbf{B}) = (V \cup V', E(\mathbf{B}))\,.$$

Then clearly, G is a bijection that puts the sets $\{0,\frac{1}{2},1\}^{m\times n}$ and $\mathcal{G}(m,n)$ into one-to-one correspondence. We show that a matrix $\mathbf{B} \in \{0,\frac{1}{2},1\}^{m\times n}$ can be realized as a bipartite rank matrix if and only if the corresponding bipartite graph $G(\mathbf{B}) \in \mathcal{G}(m,n)$ is acyclic.

First suppose $\mathbf{B} = \mathbf{B}_f(\underline{\mathbf{x}},\underline{\mathbf{x}}')$ for some $f : \mathcal{X}\rightarrow\mathbb{R}$, $\underline{\mathbf{x}} \in \mathcal{X}^m$, $\underline{\mathbf{x}}' \in \mathcal{X}^n$, and let if possible $G(\mathbf{B})$ contain a cycle, say

$$(v_{i_1} \leftarrow v'_{j_1} - v_{i_2} - v'_{j_2} - \ \ldots\ - v_{i_k} - v'_{j_k} - v_{i_1})\,.$$

Then, from the definition of a bipartite rank matrix, we get

$$f(\mathbf{x}_{i_1}) < f(\mathbf{x}'_{j_1}) = f(\mathbf{x}_{i_2}) = f(\mathbf{x}'_{j_2}) = \ldots = f(\mathbf{x}_{i_k}) = f(\mathbf{x}'_{j_k}) = f(\mathbf{x}_{i_1})\,,$$

which is a contradiction.

To prove the other direction, let $\mathbf{B} \in \{0,\frac{1}{2},1\}^{m\times n}$ be such that $G(\mathbf{B})$ is acyclic. Let $G'(\mathbf{B})$ denote the directed graph obtained by collapsing together vertices in $G(\mathbf{B})$ that are connected by an undirected edge. Then it is easily verified that $G'(\mathbf{B})$ does not contain any directed cycles, and therefore there exists a complete order on the vertices of $G'(\mathbf{B})$ that is consistent with the partial order defined by the edges of $G'(\mathbf{B})$ (topological sorting; see, for example, Cormen et al., 2001, Section 22.4). This implies a unique order on the vertices of $G(\mathbf{B})$ (in which vertices connected by undirected edges are assigned the same position in the ordering). For any $\underline{\mathbf{x}} \in \mathcal{X}^m$, $\underline{\mathbf{x}}' \in \mathcal{X}^n$, identifying $\underline{\mathbf{x}},\underline{\mathbf{x}}'$ with the vertex sets V, V' of $G(\mathbf{B})$ therefore gives a unique order on $\mathbf{x}_1,\ldots,\mathbf{x}_m,\mathbf{x}'_1,\ldots,\mathbf{x}'_n$. It can be verified

that defining $f : \mathcal{X} \to \mathbb{R}$ such that it respects this order then gives $\mathbf{B} = \mathbf{B}_f(\underline{\mathbf{x}}, \underline{\mathbf{x}}')$.

Part 2. Consider again the bijection $G : \{0, \frac{1}{2}, 1\}^{m \times n} \to \mathcal{G}(m, n)$ defined in Part 1 above. We show that a matrix $\mathbf{B} \in \{0, \frac{1}{2}, 1\}^{m \times n}$ does not contain a sub-matrix of any of the forms shown in Table 2 if and only if the corresponding bipartite graph $G(\mathbf{B}) \in \mathcal{G}(m, n)$ is acyclic; the desired result then follows by Part 1 of the theorem.

We first note that the condition that $\mathbf{B} \in \{0, \frac{1}{2}, 1\}^{m \times n}$ not contain a sub-matrix of any of the forms shown in Table 2 is equivalent to the condition that the corresponding mixed (m, n)-bipartite graph $G(\mathbf{B}) \in \mathcal{G}(m, n)$ not contain any 4-cycles.

Now, to prove the first direction, let $\mathbf{B} \in \{0, \frac{1}{2}, 1\}^{m \times n}$ not contain a sub-matrix of any of the forms shown in Table 2. As noted above, this means $G(\mathbf{B})$ does not contain any 4-cycles. Let, if possible, $G(\mathbf{B})$ contain a cycle of length $2k$, say

$$(v_{i_1} \leftarrow v'_{j_1} \text{ — } v_{i_2} \text{ — } v'_{j_2} \text{ — } \ldots \text{ — } v_{i_k} \text{ — } v'_{j_k} \text{ — } v_{i_1}) .$$

Now consider v_{i_1}, v'_{j_2}. Since $G(\mathbf{B})$ is a complete bipartite graph, there must be an edge between these vertices. If $G(\mathbf{B})$ contained the edge $(v_{i_1} \rightarrow v'_{j_2})$, it would contain the 4-cycle

$$(v_{i_1} \leftarrow v'_{j_1} \text{ — } v_{i_2} \text{ — } v'_{j_2} \leftarrow v_{i_1}) ,$$

which would be a contradiction. Similarly, if $G(\mathbf{B})$ contained the edge $(v_{i_1} \text{ — } v'_{j_2})$, it would contain the 4-cycle

$$(v_{i_1} \leftarrow v'_{j_1} \text{ — } v_{i_2} \text{ — } v'_{j_2} \text{ — } v_{i_1}) ,$$

which would again be a contradiction. Therefore, $G(\mathbf{B})$ must contain the edge $(v_{i_1} \leftarrow v'_{j_2})$. However, this means $G(\mathbf{B})$ must contain a $2(k-1)$-cycle, namely,

$$(v_{i_1} \leftarrow v'_{j_2} \text{ — } v_{i_3} \text{ — } v'_{j_3} \text{ — } \ldots \text{ — } v_{i_k} \text{ — } v'_{j_k} \text{ — } v_{i_1}) .$$

By a recursive argument, we eventually get that $G(\mathbf{B})$ must contain a 4-cycle, which is a contradiction.

To prove the other direction, let $\mathbf{B} \in \{0, \frac{1}{2}, 1\}^{m \times n}$ be such that $G(\mathbf{B})$ is acyclic. Then it follows trivially that $G(\mathbf{B})$ does not contain a 4-cycle, and therefore, by the above observation, $\mathbf{B}$ does not contain a sub-matrix of any of the forms shown in Table 2. ■

We discuss further properties of the bipartite rank-shatter coefficients in Section 4.3; we first present below our uniform convergence result in terms of these coefficients.

4.2 Uniform Convergence Bound

The following is the main result of this section:

Theorem 17 *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 positive labels in $\underline{y}$, and $n = M - m$ the number of negative labels in $\underline{y}$. Then for any $\epsilon > 0$,*

$$\begin{aligned} \mathbf{P}_{S_X | S_Y = \underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - A(f) \right| \geq \epsilon \right\} &\leq 4 \cdot r(\mathcal{F}, 2m, 2n) \cdot e^{-mn\epsilon^2/8(m+n)} \\ &= 4 \cdot r\left(\mathcal{F},\, 2\rho(\underline{y})M,\, 2(1-\rho(\underline{y}))M\right) \cdot e^{-\rho(\underline{y})(1-\rho(\underline{y}))M\epsilon^2/8} , \end{aligned}$$

where $\rho(\underline{y})$ denotes the positive skew of $\underline{y}$ defined in Eq. (6).

The proof is adapted from proofs of uniform convergence for the classification error rate (see, for example, Anthony and Bartlett, 1999; Devroye et al., 1996). The main difference is that since the AUC cannot be expressed as a sum of independent random variables, more powerful inequalities are required. In particular, a result of Devroye (1991) is required to bound the variance of the AUC that appears after an application of Chebyshev's inequality; the application of this result to the AUC requires the same reasoning that was used to apply McDiarmid's inequality in deriving the large deviation result of Theorem 5. Similarly, McDiarmid's inequality is required in the final step of the proof where Hoeffding's inequality sufficed in the case of classification. Complete details of the proof are given in Appendix A.

As in the case of the large deviation bound of Section 3, we note that the result of Theorem 17 can be strengthened so that the conditioning is only on the numbers m and n of positive and negative labels, and not on the specific label vector $\underline{y}$. From Theorem 17, we can derive a confidence interval interpretation of the bound as follows:

Corollary 18 *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 positive labels in $\underline{y}$, and $n = M - m$ the number of negative labels in $\underline{y}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f;S) - A(f) \right| \geq \sqrt{\frac{8(m+n)\left(\ln r(\mathcal{F}, 2m, 2n) + \ln\left(\frac{4}{\delta}\right)\right)}{mn}} \right\} \quad \leq \quad \delta \,.$$

Proof This follows directly from Theorem 17 by setting $4 \cdot r(\mathcal{F}, 2m, 2n) \cdot e^{-mn\epsilon^2/8(m+n)} = \delta$ and solving for ϵ. ■

Again, as in the case of the large deviation bound, the confidence interval above can be generalized to remove the conditioning on the label vector completely:

Theorem 19 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $M \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

$$\mathbf{P}_{S \sim \mathcal{D}^M} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f;S) - A(f) \right| \geq \sqrt{\frac{8\left(\ln r\left(\mathcal{F},\, 2\rho(S_Y)M,\, 2(1-\rho(S_Y))M\right) + \ln\left(\frac{4}{\delta}\right)\right)}{\rho(S_Y)(1-\rho(S_Y))M}} \right\} \quad \leq \quad \delta \,.$$

4.3 Properties of Bipartite Rank-Shatter Coefficients

As discussed in Section 4.1, we have $r(\mathcal{F}, m, n) \leq \psi(m, n)$, where $\psi(m, n)$ is the number of matrices in $\{0, \frac{1}{2}, 1\}^{m \times n}$ that can be realized as a bipartite rank matrix. The number $\psi(m, n)$ is strictly smaller than 3^{mn}; indeed, $\psi(m, n) = O(e^{(m+n)(\ln(m+n)+1)})$. (To see this, note that the number of distinct bipartite rank matrices of size $m \times n$ is bounded above by the total number of permutations of $(m + n)$ objects, allowing for objects to be placed at the same position. This number is equal to $(m + n)!\, 2^{(m+n-1)} = O(e^{(m+n)(\ln(m+n)+1)})$.) Nevertheless, $\psi(m, n)$ is still very large; in particular, $\psi(m, n) \geq 3^{\max(m,n)}$. (To see this, note that choosing any column vector in $\{0, \frac{1}{2}, 1\}^m$ and replicating it along the n columns or

choosing any row vector in $\{0, \frac{1}{2}, 1\}^n$ and replicating it along the m rows results in a matrix that does not contain a sub-matrix of any of the forms shown in Table 2. The conclusion then follows from Theorem 16 (Part 2).)

For the bound of Theorem 17 to be meaningful, one needs an upper bound on $r(\mathcal{F}, m, n)$ that is at least slightly smaller than $e^{mn/8(m+n)}$. Below we provide one method for deriving upper bounds on $r(\mathcal{F}, m, n)$; taking $\mathcal{Y}^* = \{-1, 0, +1\}$, we extend slightly the standard VC-dimension related shatter coefficients studied in binary classification to $\mathcal{Y}^*$-valued function classes, and then derive an upper bound on the bipartite rank-shatter coefficients $r(\mathcal{F}, m, n)$ of a class of ranking functions $\mathcal{F}$ in terms of the shatter coefficients of a class of $\mathcal{Y}^*$-valued functions derived from $\mathcal{F}$.

Definition 20 (Shatter coefficient) *Let $\mathcal{Y}^* = \{-1, 0, +1\}$, and let $\mathcal{H}$ be a class of $\mathcal{Y}^*$-valued functions on $\mathcal{X}$. Let $N \in \mathbb{N}$. Define the N-th* shatter coefficient *of $\mathcal{H}$, denoted by $s(\mathcal{H}, N)$, as follows:*

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

Clearly, $s(\mathcal{H}, N) \leq 3^N$ for all N. Next we define a series of $\mathcal{Y}^*$-valued function classes derived from a given ranking function class. Only the second function class is used in this section; the other two are needed in Section 4.4. Note that we take

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

Definition 21 (Function classes) *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. Define the following classes of $\mathcal{Y}^*$-valued functions derived from $\mathcal{F}$:*

$$1.\quad \bar{\mathcal{F}} \;=\; \{\bar{f} : \mathcal{X} \to \mathcal{Y}^* \mid \bar{f}(\mathbf{x}) = \mathrm{sign}(f(\mathbf{x})) \text{ for some } f \in \mathcal{F}\} \tag{20}$$

$$2.\quad \tilde{\mathcal{F}} \;=\; \{\tilde{f} : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}^* \mid \tilde{f}(\mathbf{x}, \mathbf{x}') = \mathrm{sign}(f(\mathbf{x}) - f(\mathbf{x}')) \text{ for some } f \in \mathcal{F}\} \tag{21}$$

$$3.\quad \check{\mathcal{F}} \;=\; \{\check{f}_{\mathbf{z}} : \mathcal{X} \to \mathcal{Y}^* \mid \check{f}_{\mathbf{z}}(\mathbf{x}) = \mathrm{sign}(f(\mathbf{x}) - f(\mathbf{z})) \text{ for some } f \in \mathcal{F}, \mathbf{z} \in \mathcal{X}\} \tag{22}$$

The following result gives an upper bound on the bipartite rank-shatter coefficients of a class of ranking functions $\mathcal{F}$ in terms of the standard shatter coefficients of $\tilde{\mathcal{F}}$:

Theorem 22 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $\tilde{\mathcal{F}}$ be the class of $\mathcal{Y}^*$-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[4]

$$r(\mathcal{F}, m, n) \;=\; \max_{\underline{\mathbf{x}} \in \mathcal{X}^m, \underline{\mathbf{x}}' \in \mathcal{X}^n} \left| \left\{ \left[\mathbf{I}_{\{f(\mathbf{x}_i) > f(\mathbf{x}'_j)\}} + \frac{1}{2} \mathbf{I}_{\{f(\mathbf{x}_i) = f(\mathbf{x}'_j)\}} \right] \;\middle|\; f \in \mathcal{F} \right\} \right|$$

4. We use the notation $[a_{ij}]$ to denote a matrix whose $(i,j)^{\text{th}}$ element is a_{ij}. The dimensions of such a matrix should be clear from context.

$$
\begin{aligned}
&= \max_{\underline{\mathbf{x}}\in\mathcal{X}^m,\underline{\mathbf{x}}'\in\mathcal{X}^n} \left|\left\{\left[\mathbf{I}_{\{\tilde{f}(\mathbf{x}_i,\mathbf{x}'_j)=+1\}} + \frac{1}{2}\mathbf{I}_{\{\tilde{f}(\mathbf{x}_i,\mathbf{x}'_j)=0\}}\right] \;\middle|\; \tilde{f}\in\tilde{\mathcal{F}}\right\}\right| \\
&= \max_{\underline{\mathbf{x}}\in\mathcal{X}^m,\underline{\mathbf{x}}'\in\mathcal{X}^n} \left|\left\{\left[\tilde{f}(\mathbf{x}_i,\mathbf{x}'_j)\right] \;\middle|\; \tilde{f}\in\tilde{\mathcal{F}}\right\}\right| \\
&\leq \max_{\mathbf{X},\mathbf{X}'\in\mathcal{X}^{m\times n}} \left|\left\{\left[\tilde{f}(\mathbf{x}_{ij},\mathbf{x}'_{ij})\right] \;\middle|\; \tilde{f}\in\tilde{\mathcal{F}}\right\}\right| \\
&= \max_{\underline{\mathbf{x}},\underline{\mathbf{x}'}\in\mathcal{X}^{mn}} \left|\left\{\left(\tilde{f}(\mathbf{x}_1,\mathbf{x}'_1),\ldots,\tilde{f}(\mathbf{x}_{mn},\mathbf{x}'_{mn})\right) \;\middle|\; \tilde{f}\in\tilde{\mathcal{F}}\right\}\right| \\
&= s(\tilde{\mathcal{F}},mn)\,.
\end{aligned}
$$

■

Below we make use of the above result to derive polynomial upper bounds on the bipartite rank-shatter coefficients for linear and higher-order polynomial ranking functions. We note that the same method can be used to establish similar upper bounds for other algebraically well-behaved function classes.

Lemma 23 *For $d\in\mathbb{N}$, let $\mathcal{F}_{\mathrm{lin}(d)}$ denote the class of linear ranking functions on $\mathbb{R}^d$:*

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

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

$$s(\tilde{\mathcal{F}}_{\mathrm{lin}(d)},N) \leq \left(\frac{2eN}{d}\right)^d.$$

Proof We have,

$$\tilde{\mathcal{F}}_{\mathrm{lin}(d)} = \{\tilde{f}:\mathbb{R}^d\times\mathbb{R}^d\to\mathcal{Y}^* \mid \tilde{f}(\mathbf{x},\mathbf{x}')=\mathrm{sign}(\mathbf{w}\cdot(\mathbf{x}-\mathbf{x}')) \text{ for some } \mathbf{w}\in\mathbb{R}^d\}\,.$$

Let $(\mathbf{x}_1,\mathbf{x}'_1),\ldots,(\mathbf{x}_N,\mathbf{x}'_N)$ be any N points in $\mathbb{R}^d\times\mathbb{R}^d$, and consider the 'dual' weight space corresponding to $\mathbf{w}\in\mathbb{R}^d$. Each point $(\mathbf{x}_i,\mathbf{x}'_i)$ defines a hyperplane $(\mathbf{x}_i-\mathbf{x}'_i)$ in this space; the N points thus give rise to an arrangement of N hyperplanes in $\mathbb{R}^d$. It is easily seen that the number of sign patterns $(\tilde{f}(\mathbf{x}_1,\mathbf{x}'_1),\ldots,\tilde{f}(\mathbf{x}_N,\mathbf{x}'_N))$ that can be realized by functions $\tilde{f}\in\tilde{\mathcal{F}}_{\mathrm{lin}(d)}$ is equal to the total number of faces of this arrangement (Matoušek, 2002), which is at most (Buck, 1943)

$$\sum_{k=0}^{d}\sum_{i=d-k}^{d}\binom{i}{d-k}\binom{N}{i} = \sum_{i=0}^{d}2^i\binom{N}{i} \leq \left(\frac{2eN}{d}\right)^d.$$

Since the N points were arbitrary, the result follows. ■

Theorem 24 *For $d\in\mathbb{N}$, let $\mathcal{F}_{\mathrm{lin}(d)}$ denote the class of linear ranking functions on $\mathbb{R}^d$ (defined in Lemma 23 above). Then for all $m,n\in\mathbb{N}$,*

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

Proof This follows immediately from Lemma 23 and Theorem 22. ■

Lemma 25 *For $d, q \in \mathbb{N}$, let $\mathcal{F}_{\mathrm{poly}(d,q)}$ denote the class of polynomial ranking functions on $\mathbb{R}^d$ with degree less than or equal to q. Then for all $N \in \mathbb{N}$,*

$$s(\tilde{\mathcal{F}}_{\mathrm{poly}(d,q)}, N) \;\leq\; \left(\frac{2eN}{C(d,q)}\right)^{C(d,q)},$$

where

$$C(d,q) \;=\; \sum_{i=1}^{q}\left(\binom{d}{i}\sum_{j=1}^{q}\binom{j-1}{i-1}\right). \tag{23}$$

Proof We have,

$$\tilde{\mathcal{F}}_{\mathrm{poly}(d,q)} \;=\; \{\tilde{f} : \mathbb{R}^d \times \mathbb{R}^d \to \mathcal{Y}^* \mid \tilde{f}(\mathbf{x}, \mathbf{x}') = \mathrm{sign}(f(\mathbf{x}) - f(\mathbf{x}')) \text{ for some } f \in \mathcal{F}_{\mathrm{poly}(d,q)}\}.$$

Let $(\mathbf{x}_1, \mathbf{x}'_1), \ldots, (\mathbf{x}_N, \mathbf{x}'_N)$ be any N points in $\mathbb{R}^d \times \mathbb{R}^d$. For any $f \in \mathcal{F}_{\mathrm{poly}(d,q)}$, $(f(\mathbf{x}) - f(\mathbf{x}'))$ is a linear combination of $C(d,q)$ basis functions of the form $(g_k(\mathbf{x}) - g_k(\mathbf{x}'))$, $1 \leq k \leq C(d,q)$, each $g_k(\mathbf{x})$ being a product of 1 to q components of $\mathbf{x}$. Denote $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_{C(d,q)}(\mathbf{x})) \in \mathbb{R}^{C(d,q)}$. Then each point $(\mathbf{x}_i, \mathbf{x}'_i)$ defines a hyperplane $(\mathbf{g}(\mathbf{x}_i) - \mathbf{g}(\mathbf{x}'_i))$ in $\mathbb{R}^{C(d,q)}$; the N points thus give rise to an arrangement of N hyperplanes in $\mathbb{R}^{C(d,q)}$. It is easily seen that the number of sign patterns $(\tilde{f}(\mathbf{x}_1, \mathbf{x}'_1), \ldots, \tilde{f}(\mathbf{x}_N, \mathbf{x}'_N))$ that can be realized by functions $\tilde{f} \in \tilde{\mathcal{F}}_{\mathrm{poly}(d,q)}$ is equal to the total number of faces of this arrangement (Matoušek, 2002), which is at most (Buck, 1943)

$$\left(\frac{2eN}{C(d,q)}\right)^{C(d,q)}.$$

Since the N points were arbitrary, the result follows. ■

Theorem 26 *For $d, q \in \mathbb{N}$, let $\mathcal{F}_{\mathrm{poly}(d,q)}$ denote the class of polynomial ranking functions on $\mathbb{R}^d$ with degree less than or equal to q. Then for all $m, n \in \mathbb{N}$,*

$$r(\mathcal{F}_{\mathrm{poly}(d,q)}, m, n) \;\leq\; \left(\frac{2emn}{C(d,q)}\right)^{C(d,q)},$$

where $C(d,q)$ is as defined in Eq. (23).

Proof This follows immediately from Lemma 25 and Theorem 22. ■

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, namely the ranking loss for the bipartite ranking problem. As pointed out by Cortes and Mohri (2004), the bipartite ranking loss is equal to one minus the AUC; the uniform convergence bound of Freund et al. (2003) therefore implies a uniform convergence bound for the AUC.[5] 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 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.[6]

Theorem 27 (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 positive labels in $\underline{y}$, and $n = M - m$ the number of negative labels in $\underline{y}$. Then for any $0 < \delta \leq 1$,*

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

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

The proof follows that of Freund et al. (2003); for completeness, we give details in Appendix B. We now compare the uniform convergence bound derived in Section 4.2 with that of Freund et al. for a simple function class for which the quantities involved in both bounds (namely, $r(\mathcal{F}, 2m, 2n)$ and $s(\check{\mathcal{F}}, 2m), s(\check{\mathcal{F}}, 2n)$) can be characterized exactly. Specifically, consider the function class $\mathcal{F}_{\mathrm{lin}(1)}$ of linear ranking functions on $\mathbb{R}$, given by

$$\mathcal{F}_{\mathrm{lin}(1)} \quad = \quad \{f : \mathbb{R}\to\mathbb{R} \mid f(x) = wx + b \text{ for some } w \in \mathbb{R}, b \in \mathbb{R}\}\,.$$

Although $\mathcal{F}_{\mathrm{lin}(1)}$ is an infinite function class, it is easy to verify that $r(\mathcal{F}_{\mathrm{lin}(1)}, m, n) = 3$ for all $m, n \in \mathbb{N}$. (To see this, note that for any set of $m + n$ distinct points in $\mathbb{R}$, one can obtain exactly three different ranking behaviours with functions in $\mathcal{F}_{\mathrm{lin}(1)}$: one by setting $w > 0$, another by setting $w < 0$, and the third by setting $w = 0$.) On the other hand, $s(\check{\mathcal{F}}_{\mathrm{lin}(1)}, N) = 4N + 1$ for all $N \geq 2$, since $\check{\mathcal{F}}_{\mathrm{lin}(1)} = \bar{\mathcal{F}}_{\mathrm{lin}(1)}$ (see Eq. (20)) and, as is easily verified, the number of sign patterns on $N \geq 2$ distinct points in $\mathbb{R}$ that can be realized by functions in $\bar{\mathcal{F}}_{\mathrm{lin}(1)}$ is $4N + 1$. We thus get from our result (Corollary 18) that

$$\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{ \sup_{f\in\mathcal{F}_{\mathrm{lin}(1)}} \left|\hat{A}(f;S) - A(f)\right| \geq \sqrt{\frac{8(m+n)\left(\ln 3 + \ln\left(\frac{4}{\delta}\right)\right)}{mn}} \right\} \quad \leq \quad \delta\,,$$

5. As in the AUC definition of (Cortes and Mohri, 2004), the ranking loss defined in (Freund et al., 2003) does not account for ties; this is easily remedied.
6. The result in (Freund et al., 2003) was stated in terms of the VC dimension, but the basic result can be stated in terms of shatter coefficients. Due to our AUC definition which accounts for ties, the standard shatter coefficients are replaced here with the extended shatter coefficients defined above for $\mathcal{Y}^*$-valued function classes.

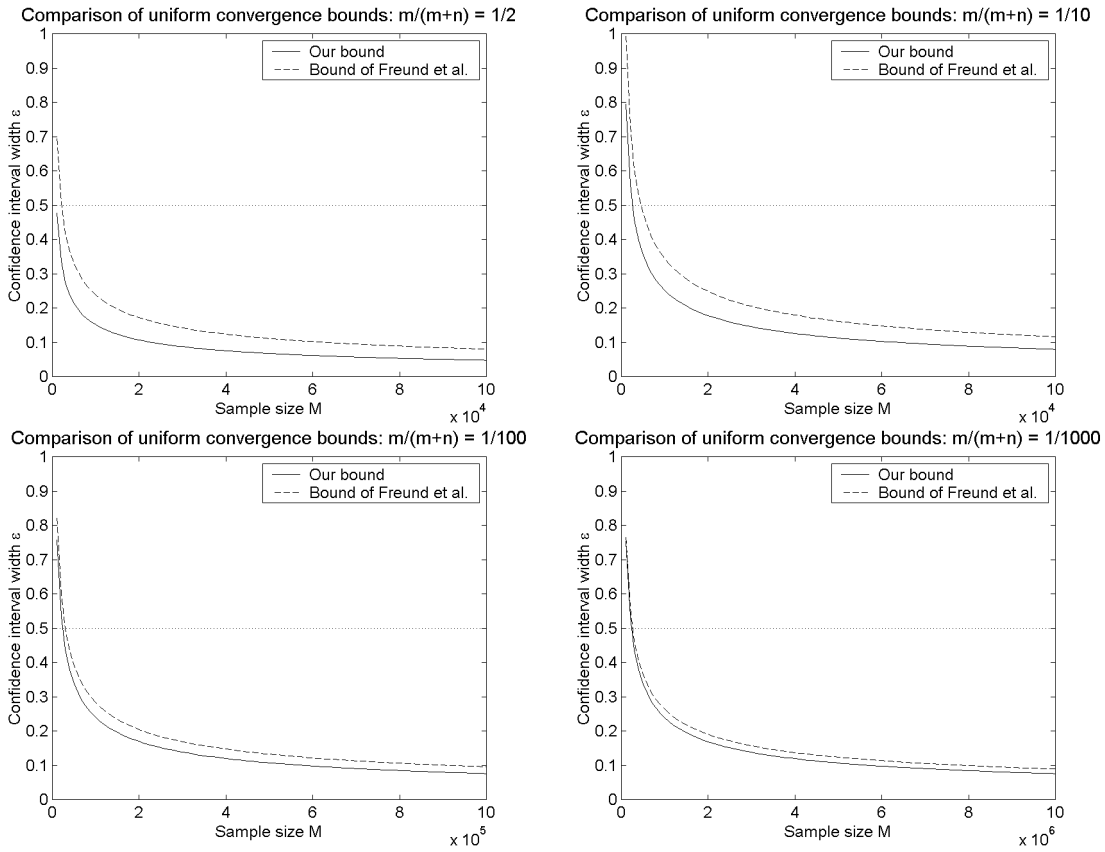

Figure 3: A comparison of our uniform convergence bound with that of Freund et al. (2003) for the class of linear ranking functions on $\mathbb{R}$. The plots are for $\delta = 0.01$ and show how the confidence interval size ϵ given by the two bounds varies with the sample size $M = m + n$, for various values of $m/(m+n)$. In all cases where the bounds are meaningful ($\epsilon < 0.5$), our bound is tighter.

and from the result of Freund et al. (Theorem 27) that

$$\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{ \sup_{f\in\mathcal{F}_{\mathrm{lin}(1)}} \left|\hat{A}(f;S) - A(f)\right| \geq 2\sqrt{\frac{\ln(8m+1)+\ln\left(\frac{12}{\delta}\right)}{m}} + 2\sqrt{\frac{\ln(8n+1)+\ln\left(\frac{12}{\delta}\right)}{n}} \right\} \;\leq\; \delta\,.$$

The above bounds are plotted in Figure 3 for $\delta = 0.01$ and various values of $m/(m+n)$. As can be seen, the bound provided by our result is considerably tighter.

Table 3: The number of positive examples m and the number of negative examples n in different training sequences used in the experiments described in Section 4.5.

Training sequence	m	n
0	50	50
1	20	80
2	10	90
3	100	100
4	40	160
5	20	180
6	150	150
7	60	240
8	30	270
9	200	200
10	80	320
11	40	360

4.5 Correctness of Functional Shape of Bound

Although our bound seems to be tighter than the previous bound of Freund et al. (2003), it is still in general loose for practical use. However, the bound can be a valuable analysis tool, as well as a useful tool for model selection, if it displays a correct functional dependency on the training sample size parameters m and n. In this section we give an empirical assessment of the correctness of the functional shape of our bound.

We generated data points in $d = 20$ dimensions ($\mathcal{X} = \mathbb{R}^{20}$) as follows. We took $\mathcal{D}_{+1}$ and $\mathcal{D}_{-1}$ to be mixtures of two 20-dimensional Gaussians each, where each of the elements of both the means and the (diagonal) covariances of the Gaussians were chosen randomly from a uniform distribution on the interval $(0, 1)$. Twelve training sequences of varying sizes were generated by drawing m points from $\mathcal{D}_{+1}$ and n points from $\mathcal{D}_{-1}$ for various values of m and n (see Table 3).[7] Similarly, a test sequence was generated by drawing 2500 points from $\mathcal{D}_{+1}$ and 2500 points from $\mathcal{D}_{-1}$. For each training sequence, a linear ranking function in $\mathcal{F}_{\mathrm{lin}(20)}$ was learned using the RankBoost algorithm of Freund et al. (2003). The training AUC of the learned ranking function, its AUC on the independent test sequence, and the lower bound on its expected ranking accuracy obtained from our uniform convergence result (using Corollary 18, at a confidence level $\delta = 0.01$) were then calculated. Since we do not have a means to characterize $r(\mathcal{F}_{\mathrm{lin}(20)}, m, n)$ exactly, we used the (loose) bound provided by Theorem 24 in calculating the lower bound on the expected accuracy. The results are shown in Figure 4. As can be seen, the functional shape of the bound is roughly in accordance with that of the test AUC, suggesting that the bound does indeed display a correct functional dependency and therefore can be useful as an analysis and model selection tool.

7. To sample points from Gaussian mixtures we made use of the NETLAB toolbox written by Ian Nabney and Christopher Bishop, available from `http://www.ncrg.aston.ac.uk/netlab/` .

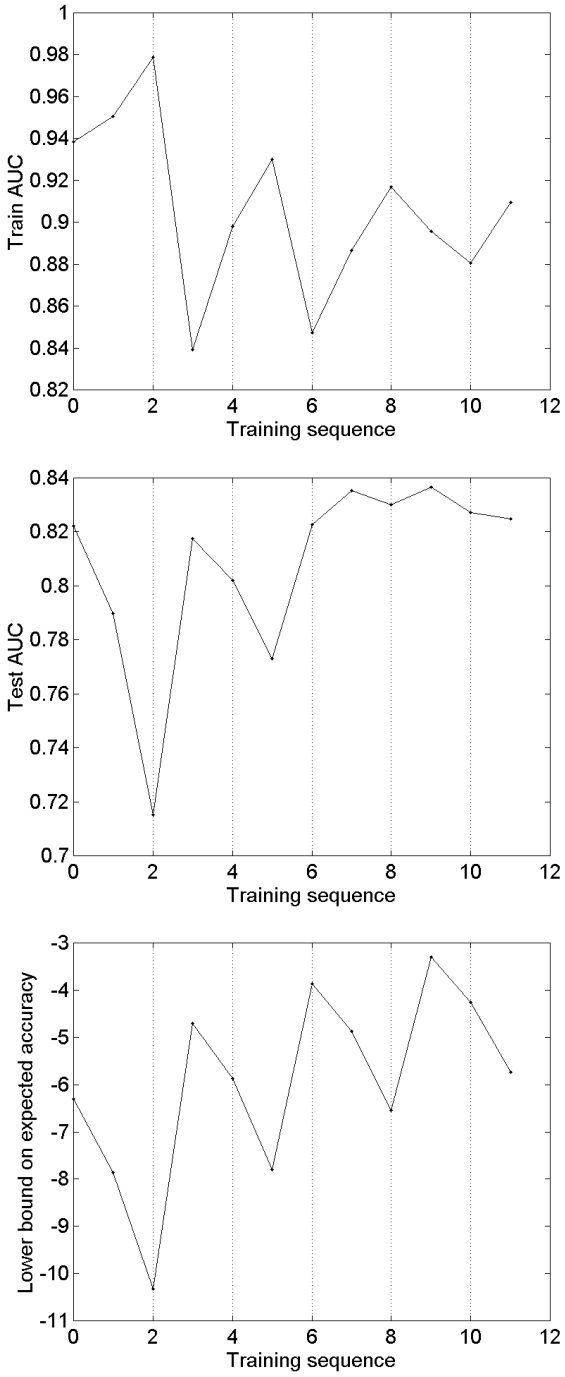

Figure 4: The training AUC (top), test AUC (middle), and lower bound on expected ranking accuracy (bottom) of linear ranking functions learned from training sequences of different sizes (see Table 3). The functional shape of the bound is roughly in accordance with that of the test AUC.

5. Conclusion and Open Questions

We have derived geralization bounds for the area under the ROC curve (AUC), a quantity used as an evaluation criterion for the bipartite ranking problem. We have derived 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. Both our bounds are distribution-free.

Our large deviation result for the AUC parallels the classical large deviation result for the classification error rate obtained via Hoeffding's inequality. A comparison with the large deviation result for the error rate suggests that, in the distribution-free setting, 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.

Our uniform convergence bound for the AUC 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 real-valued function classes and play the same role in our result as do the standard VC-dimension related shatter coefficients in uniform convergence results for the classification error rate.

For the case of linear ranking functions on $\mathbb{R}$, for which we could compute the bipartite rank-shatter coefficients exactly, we have shown that our uniform convergence bound is considerably tighter than a recent uniform convergence bound derived by Freund et al. (2003), which is expressed directly in terms of standard shatter coefficients from results for classification. This suggests that the bipartite rank-shatter coefficients we have introduced may be a more appropriate complexity measure for studying the bipartite ranking problem. However, in order to take advantage of our results, one needs to be able to characterize these coefficients for the class of ranking functions of interest. The biggest open question that arises from our study is, for what other function classes $\mathcal{F}$ can the bipartite rank-shatter coefficients $r(\mathcal{F}, m, n)$ be characterized? We have derived in Theorem 22 a general upper bound on the bipartite rank-shatter coefficients of a function class $\mathcal{F}$ in terms of the standard shatter coefficients of the function class $\tilde{\mathcal{F}}$ (see Eq. (21)); this allows us to establish a polynomial upper bound on the bipartite rank-shatter coefficients for linear and higher-order polynomial ranking functions on $\mathbb{R}^d$ and other algebraically well-behaved function classes. However, this upper bound is inherently loose (see proof of Theorem 22). Is it possible to find tighter upper bounds on $r(\mathcal{F}, m, n)$ than that given by Theorem 22?

Our study also raises several other interesting questions. First, can we establish analogous complexity measures and generalization bounds for other forms of ranking problems (*i.e.*, other than bipartite)? Second, do there exist data-dependent bounds for ranking, analogous to existing margin bounds for classification? Finally, it also remains an open question whether tighter (or alternative) generalization bounds for the AUC can be derived using different proof techniques. A possible route for deriving an alternative large deviation bound for the AUC could be via the theory of U-statistics (de la Peña 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., 2005).

Acknowledgements

We would like to thank the anonymous reviewers of our work for many useful suggestions and for pointing us to the statistical literature on ranks. We are also very grateful to an anonymous reviewer of an earlier version of part of this work for helping us identify an important mistake in our earlier results. This research was supported in part by NSF ITR grants IIS 00-85980 and IIS 00-85836 and a grant from the ONR-TRECC program.

References

Shivani Agarwal, Thore Graepel, Ralf Herbrich, and Dan Roth. A large deviation bound for the area under the ROC curve. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005a.

Shivani Agarwal, Sariel Har-Peled, and Dan Roth. A uniform convergence bound for the area under the ROC curve. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005b.

Martin Anthony and Peter Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge University Press, 1999.

Z. W. Birnbaum and O. M. Klose. Bounds for the variance of the Mann-Whitney statistic. *Annals of Mathematical Statistics*, 38, 1957.

R. C. Buck. Partition of space. *American Mathematical Monthly*, 50:2541–544, 1943.

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

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, second edition, 2001.

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.

Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the ROC curve. In *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.

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.

D. Van Dantzig. On the consistency and power of Wilcoxon's two sample test. In *Koninklijke Nederlandse Akademie van Weterschappen, Series A*, volume 54, 1915.

Víctor H. de la Peña 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*, 2005. To appear.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

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.

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

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

Jiří Matoušek. *Lectures on Discrete Geometry*. Springer-Verlag, New York, 2002.

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

Saharon Rosset. Model selection via the AUC. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

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

Appendix A. Proof of Theorem 17

We shall 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 28 (Devroye, 1991; Devroye et al., 1996, Theorem 9.3) *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

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

Proof [of Theorem 17]
The proof is adapted from proofs of uniform convergence for the classification error rate given in (Anthony and Bartlett, 1999; Devroye et al., 1996). It consists of four steps.

Step 1. Symmetrization by a ghost sample.

For each $k \in \{1, \ldots, M\}$, define the random variable $\tilde{X}_k$ such that $X_k, \tilde{X}_k$ are independent and identically distributed. Let $\tilde{S}_X = (\tilde{X}_1, \ldots, \tilde{X}_M)$, and denote by $\tilde{S}$ the joint sequence $(\tilde{S}_X, \underline{y})$. Then for any $\epsilon > 0$ satisfying $mn\epsilon^2/(m+n) \geq 2$, we have

$$\mathbf{P}_{S_X|S_Y=\underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - A(f) \right| \geq \epsilon \right\} \quad \leq \quad 2\mathbf{P}_{S_X \tilde{S}_X|S_Y=\underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - \hat{A}(f; \tilde{S}) \right| \geq \frac{\epsilon}{2} \right\} .$$

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

$$\begin{aligned}
&\mathbf{P}_{S_X \tilde{S}_X|S_Y=\underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - \hat{A}(f; \tilde{S}) \right| \geq \frac{\epsilon}{2} \right\} \\
&\quad \geq \quad \mathbf{P}_{S_X \tilde{S}_X|S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S^*; S) - \hat{A}(f_S^*; \tilde{S}) \right| \geq \frac{\epsilon}{2} \right\} \\
&\quad \geq \quad \mathbf{P}_{S_X \tilde{S}_X|S_Y=\underline{y}} \left\{ \left\{ \left| \hat{A}(f_S^*; S) - A(f_S^*) \right| \geq \epsilon \right\} \cap \left\{ \left| \hat{A}(f_S^*; \tilde{S}) - A(f_S^*) \right| \leq \frac{\epsilon}{2} \right\} \right\} \\
&\quad = \quad \mathbf{E}_{S_X|S_Y=\underline{y}} \left\{ \mathbf{I}_{\{|\hat{A}(f_S^*; S) - A(f_S^*)| \geq \epsilon\}} \mathbf{P}_{\tilde{S}_X|S_X, S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S^*; \tilde{S}) - A(f_S^*) \right| \leq \frac{\epsilon}{2} \right\} \right\} . \quad (24)
\end{aligned}$$

The conditional probability inside can be bounded using Chebyshev's inequality (and Lemma 2):

$$\mathbf{P}_{\tilde{S}_X|S_X, S_Y=\underline{y}} \left\{ \left| \hat{A}(f_S^*; \tilde{S}) - A(f_S^*) \right| \leq \frac{\epsilon}{2} \right\} \quad \geq \quad 1 - \frac{\mathbf{Var}_{\tilde{S}_X|S_X, S_Y=\underline{y}} \left\{ \hat{A}(f_S^*; \tilde{S}) \right\}}{\epsilon^2/4} .$$

Now, by the same reasoning as in the proof of Theorem 5, a change in the value of a single random variable $\tilde{X}_k$ can cause a change of at most $1/m$ in $\hat{A}(f_S^*; \tilde{S})$ for $k : y_k = +1$, and a

change of at most $1/n$ for $k : y_k = -1$. Thus, by Theorem 28, we have

$$\mathbf{Var}_{\tilde{S}_X|S_X,S_Y=\underline{y}}\left\{\hat{A}(f_S^*;\tilde{S})\right\} \;\leq\; \frac{1}{4}\left(\sum_{\{i:y_i=+1\}}\left(\frac{1}{m}\right)^2+\sum_{\{j:y_j=-1\}}\left(\frac{1}{n}\right)^2\right) \;=\; \frac{m+n}{4mn}\,.$$

This gives

$$\mathbf{P}_{\tilde{S}_X|S_X,S_Y=\underline{y}}\left\{\left|\hat{A}(f_S^*;\tilde{S})-A(f_S^*)\right|\leq\frac{\epsilon}{2}\right\} \;\geq\; 1-\frac{m+n}{mn\epsilon^2} \;\geq\; \frac{1}{2}\,,$$

whenever $mn\epsilon^2/(m+n)\geq 2$. Thus, from Eq. (24) and the definition of f_S^*, we have

$$\begin{aligned}\mathbf{P}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S)-\hat{A}(f;\tilde{S})\right|\geq\frac{\epsilon}{2}\right\} &\geq \frac{1}{2}\mathbf{E}_{S_X|S_Y=\underline{y}}\left\{\mathbf{I}_{\{|\hat{A}(f_S^*;S)-A(f_S^*)|\geq\epsilon\}}\right\}\\ &= \frac{1}{2}\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\left|\hat{A}(f_S^*;S)-A(f_S^*)\right|\geq\epsilon\right\}\\ &\geq \frac{1}{2}\mathbf{P}_{S_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S)-A(f)\right|\geq\epsilon\right\}.\end{aligned}$$

Step 2. Permutations.

Let Γ_M be the set of all permutations of $\{X_1,\ldots,X_M,\tilde{X}_1,\ldots,\tilde{X}_M\}$ that swap X_k and $\tilde{X}_k$, for all k in some subset of $\{1,\ldots,M\}$. In other words, for all $\sigma\in\Gamma_M$ and $k\in\{1,\ldots,M\}$, either $\sigma(X_k)=X_k$, in which case $\sigma(\tilde{X}_k)=\tilde{X}_k$, or $\sigma(X_k)=\tilde{X}_k$, in which case $\sigma(\tilde{X}_k)=X_k$. Now, define

$$\begin{aligned}\beta_f(X_1,\ldots,X_M,\tilde{X}_1,\ldots,\tilde{X}_M) \;\equiv\; \frac{1}{mn}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\Bigg(&\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}}+\frac{1}{2}\mathbf{I}_{\{f(X_i)=f(X_j)\}}\right)\\ &-\left(\mathbf{I}_{\{f(\tilde{X}_i)>f(\tilde{X}_j)\}}+\frac{1}{2}\mathbf{I}_{\{f(\tilde{X}_i)=f(\tilde{X}_j)\}}\right)\Bigg).\end{aligned}$$

Then clearly, since $X_k,\tilde{X}_k$ are i.i.d. for each k, for any $\sigma\in\Gamma_M$ we have that the distribution of

$$\sup_{f\in\mathcal{F}}\left|\beta_f(X_1,\ldots,X_M,\tilde{X}_1,\ldots,\tilde{X}_M)\right|$$

is the same as the distribution of

$$\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(X_1),\ldots,\sigma(X_M),\sigma(\tilde{X}_1),\ldots,\sigma(\tilde{X}_M))\right|\,.$$

Therefore, using $\mathcal{U}(D)$ to denote the uniform distribution over a discrete set D, we have the following:

$$\mathbf{P}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\hat{A}(f;S)-\hat{A}(f;\tilde{S})\right|\geq\frac{\epsilon}{2}\right\}$$

$$
\begin{aligned}
&= \mathbf{P}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(X_1,\ldots,X_M,\tilde{X}_1,\ldots,\tilde{X}_M)\right| \geq \frac{\epsilon}{2}\right\} \\
&= \frac{1}{|\Gamma_M|}\sum_{\sigma\in\Gamma_M}\mathbf{P}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(X_1),\ldots,\sigma(X_M),\sigma(\tilde{X}_1),\ldots,\sigma(\tilde{X}_M))\right| \geq \frac{\epsilon}{2}\right\} \\
&= \frac{1}{|\Gamma_M|}\sum_{\sigma\in\Gamma_M}\mathbf{E}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\mathbf{I}_{\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(X_1),\ldots,\sigma(X_M),\sigma(\tilde{X}_1),\ldots,\sigma(\tilde{X}_M))\right|\geq\frac{\epsilon}{2}\right\}}\right\} \\
&= \mathbf{E}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\frac{1}{|\Gamma_M|}\sum_{\sigma\in\Gamma_M}\mathbf{I}_{\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(X_1),\ldots,\sigma(X_M),\sigma(\tilde{X}_1),\ldots,\sigma(\tilde{X}_M))\right|\geq\frac{\epsilon}{2}\right\}}\right\} \\
&= \mathbf{E}_{S_X\tilde{S}_X|S_Y=\underline{y}}\left\{\mathbf{P}_{\sigma\sim\mathcal{U}(\Gamma_M)}\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(X_1),\ldots,\sigma(X_M),\sigma(\tilde{X}_1),\ldots,\sigma(\tilde{X}_M))\right| \geq \frac{\epsilon}{2}\right\}\right\} \\
&\leq \max_{\underline{\mathbf{x}},\underline{\tilde{\mathbf{x}}}\in\mathcal{X}^M}\mathbf{P}_{\sigma\sim\mathcal{U}(\Gamma_M)}\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(\mathbf{x}_1),\ldots,\sigma(\mathbf{x}_M),\sigma(\tilde{\mathbf{x}}_1),\ldots,\sigma(\tilde{\mathbf{x}}_M))\right| \geq \frac{\epsilon}{2}\right\}.
\end{aligned}
$$

Step 3. Reduction to a finite class.

We wish to bound the quantity on the right hand side above. From the definition of bipartite rank matrices (Definition 14), it follows that for any $\underline{\mathbf{x}},\underline{\tilde{\mathbf{x}}}\in\mathcal{X}^M$, as f ranges over $\mathcal{F}$, the number of different random variables

$$\left|\beta_f(\sigma(\mathbf{x}_1),\ldots,\sigma(\mathbf{x}_M),\sigma(\tilde{\mathbf{x}}_1),\ldots,\sigma(\tilde{\mathbf{x}}_M))\right|$$

is at most the number of different bipartite rank matrices $\mathbf{B}_f(\underline{\mathbf{z}},\underline{\mathbf{z}}')$ that can be realized by functions in $\mathcal{F}$, where $\underline{\mathbf{z}}\in\mathcal{X}^{2m}$ contains $\mathbf{x}_i,\tilde{\mathbf{x}}_i$ for $i:y_i=+1$ and $\underline{\mathbf{z}}'\in\mathcal{X}'^{2n}$ contains $\mathbf{x}_j,\tilde{\mathbf{x}}_j$ for $j:y_j=-1$. This number, by definition, cannot exceed $r(\mathcal{F},2m,2n)$ (see the definition of bipartite rank-shatter coefficients, Definition 15). Therefore, the supremum in the above probability is a maximum of at most $r(\mathcal{F},2m,2n)$ random variables. Thus, by the union bound, we get for any $\underline{\mathbf{x}},\underline{\tilde{\mathbf{x}}}\in\mathcal{X}^M$,

$$
\begin{aligned}
&\mathbf{P}_{\sigma\sim\mathcal{U}(\Gamma_M)}\left\{\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(\mathbf{x}_1),\ldots,\sigma(\mathbf{x}_M),\sigma(\tilde{\mathbf{x}}_1),\ldots,\sigma(\tilde{\mathbf{x}}_M))\right| \geq \frac{\epsilon}{2}\right\} \\
&\quad\leq\ r(\mathcal{F},2m,2n)\cdot\sup_{f\in\mathcal{F}}\mathbf{P}_{\sigma\sim\mathcal{U}(\Gamma_M)}\left\{\left|\beta_f(\sigma(\mathbf{x}_1),\ldots,\sigma(\mathbf{x}_M),\sigma(\tilde{\mathbf{x}}_1),\ldots,\sigma(\tilde{\mathbf{x}}_M))\right| \geq \frac{\epsilon}{2}\right\}.
\end{aligned}
$$

Step 4. McDiarmid's inequality.

Notice that for any $\underline{\mathbf{x}},\underline{\tilde{\mathbf{x}}}\in\mathcal{X}^M$, we can write

$$
\begin{aligned}
&\mathbf{P}_{\sigma\sim\mathcal{U}(\Gamma_M)}\left\{\left|\beta_f(\sigma(\mathbf{x}_1),\ldots,\sigma(\mathbf{x}_M),\sigma(\tilde{\mathbf{x}}_1),\ldots,\sigma(\tilde{\mathbf{x}}_M))\right| \geq \frac{\epsilon}{2}\right\} \\
&\quad=\ \mathbf{P}_{\underline{W}\sim\mathcal{U}(\prod_{k=1}^M\{\mathbf{x}_k,\tilde{\mathbf{x}}_k\})}\left\{\left|\beta_f(W_1,\ldots,W_M,\tilde{W}_1,\ldots,\tilde{W}_M)\right| \geq \frac{\epsilon}{2}\right\},
\end{aligned}
$$

where $\underline{W}=(W_1,\ldots,W_M)$ and $\tilde{W}_k=\begin{cases}\tilde{\mathbf{x}}_k, & \text{if } W_k=\mathbf{x}_k\\ \mathbf{x}_k, & \text{if } W_k=\tilde{\mathbf{x}}_k\end{cases}$.

Now, for $f : \mathcal{X} \to \mathbb{R}$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, let

$$\alpha(f; \mathbf{x}, \mathbf{x}') \equiv \mathbf{I}_{\{f(\mathbf{x}) > f(\mathbf{x}')\}} + \frac{1}{2}\mathbf{I}_{\{f(\mathbf{x}) = f(\mathbf{x}')\}} .$$

Then for any $f \in \mathcal{F}$,

$$\begin{aligned}
&\mathbf{E}_{\underline{W} \sim \mathcal{U}(\prod_{k=1}^{M} \{\mathbf{x}_k, \tilde{\mathbf{x}}_k\})} \left\{ \beta_f(W_1, \ldots, W_M, \tilde{W}_1, \ldots, \tilde{W}_M) \right\} \\
&= \frac{1}{mn} \sum_{\{i: y_i = +1\}} \sum_{\{j: y_j = -1\}} \mathbf{E}_{W_i \sim \mathcal{U}(\{\mathbf{x}_i, \tilde{\mathbf{x}}_i\}), W_j \sim \mathcal{U}(\{\mathbf{x}_j, \tilde{\mathbf{x}}_j\})} \left\{ \alpha(f; W_i, W_j) - \alpha(f; \tilde{W}_i, \tilde{W}_j) \right\} \\
&= \frac{1}{mn} \sum_{\{i: y_i = +1\}} \sum_{\{j: y_j = -1\}} \frac{1}{4} \Big[\Big(\alpha(f; \mathbf{x}_i, \mathbf{x}_j) - \alpha(f; \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \Big) + \Big(\alpha(f; \tilde{\mathbf{x}}_i, \mathbf{x}_j) - \alpha(f; \mathbf{x}_i, \tilde{\mathbf{x}}_j) \Big) + \\
&\qquad \Big(\alpha(f; \mathbf{x}_i, \tilde{\mathbf{x}}_j) - \alpha(f; \tilde{\mathbf{x}}_i, \mathbf{x}_j) \Big) + \Big(\alpha(f; \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) - \alpha(f; \mathbf{x}_i, \mathbf{x}_j) \Big) \Big] \\
&= 0 .
\end{aligned}$$

Also, it can be verified that for any $f \in \mathcal{F}$, a change in the value of a single random variable W_k can bring a change of at most $2/m$ in the value of

$$\beta_f(W_1, \ldots, W_M, \tilde{W}_1, \ldots, \tilde{W}_M)$$

for $k : y_k = +1$, and a change of at most $2/n$ for $k : y_k = -1$. Therefore, by McDiarmid's inequality (Theorem 3), it follows that for any $f \in \mathcal{F}$,

$$\begin{aligned}
&\mathbf{P}_{\underline{W} \sim \mathcal{U}(\prod_{k=1}^{M} \{\mathbf{x}_k, \tilde{\mathbf{x}}_k\})} \left\{ \left| \beta_f(W_1, \ldots, W_M, \tilde{W}_1, \ldots, \tilde{W}_M) \right| \geq \frac{\epsilon}{2} \right\} \\
&\leq 2e^{-2\epsilon^2 / 4(m(\frac{2}{m})^2 + n(\frac{2}{n})^2)} \\
&= 2e^{-mn\epsilon^2/8(m+n)} .
\end{aligned}$$

Putting everything together, we get that

$$\mathbf{P}_{S_X | S_Y = \underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - A(f) \right| \geq \epsilon \right\} \leq 4 \cdot r(\mathcal{F}, 2m, 2n) \cdot e^{-mn\epsilon^2/8(m+n)} ,$$

for $mn\epsilon^2/(m+n) \geq 2$. In the other case, *i.e.*, for $mn\epsilon^2/(m+n) < 2$, the bound is greater than one and therefore holds trivially. ■

Appendix B. Proof of Theorem 27

We shall need to extend the notion of error rate to $\mathcal{Y}^*$-valued functions (recall that $\mathcal{Y}^* = \{-1, 0, +1\}$). Given a function $h : \mathcal{X} \to \mathcal{Y}^*$ and a data sequence $T = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$, let the empirical error rate of h with respect to T be denoted by $\hat{L}^*(h; T)$ and defined as

$$\hat{L}^*(h;T) \;=\; \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{I}_{\{h(\mathbf{x}_i)\neq 0\}}\mathbf{I}_{\{h(\mathbf{x}_i)\neq y_i\}} + \frac{1}{2}\mathbf{I}_{\{h(\mathbf{x}_i)=0\}}\right\}. \tag{25}$$

Similarly, for an underlying distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, let the expected error rate of h be denoted by $L^*(h)$ and defined as

$$L^*(h) \;=\; \mathbf{E}_{XY\sim\mathcal{D}}\left\{\mathbf{I}_{\{h(X)\neq 0\}}\mathbf{I}_{\{h(X)\neq Y\}} + \frac{1}{2}\mathbf{I}_{\{h(X)=0\}}\right\}. \tag{26}$$

Then, following the proof of a similar result given in (Vapnik, 1982) for binary-valued functions, it can be shown that if $\mathcal{H}$ is a class of $\mathcal{Y}^*$-valued functions on $\mathcal{X}$ and $M \in \mathbb{N}$, then for any $\epsilon > 0$,

$$\mathbf{P}_{S\sim\mathcal{D}^M}\left\{\sup_{h\in\mathcal{H}}\left|\hat{L}^*(h;S) - L^*(h)\right| \geq \epsilon\right\} \;\leq\; 6s(\mathcal{H}, 2M)e^{-M\epsilon^2/4}. \tag{27}$$

Proof [of Theorem 27]
To keep notation concise, for $f : \mathcal{X} \to \mathbb{R}$ and $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, let

$$\eta(f;\mathbf{x},\mathbf{x}') \;\equiv\; \mathbf{I}_{\{f(\mathbf{x})<f(\mathbf{x}')\}} + \frac{1}{2}\mathbf{I}_{\{f(\mathbf{x})=f(\mathbf{x}')\}},$$

and for $h : \mathcal{X} \to \mathcal{Y}^*$, $\mathbf{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, let

$$\nu(h;\mathbf{x},y) \;\equiv\; \mathbf{I}_{\{h(\mathbf{x})\neq 0\}}\mathbf{I}_{\{h(\mathbf{x})\neq y\}} + \frac{1}{2}\mathbf{I}_{\{h(\mathbf{x})=0\}}.$$

Now, given $S_Y = \underline{y}$, we have for all $f \in \mathcal{F}$

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

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

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

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

where $S^{(m)}_{+1}$ and $S^{(n)}_{-1}$ denote the subsequences of S containing the m positive and n negative examples, respectively. Now, from the confidence interval interpretation of the result given in Eq. (27), we have

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

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

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