A Uniform Convergence Bound for the Area Under the ROC Curve

Shivani Agarwal, Sariel Har-Peled and Dan Roth
Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Avenue
Urbana, IL 61801, USA
{sagarwal,sariel,danr}@cs.uiuc.edu

Abstract

The area under the ROC curve (AUC) has been advocated as an evaluation criterion for the bipartite ranking problem. We study uniform convergence properties of the AUC; in particular, we derive a distribution-free uniform convergence bound for the AUC which serves 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. Our 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. [9] for a quantity closely related to the AUC shows that the bound provided by our result can be considerably tighter.

1 INTRODUCTION

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

The problem of ranking has been studied from a learning perspective under a variety of settings [4, 10, 6, 9]. 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, training examples consist of documents labeled as relevant (positive) or irrelevant (negative). This form of ranking problem corresponds to the 'bipartite feedback' case of [9]; we therefore 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 $\mathcal{X}$ and a set of two class labels $\mathcal{Y} = \{-1, +1\}$. 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.

The *area under the ROC curve* (AUC) has recently gained attention as an evaluation criterion for the bipartite ranking problem [5]. Given a ranking function $f : \mathcal{X} \to \mathbb{R}$ and a data sequence $T = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ containing m positive and n negative examples, the AUC of f with respect to T, denoted $\hat{A}(f; T)$, can be expressed as the following Wilcoxon-Mann-Whitney statistic [5]:

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

where $\mathbf{I}_{\{\cdot\}}$ denotes the indicator variable whose value is one if its argument is true and zero otherwise. The AUC of f with respect to T is thus simply the fraction of positive-negative pairs in T that are ranked correctly by f, assuming that ties are broken uniformly at random.[1]

The AUC is an empirical quantity that evaluates a ranking function with respect to a particular data sequence. What

[1] In [5], a slightly simpler form of the Wilcoxon-Mann-Whitney statistic is used, which does not account for ties.

does the empirical AUC tell us about the expected performance of a ranking function on future examples? This is the question we consider. 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 question is addressed in [1]; we address the second question 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 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. Properties of the bipartite rank-shatter coefficients are discussed in Section 4. Section 5 compares our result with a recent uniform convergence result derived by Freund et al. [9] for a quantity closely related to the AUC. We conclude with some open questions in Section 6.

2 EXPECTED RANKING ACCURACY

We begin by introducing some 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 $\mathcal{D}_{y_1} \times \ldots \times \mathcal{D}_{y_N}$.[2] 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.

[2]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 $\mathcal{D}_{+1}^m \times \mathcal{D}_{-1}^n$.

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:*

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

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 following simple lemma shows that the empirical AUC of a ranking function f is an unbiased estimator of the expected ranking accuracy of f:

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

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

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. (1)) and linearity of expectation, we have

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

□

3 UNIFORM CONVERGENCE BOUND

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 $\epsilon > 0$,

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

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.

Definition 2 (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*

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

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

Definition 3 (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:*

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

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

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

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 (proof omitted due to lack of space):

Theorem 1. *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*

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 1.*

We discuss further properties of the bipartite rank-shatter coefficients in Section 4; we first present below our uniform convergence result in terms of these coefficients. The following can be viewed as the main result of this paper. We note that our results are all distribution-free, in the sense that they hold for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.

Table 1: Sub-matrices that cannot appear in a bipartite rank matrix.

$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	$\begin{bmatrix}½ & 0\\ 0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & ½\\ 0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\ 0 & ½\end{bmatrix}$	$\begin{bmatrix}1 & 0\\ ½ & 1\end{bmatrix}$	$\begin{bmatrix}½ & ½\\ 0 & 1\end{bmatrix}$	$\begin{bmatrix}½ & 0\\ 0 & ½\end{bmatrix}$	$\begin{bmatrix}½ & 0\\ ½ & 1\end{bmatrix}$	$\begin{bmatrix}1 & ½\\ 0 & ½\end{bmatrix}$	$\begin{bmatrix}1 & ½\\ ½ & 1\end{bmatrix}$
$\begin{bmatrix}1 & 0\\ ½ & ½\end{bmatrix}$	$\begin{bmatrix}1 & ½\\ ½ & ½\end{bmatrix}$	$\begin{bmatrix}½ & 0\\ ½ & ½\end{bmatrix}$	$\begin{bmatrix}½ & ½\\ ½ & 1\end{bmatrix}$	$\begin{bmatrix}½ & ½\\ 0 & ½\end{bmatrix}$	$\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$	$\begin{bmatrix}½ & 1\\ 1 & 0\end{bmatrix}$	$\begin{bmatrix}0 & ½\\ 1 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 1\\ 1 & ½\end{bmatrix}$	$\begin{bmatrix}0 & 1\\ ½ & 0\end{bmatrix}$
$\begin{bmatrix}½ & ½\\ 1 & 0\end{bmatrix}$	$\begin{bmatrix}½ & 1\\ 1 & ½\end{bmatrix}$	$\begin{bmatrix}½ & 1\\ ½ & 0\end{bmatrix}$	$\begin{bmatrix}0 & ½\\ 1 & ½\end{bmatrix}$	$\begin{bmatrix}0 & ½\\ ½ & 0\end{bmatrix}$	$\begin{bmatrix}0 & 1\\ ½ & ½\end{bmatrix}$	$\begin{bmatrix}0 & ½\\ ½ & ½\end{bmatrix}$	$\begin{bmatrix}½ & 1\\ ½ & ½\end{bmatrix}$	$\begin{bmatrix}½ & ½\\ ½ & 0\end{bmatrix}$	$\begin{bmatrix}½ & ½\\ 1 & ½\end{bmatrix}$

Theorem 2. *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 $\epsilon > 0$,*

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

The proof is adapted from uniform convergence proofs for the classification error rate (see, for example, [2, 8]). 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 [7] is required to bound the variance of the AUC that appears after an application of Chebyshev's inequality, and McDiarmid's inequality [12] is required in the final step of the proof where Hoeffding's inequality sufficed in the case of classification. Details are given in Appendix A.

We note that the result of Theorem 2 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 $\underline{y}$.[3] From Theorem 2, we can derive a confidence interval interpretation of the bound as follows:

Corollary 1. *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 $0 < \delta \leq 1$,*

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

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

As in the case of the large deviation bound of [1], the confidence interval above can be generalized to remove the conditioning on the label vector completely (we note that Theorem 2 cannot be generalized in this manner):

Theorem 3. *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$, and let $M \in \mathbb{N}$. Then for any $0 < \delta \leq 1$,*

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

where $\rho(S_Y)$ denotes the proportion of positive labels in S_Y.

[3] Our thanks to an anonymous reviewer for pointing this out.

4 PROPERTIES OF BIPARTITE RANK-SHATTER COEFFICIENTS

As discussed above, 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}, but 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 1. The conclusion then follows from Theorem 1 (Part 2).) For the bound of Theorem 2 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 shatter coefficients studied in 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 4 (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:*

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

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 5. Note that we take

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

Definition 5 (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 $\mathcal{F}$:*

$$1.\quad \bar{\mathcal{F}} \;=\; \Big\{\bar{f}:\mathcal{X}\to\mathcal{Y}^* \mid \bar{f}(\mathbf{x}) = \text{sign}(f(\mathbf{x})) \text{ for some } f\in\mathcal{F}\Big\} \tag{2}$$

$$2.\quad \tilde{\mathcal{F}} \;=\; \Big\{\tilde{f}:\mathcal{X}\times\mathcal{X}\to\mathcal{Y}^* \mid \tilde{f}(\mathbf{x},\mathbf{x}') = \text{sign}(f(\mathbf{x})-f(\mathbf{x}')) \text{ for some } f\in\mathcal{F}\Big\} \tag{3}$$

$$3.\quad \check{\mathcal{F}} \;=\; \Big\{\check{f}_{\mathbf{z}}:\mathcal{X}\to\mathcal{Y}^* \mid \check{f}_{\mathbf{z}}(\mathbf{x}) = \text{sign}(f(\mathbf{x})-f(\mathbf{z})) \text{ for some } f\in\mathcal{F}, \mathbf{z}\in\mathcal{X}\Big\} \tag{4}$$

Theorem 4. *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. (3). 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]

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

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

Lemma 2. *For $d\in\mathbb{N}$, let $\mathcal{F}_{\text{lin}(d)}$ denote the class of linear ranking functions on $\mathbb{R}^d$:*

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

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

$$s(\tilde{\mathcal{F}}_{\text{lin}(d)},N) \;\leq\; (2eN/d)^d.$$

Proof. We have,

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

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 $\tilde{f}\in\tilde{\mathcal{F}}$ is equal to the total number of faces of this arrangement [11], which is at most [3]

$$\sum_{k=0}^{d}\sum_{i=d-k}^{d}\binom{i}{d-k}\binom{N}{i} \;=\; \sum_{i=0}^{d}2^i\binom{N}{i} \;\leq\; (2eN/d)^d\,.$$

Since the N points were arbitrary, the result follows. $\square$

Theorem 5. *For $d\in\mathbb{N}$, let $\mathcal{F}_{\text{lin}(d)}$ denote the class of linear ranking functions on $\mathbb{R}^d$ (defined in Lemma 2 above). Then for all $m,n\in\mathbb{N}$,*

$$r(\mathcal{F}_{\text{lin}(d)},m,n) \;\leq\; (2emn/d)^d.$$

Proof. This follows immediately from Theorem 4 and Lemma 2. $\square$

[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.

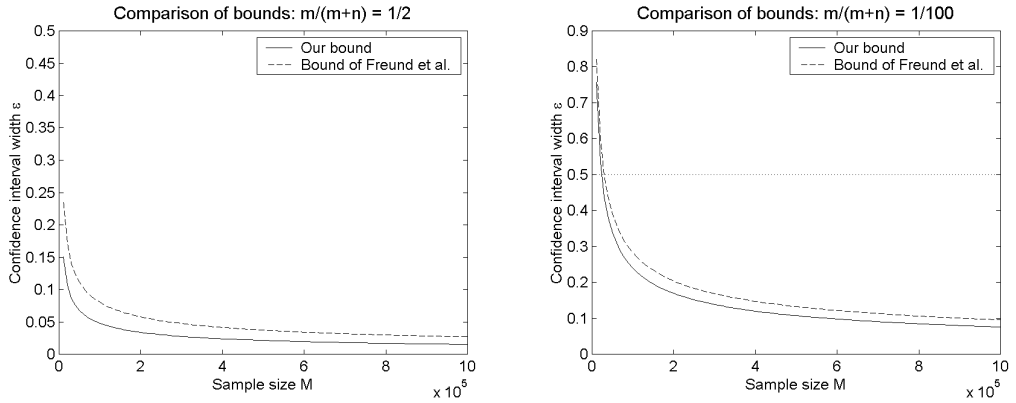

Figure 1: A comparison of our uniform convergence bound with that of [9] for the class of linear ranking functions on $\mathbb{R}$. The plots are for $\delta = 0.01$ and show how the confidence interval width ϵ given by the two bounds varies with the sample size M, for various values of $m/(m+n)$. In all cases where the bounds are meaningful ($\epsilon < 0.5$), our bound is tighter.

5 COMPARISON WITH BOUND OF FREUND ET AL.

Freund et al. [9] 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 in [5], the bipartite ranking loss is equal to one minus the AUC; the uniform convergence bound of [9] therefore implies a uniform convergence bound for the AUC.[5] Although the result in [9] 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 [9], the result is given in the form of a confidence interval.[6]

Theorem 6 (Generalization of [9], 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 $0 < \delta \leq 1$,*

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

where $\check{\mathcal{F}}$ is the class of $\mathcal{Y}^*$-valued functions on $\mathcal{X}$ defined by Eq. (4).

[5] As in the AUC definition of [5], the ranking loss defined in [9] does not account for ties; this is easily remedied.

[6] The result in [9] 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.

The proof follows that of [9] and is omitted. We now compare the uniform convergence bound derived in Section 3 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 $\mathcal{F}_{\text{lin}(1)}$ of linear ranking functions on $\mathbb{R}$, given by

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

Although $\mathcal{F}_{\text{lin}(1)}$ is an infinite function class, it is easy to verify that $r(\mathcal{F}_{\text{lin}(1)}, m, n) = 3$ 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 $\mathcal{F}_{\text{lin}(1)}$: one by setting $w > 0$, another by setting $w < 0$, and the third by setting $w = 0$.) On the other hand, $s(\check{\mathcal{F}}_{\text{lin}(1)}, N) = 4N + 1$ for all $N \geq 2$, since $\check{\mathcal{F}}_{\text{lin}(1)} = \bar{\mathcal{F}}_{\text{lin}(1)}$ (see Eq. (2)) 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 $\bar{\mathcal{F}}_{\text{lin}(1)}$ is $4N + 1$. We thus get from our result (Corollary 1) that

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

and from the result of Freund et al. (Theorem 6) that

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

The above bounds are plotted in Figure 1 for $\delta = 0.01$ and various values of $m/(m+n)$. As can be seen, the bound provided by our result is considerably tighter.

6 CONCLUSION & OPEN QUESTIONS

We have derived a distribution-free uniform convergence bound for the area under the ROC curve (AUC), a quantity used as an evaluation criterion for the bipartite ranking problem. Our bound is expressed in terms of a new set of combinatorial parameters that we have termed the bipartite rank-shatter coefficients. These coefficients define a new measure of complexity for 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 bound of Freund et al. [9], 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 4 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. (3)); this allows us to establish a polynomial upper bound on the bipartite rank-shatter coefficients for linear ranking functions on $\mathbb{R}^d$ and other algebraically well-behaved function classes. However, this upper bound is inherently loose (see proof of Theorem 4). Is it possible to find tighter upper bounds on $r(\mathcal{F}, m, n)$ than that given by Theorem 4?

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 generalization bounds for the AUC can be derived using different proof techniques.

Acknowledgements

We would like to thank Thore Graepel and Ralf Herbrich for discussions related to this work and for pointing out to us the graph-based interpretation of bipartite rank matrices used in Theorem 1. We are also very grateful to an anonymous reviewer, and to Thore Graepel and Ralf Herbrich again, for helping us identify an important mistake in an earlier version of our 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

[1] 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, 2005. To appear.

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

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

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

[5] Corinna Cortes and Mehryar Mohri. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems 16*. MIT Press, 2004.

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

[7] 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.

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

[9] 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.

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

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

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

A Proof of Theorem 2

Our proof makes use of the following two results [12, 7] that bound the probability of a large deviation and the variance, respectively, of any function of a sample for which a single change in the sample has limited effect:

Theorem 7 (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*

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

Then for any $\epsilon > 0$,

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

Theorem 8 (Devroye, 1991; Devroye et al., 1996, Theorem 9.3). *Under the conditions of Theorem 7,*

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

The following lemma establishes that a change in a single instance in a data sequence has a limited effect on the AUC of a ranking function with respect to the data sequence:

Lemma 3. *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. Let m be the number of positive labels in $\underline{y}$ and n the number of negative labels in $\underline{y}$. Let $\phi : \mathcal{X}^N \to \mathbb{R}$ be defined 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 all $\mathbf{x}_i, \mathbf{x}'_k \in \mathcal{X}$,

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

where $c_k = 1/m$ if $y_k = +1$ and $c_k = 1/n$ if $y_k = -1$.

Proof. For each k such that $y_k = +1$, we have

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

The case $y_k = -1$ can be proved similarly. □

We are now ready to give the main proof:

Proof (of Theorem 2). The proof is adapted from proofs of uniform convergence for the classification error rate given in [2, 8]. It consists of four steps.

Step 1. First 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

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

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

$$\begin{aligned}
&\mathbf{P}_{S_X \tilde{S}_X | S_Y = \underline{y}} \left\{ \sup_{f \in \mathcal{F}} \left| \hat{A}(f; S) - \hat{A}(f; \tilde{S}) \right| \geq \frac{\epsilon}{2} \right\} \\
&\quad\geq\; \mathbf{P}_{S_X \tilde{S}_X | S_Y = \underline{y}} \left\{ \left| \hat{A}(f^*_S; S) - \hat{A}(f^*_S; \tilde{S}) \right| \geq \frac{\epsilon}{2} \right\} \\
&\quad\geq\; \mathbf{P}_{S_X \tilde{S}_X | S_Y = \underline{y}} \left\{ \left\{ \left| \hat{A}(f^*_S; S) - A(f^*_S) \right| \geq \epsilon \right\} \cap \left\{ \left| \hat{A}(f^*_S; \tilde{S}) - A(f^*_S) \right| \leq \frac{\epsilon}{2} \right\} \right\} \\
&\quad=\; \mathbf{E}_{S_X | S_Y = \underline{y}} \left\{ \mathbf{I}_{\{|\hat{A}(f^*_S; S) - A(f^*_S)| \geq \epsilon\}} \times \mathbf{P}_{\tilde{S}_X | S_X, S_Y = \underline{y}} \left\{ \left| \hat{A}(f^*_S; \tilde{S}) - A(f^*_S) \right| \leq \frac{\epsilon}{2} \right\} \right\} \,. \qquad (5)
\end{aligned}$$

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

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

Now, by Lemma 3 and Theorem 8, we have

$$\begin{aligned}
&\mathbf{Var}_{\tilde{S}_X | S_X, S_Y = \underline{y}} \left\{ \hat{A}(f^*_S; \tilde{S}) \right\} \\
&\qquad\leq\; \frac{1}{4} \left(m \Big(\frac{1}{m}\Big)^2 + n \Big(\frac{1}{n}\Big)^2 \right) \;=\; \frac{m+n}{4mn} \,.
\end{aligned}$$

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. (5) and the definition of f^*_S, we have

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

Step 2. Second symmetrization by 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 $\sigma(\tilde{X}_k) = X_k$. Denote $\sigma(S_X) = (\sigma(X_1), \ldots, \sigma(X_M))$, and $\sigma(\tilde{S}_X) = (\sigma(\tilde{X}_1), \ldots, \sigma(\tilde{X}_M))$. Now, define

$$\beta_f(S_X,\tilde{S}_X) \equiv \frac{1}{mn}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\left(\left(\mathbf{I}_{\{f(X_i)>f(X_j)\}}+\frac{1}{2}\mathbf{I}_{\{f(X_i)=f(X_j)\}}\right)-\left(\mathbf{I}_{\{f(\tilde{X}_i)>f(\tilde{X}_j)\}}+\frac{1}{2}\mathbf{I}_{\{f(\tilde{X}_i)=f(\tilde{X}_j)\}}\right)\right).$$

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

$$\sup_{f\in\mathcal{F}}\left|\beta_f(S_X,\tilde{S}_X)\right|$$

is the same as the distribution of

$$\sup_{f\in\mathcal{F}}\left|\beta_f(\sigma(S_X),\sigma(\tilde{S}_X))\right|.$$

Therefore, using $\mathcal{U}(D)$ to denote the uniform distribution over a discrete set D, we have the following (note that except where specified otherwise, all probabilities and expectations below are with respect to the distribution $\mathcal{D}_{S_X\tilde{S}_X|S_Y=\underline{y}}$):

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

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 2), 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

$$\left|\beta_f(\sigma(\underline{\mathbf{x}}),\sigma(\underline{\tilde{\mathbf{x}}}))\right|$$

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 3). 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 $\underline{\mathbf{x}},\underline{\tilde{\mathbf{x}}}\in\mathcal{X}^M$,

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

Step 4. McDiarmid's inequality.

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

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

where $\underline{W}=(W_1,\ldots,W_M)$, $\underline{\tilde{W}}=(\tilde{W}_1,\ldots,\tilde{W}_M)$ 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 any $f\in\mathcal{F}$,

$$\mathbf{E}_{\underline{W}\sim\mathcal{U}(\prod_{k=1}^M\{\mathbf{x}_k,\tilde{\mathbf{x}}_k\})}\left\{\beta_f(\underline{W},\underline{\tilde{W}})\right\}\ =\ 0,$$

since for all $i:y_i=+1$ and $j:y_j=-1$,

$$\begin{aligned}
&\mathbf{E}_{W_i\sim\mathcal{U}(\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}),W_j\sim\mathcal{U}(\{\mathbf{x}_j,\tilde{\mathbf{x}}_j\})}\left\{\mathbf{I}_{\{f(W_i)>f(W_j)\}}-\mathbf{I}_{\{f(\tilde{W}_i)>f(\tilde{W}_j)\}}\right\}\\
&\quad=\frac{1}{4}\Big(\left(\mathbf{I}_{\{f(\mathbf{x}_i)>f(\mathbf{x}_j)\}}-\mathbf{I}_{\{f(\tilde{\mathbf{x}}_i)>f(\tilde{\mathbf{x}}_j)\}}\right)+\\
&\qquad\left(\mathbf{I}_{\{f(\tilde{\mathbf{x}}_i)>f(\mathbf{x}_j)\}}-\mathbf{I}_{\{f(\mathbf{x}_i)>f(\tilde{\mathbf{x}}_j)\}}\right)+\\
&\qquad\left(\mathbf{I}_{\{f(\mathbf{x}_i)>f(\tilde{\mathbf{x}}_j)\}}-\mathbf{I}_{\{f(\tilde{\mathbf{x}}_i)>f(\mathbf{x}_j)\}}\right)+\\
&\qquad\left(\mathbf{I}_{\{f(\tilde{\mathbf{x}}_i)>f(\tilde{\mathbf{x}}_j)\}}-\mathbf{I}_{\{f(\mathbf{x}_i)>f(\mathbf{x}_j)\}}\right)\Big)\\
&\quad=0,
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\mathbf{E}_{W_i\sim\mathcal{U}(\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}),W_j\sim\mathcal{U}(\{\mathbf{x}_j,\tilde{\mathbf{x}}_j\})}\left\{\mathbf{I}_{\{f(W_i)=f(W_j)\}}-\mathbf{I}_{\{f(\tilde{W}_i)=f(\tilde{W}_j)\}}\right\}\\
&\quad=0.
\end{aligned}$$

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 about a change of at most $2/m$ in the value of

$$\beta_f(\underline{W},\underline{\tilde{W}})$$

for $k:y_k=+1$, and a change of at most $2/n$ for $k:y_k=-1$. Therefore, by McDiarmid's inequality (Theorem 7), it follows that for any $f\in\mathcal{F}$,

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

Putting everything together, we get that

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

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. □