Journal of Machine Learning Research 10 (2009) 2233-2271 Submitted 4/08; Revised 10/08; Published 10/09

The P-Norm Push: A Simple Convex Ranking Algorithm that Concentrates at the Top of the List

Cynthia Rudin[*] RUDIN@MIT.EDU
MIT Sloan School of Management
Cambridge, MA 02142

Editor: Michael Collins

Abstract

We are interested in supervised ranking algorithms that perform especially well near the top of the ranked list, and are only required to perform sufficiently well on the rest of the list. In this work, we provide a general form of convex objective that gives high-scoring examples more importance. This "push" near the top of the list can be chosen arbitrarily large or small, based on the preference of the user. We choose ℓ_p-norms to provide a specific type of push; if the user sets p larger, the objective concentrates harder on the top of the list. We derive a generalization bound based on the p-norm objective, working around the natural asymmetry of the problem. We then derive a boosting-style algorithm for the problem of ranking with a push at the top. The usefulness of the algorithm is illustrated through experiments on repository data. We prove that the minimizer of the algorithm's objective is unique in a specific sense. Furthermore, we illustrate how our objective is related to quality measurements for information retrieval.

Keywords: ranking, RankBoost, generalization bounds, ROC, information retrieval

1. Introduction

The problem of supervised ranking is useful in many application domains, for instance, maintenance operations to be performed in a specific order, natural language processing, information retrieval, and drug discovery. Many of these domains require the construction of a ranked list, yet often, only the top portion of the list is used in practice. For instance, in the setting of supervised movie ranking, the learning algorithm provides the user (an avid movie-goer) with a ranked list of movies based on preference data. We expect the user to examine the top portion of the list as a recommendation. It is possible that she never looks at the rest of the list, or examines it only briefly. Thus, we wish to make sure that the top portion of the list is correctly constructed. This is the problem on which we concentrate.

We present a fairly general and flexible technique for solving these types of problems. Specifically, we derive a convex objective function that places more emphasis at the top of the list. The algorithm we develop using this technique ("The P-Norm Push") is based on minimization of a specific version of this objective. The user chooses a parameter "p" in the objective, corresponding to the p of an ℓ_p norm. By varying p, one changes the degree of concentration ("push") at the top of the list. One can concentrate at the very top of the list (a big push, large p), or one can have a moderate emphasis at the top (a little push, low p), or somewhere in between. The case with no

∗. Also at Center for Computational Learning Systems, Columbia University, 475 Riverside Drive MC 7717, New York, NY 10115.

emphasis at the top (no push, $p = 1$) corresponds to a standard objective for supervised bipartite ranking, namely the exponentiated pairwise misranking error.

The P-Norm Push is motivated in the setting of supervised bipartite ranking. In the supervised bipartite ranking problem, each training instance has a label of +1 or -1; each movie is either a good movie or a bad movie. In this case, we want to push the bad movies away from the top of the list where the good movies are desired. The quality of a ranking can be determined by examining the Receiver Operator Characteristic (ROC) curve. The AUC (Area Under the ROC Curve) is precisely a constant times one minus the total standard pairwise misranking error. The accuracy measure for our problem is different; we care mostly about the leftmost portion of the ROC curve, corresponding to the top of the ranked list. We wish to make the leftmost portion of the curve higher. Thus, we choose to make a tradeoff: in order make the leftmost portion of the curve higher, we sacrifice on the total area underneath the curve. The parameter p in the P-Norm Push allows the user to directly control this tradeoff.

This problem is highly asymmetric with respect to the positive and negative classes, and is not represented by a sum of independent random variables. It is interesting to consider generalization bounds for such a problem; it is not clear how to use standard techniques that require natural symmetry with respect to the positive and negative examples, for instance, many VC bounds rely on this kind of symmetry. In this work, we present a generalization bound that uses covering numbers as a measure of complexity. This bound is designed specifically to handle these asymmetric conditions. The bound underscores an important property of algorithms that concentrate on a small portion of the domain, such as algorithms that concentrate on the top of a ranked list: these algorithms require more examples for generalization.

Recently, there has been a large amount of interest in the supervised ranking problem, and especially in the bipartite problem. Freund et al. (2003) have developed the RankBoost algorithm for the general setting. We inherit the setup of RankBoost, and our algorithm will also be a boosting-style algorithm. Oddly, Freund and Schapire's classification algorithm AdaBoost (Freund and Schapire, 1997) performs just as well for bipartite ranking as RankBoost; both algorithms achieve equally good values of the AUC (Rudin and Schapire, 2009). This is in contrast with support vector machine classifiers (Cortes and Vapnik, 1995), which do not tend to perform well for the bipartite ranking problem (Rakotomamonjy, 2004; Brefeld and Scheffer, 2005). Mozer et al. (2002) aim to manipulate specific points of the ROC curve in order to study "churn" in the telecommunications industry. Perhaps the closest algorithm to ours is the one proposed by Dekel et al. (2004), who have used a similar form of objective with different specifics to achieve a different goal, namely to rank labels in a multilabel setting. Other related works on label ranking include those of Crammer and Singer (2001) and Shalev-Shwartz and Singer (2006). The work of Yan et al. (2003) contains a brief mention of a method to optimize the lower left corner of the ROC curve, though their multi-layer perception approach is highly non-convex. There is a lot of recent work on generalization bounds (and large deviation bounds) for supervised ranking, namely, the bounds of Freund et al. (2003), Clemençon et al. (2008), Agarwal et al. (2005), Usunier et al. (2005), Hill et al. (2002), Rudin et al. (2005) and Rudin and Schapire (2009), though we were only able to adapt techniques from the latter two bounds to our particular setting, since the covering number approach can handle the natural asymmetry of our problem. There is also a body of work on ROC curves in general, for example, the estimation of confidence bands for ROC curves (Macskassy et al., 2005), and more recent works by Clemençon and Vayatis addressing statistical aspects of ranking problems (e.g., Clemençon and Vayatis, 2007, 2008).

There is a large body of literature on information retrieval (IR) that considers other quality measurements for a ranked list, including "discounted cumulative gain," "average precision" and "winner take all." In essence, the P-Norm Push algorithm can be considered as a way to interpolate between AUC maximization (no push, $p = 1$) and a quantity similar to "winner take all" (largest possible push, $p = \infty$). A simple variation of the P-Norm Push derivation can be used to derive convex objectives that are somewhat similar to the "discounted cumulative gain" as we illustrate in Section 7. Our approach yields simple smooth convex objectives that can be minimized using simple coordinate techniques. In that sense, our work complements those of Tsochantaridis et al. (2005) and Le and Smola (2007) who also minimize a convex upper bound of IR ranking measurements, but with a structured learning approach that requires optimization with exponentially many constraints; those works have suggested useful ways to combat this problem. Additionally, there are recent works (Cossock and Zhang, 2006; Zheng et al., 2007) that suggest regression approaches to optimize ranking criteria for information retrieval.

Here is the outline of the work: in Section 2, we present a general form of objective function, allowing us to incorporate a push near the top of the ranked list. In order to construct a specific case of this objective, one chooses both a loss function ℓ and a convex "price" function g. We will choose g to be a power law, $g(r) = r^p$, so that a higher power p corresponds to a larger push near the top. In Section 3 we give some examples to illustrate how the objective works. In Section 4, we provide a generalization bound for our objective with ℓ as the 0-1 loss, based on L_∞ covering numbers. The generalization bound has been improved from the conference version of this work (Rudin, 2006). In Section 5 we derive the "P-Norm Push" coordinate descent algorithm based on the objective with ℓ chosen as the exponential loss used for AdaBoost and RankBoost. Section 6 discusses uniqueness of the minimizer of the P-Norm Push algorithm's objective. We prove that the minimizer is unique in a specific sense. This result is based on conjugate duality and the theory of Bregman distances (Della Pietra et al., 2002), and is analogous to the result of Collins et al. (2002) for AdaBoost. The "primal" problem for AdaBoost can be written as relative entropy minimization. For the objective of the P-Norm Push algorithm, the problem is more difficult and the primal is not a common function. Section 7 illustrates the similarity between quality measurements used for information retrieval and our objective, and gives other variations of the objective. In Section 8, we demonstrate the P-Norm Push on repository data. Section 9 discusses open problems and future work. Sections 10 and 11 contain the major proofs from Sections 4 and 6. The P-Norm Push was recently applied to the problem of prioritizing manholes in New York City for maintenance and repair (Rudin et al., 2009).

The main contributions of this work are: a generalization bound for a learning problem that is asymmetric by design, a simple user-adjustable, easy-to-implement algorithm for supervised ranking with a "push," and a proof that the minimizer of the algorithm's objective is unique in a specific sense.

2. An Objective for Ranking with a Push

The set of instances with positive labels is $\{\mathbf{x}_i\}_{i=1,\ldots,I}$, where $\mathbf{x}_i \in \mathcal{X}$. The negative instances are $\{\tilde{\mathbf{x}}_k\}_{k=1,\ldots,K}$, where $\tilde{\mathbf{x}}_k \in \mathcal{X}$. We always use i for the index over positive instances and k for the index over negative instances. In the case of the movie ranking problem, the $\mathbf{x}_i$'s are the good movies used for training, the $\tilde{\mathbf{x}}_k$'s are the bad movies, and $\mathcal{X}$ is a database of movies. Our goal is to construct a ranking function f that gives a real valued score to each instance in $\mathcal{X}$, that is, $f : \mathcal{X} \to \mathcal{R}$. We do

not care about the actual values of each instance, only the relative values; for positive-negative pair $\mathbf{x}_i, \tilde{\mathbf{x}}_k$, we care that $f(\mathbf{x}_i) > f(\tilde{\mathbf{x}}_k)$ but it is not important to know, for example, that $f(\mathbf{x}_i) = 0.4$ and $f(\tilde{\mathbf{x}}_k) = 0.1$.

Let us now derive the general form of our objective. For a particular negative example, we wish to reduce its *Height*, which is the number of positive examples ranked beneath it. That is, for each k, we wish to make Height(k) small, where:

$$\text{Height}(k) := \sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}.$$

Let us now add the push. We want to concentrate harder on negative examples that have high scores; we want to push these examples down from the top. Since the highest scoring negative examples also achieve the largest Heights, these are the examples for which we impose a larger price. Namely, for convex, non-negative, monotonically increasing function $g : \mathcal{R}_+ \to \mathcal{R}_+$, we place the price $g(\text{Height}(k))$ on negative example k:

$$g\left(\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}\right).$$

If g is very steep, we pay an extremely large price for a high scoring negative example. Examples of steep functions include $g(r) = e^r$ and $g(r) = r^p$ for p large. Thus we have derived an objective to minimize, namely the sum of the prices for the negative examples:

$$R_{g,\mathbf{1}}(f) := \sum_{k=1}^{K} g\left(\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}\right).$$

The effect of g is to force the value of $R_{g,\mathbf{1}}$ to come mostly from the highest scoring negative examples. These high scoring negative examples are precisely the examples represented by the leftmost portion of the ROC Curve. Minimizing $R_{g,\mathbf{1}}$ should thus boost performance around high scoring negative examples and increase the leftmost portion of the ROC Curve.

It is hard to minimize $R_{g,\mathbf{1}}$ directly due to the 0-1 loss in the inner sum. Instead, we will minimize an upper bound, $R_{g,\ell}$, which incorporates $\ell : \mathcal{R} \to \mathcal{R}_+$, a convex, non-negative, monotonically decreasing upper bound on the 0-1 loss. Popular loss functions include the exponential, logistic, and hinge losses. We can now define the general form of our objective:

$$R_{g,\ell}(f) := \sum_{k=1}^{K} g\left(\sum_{i=1}^{I} \ell\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right).$$

To construct a specific version of this objective, one chooses the loss ℓ, the price function g, and an appropriate hypothesis space $\mathcal{F}$ over which to minimize $R_{g,\ell}$. In order to derive RankBoost's specific objective from $R_{g,\ell}$, we would choose ℓ as the exponential loss and g to be the identity.

For the moment, let us assume we care only about the very top of the list, that is, we wish to push the most offending negative example as far down the list as possible. Equivalently, we wish to minimize $R_{\max}$, the number of positives below the highest scoring negative example:

$$R_{\max}(f) := \max_k \sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}.$$

Minimizing this misranking error at the very top is similar to optimizing a "winner take all" loss such as $\mathbf{1}_{[\max_i f(\mathbf{x}_i) \leq \max_k f(\tilde{\mathbf{x}}_k)]}$ in that both would choose a ranked list where a negative example is not at the top of the list.

Although it is hard to minimize $R_{\max}(f)$ directly, $R_{g,\ell}$ can give us some control over $R_{\max}$. Namely, the following relationships exist between $R_{g,\ell}$, $R_{g,\mathbf{1}}$ and $R_{\max}$.

Theorem 1 *For all convex, non-negative, monotonic g and for all ℓ that are upper bounds for the 0-1 loss, we have that:*

$$Kg\left(\frac{1}{K}R_{\max}(f)\right) \leq R_{g,\mathbf{1}}(f) \leq Kg\Big(R_{\max}(f)\Big) \quad \text{and} \quad R_{g,\mathbf{1}}(f) \leq R_{g,\ell}(f).$$

Proof The proof of the first inequality follows from the monotonicity of g and Jensen's inequality for convex function g.

$$\begin{aligned} Kg\left(\frac{1}{K}R_{\max}(f)\right) &= Kg\left(\frac{1}{K}\max_k \sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) \leq Kg\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) \\ &\leq \sum_{k=1}^{K} g\left(\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) = R_{g,\mathbf{1}}(f). \end{aligned}$$

For the second inequality, we use the fact that g is monotonic:

$$\begin{aligned} R_{g,\mathbf{1}}(f) &= \sum_{k=1}^{K} g\left(\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) \leq K\max_k g\left(\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) \\ &= Kg\left(\max_k \sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) = Kg\Big(R_{\max}(f)\Big). \end{aligned}$$

Using that ℓ is an upper bound on the 0-1 loss, we have the last inequality:

$$R_{g,\mathbf{1}}(f) = \sum_{k=1}^{K} g\left(\sum_{i=1}^{I} \mathbf{1}_{[f(\mathbf{x}_i)\leq f(\tilde{\mathbf{x}}_k)]}\right) \leq \sum_{k=1}^{K} g\left(\sum_{i=1}^{I} \ell\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right) = R_{g,\ell}(f).$$

■

The fact that the function $Kg(\frac{1}{K}r)$ is monotonic in r adds credibility to our choice of objective $R_{g,\ell}$; if $R_{g,\ell}(f)$ is minimized, causing a reduction in $Kg(\frac{1}{K}R_{\max}(f))$, then $R_{\max}(f)$ will also be reduced. Thus, Theorem 1 suggests that $R_{g,\ell}$ is a reasonable quantity to minimize in order to incorporate a push at the top, for instance, in order to diminish $R_{\max}$. Also recall that if g is especially steep, for instance $g(r) = e^r$ or $g(r) = r^p$ for p large, then $g^{-1}(\sum_{k=1}^{K} g(r_k)) \approx \max_k r_k$. That is, the quantity $g^{-1}(R_{g,\mathbf{1}})$, for steep functions g, will approximate $R_{\max}$.

For most of the paper, we are considering the power law (or "p-norm") price functions $g(r) = r^p$. By allowing the user to choose p, we allow the amount of push to be specified to match the application. At the heart of this derivation, we are using ℓ_p-norms to interpolate between the ℓ_1-norm (the AUC), and the ℓ_∞-norm (the values of $R_{\max}$). In what follows, we overload notation by defining $R_{p,\ell}$ to denote $R_{g,\ell}$ where $g(r) = r^p$:

$$R_{p,\ell}(f) := \sum_{k=1}^{K} \left(\sum_{i=1}^{I} \ell\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p.$$

Thus, $R_{p,\ell}^{1/p}(f) \to R_{\max,\ell}(f)$ as $p \to \infty$, where $R_{\max,\ell}(f) := \max_k \sum_{i=1}^{I} \ell\big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\big)$.

As we will discuss, the choice of p should depend on the number of examples. More examples are needed for generalization if a larger value of p is chosen.

3. Illustrating That It Works

In this section, we will give some examples to illustrate how the objective concentrates on the top of the list when p is large, or more generally, when g is steep.

3.1 First Illustration: Swap on the Bottom vs. Swap on the Top

For our first illustration, we aim simply to show that the objective function we have derived really does care more about the top of the list than the rest. Consider the set of examples $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_8$ with vector of labels:

$$(-1, +1, -1, +1, -1, -1, +1, +1).$$

Consider scoring function f_{orig} which gives the scores: $f_{\text{orig}}(\mathbf{x}_i) = i$ for all i. Placing the labels in rank order of f_{orig} yields:

$$\text{labels in original rank order: } (-1 \quad +1 \quad -1 \quad +1 \quad -1 \quad -1 \quad +1 \quad +1) \ .$$

Using the power law $g(r) = r^4$ for the price function, we can compute the value of $R_{4,\mathbf{1}}(f_{\text{orig}})$ for this ranked list: $0^4 + 1^4 + 2^4 + 2^4 = 33$.

Now consider $f_{\text{swapOnBot}}$ which swaps the scores of a pair of examples at the bottom of the ranked list, $f_{\text{swapOnBot}}(\mathbf{x}_1) = 2$, $f_{\text{swapOnBot}}(\mathbf{x}_2) = 1$, and $f_{\text{swapOnBot}}(\mathbf{x}_i) = i$ for all other i. The new rank ordering of labels is:

$$\text{swap on the bottom: } (+1 \quad -1 \quad -1 \quad +1 \quad -1 \quad -1 \quad +1 \quad +1) \ .$$

Here a negative example is ranked above one more positive example than before. Computing the value of $R_{4,\mathbf{1}}(f_{\text{swapOnBot}})$ yields $1^4 + 1^4 + 2^4 + 2^4 = 34 > 33$; the value of $R_{4,\mathbf{1}}$ changes slightly when a swap is made at the bottom of the list, only from 33 to 34. Let us now instead consider a swap near the top of the list, so that the new set of labels is again only one swap away from the original, $f_{\text{swapOnTop}}(\mathbf{x}_6) = 7$, $f_{\text{swapOnTop}}(\mathbf{x}_7) = 6$, and $f_{\text{swapOnTop}}(\mathbf{x}_i) = i$ for all other i. The new ordering of labels is:

$$\text{swap on the top: } (-1 \quad +1 \quad -1 \quad +1 \quad -1 \quad +1 \quad -1 \quad +1) \ .$$

Here, the value of $R_{4,\mathbf{1}}(f_{\text{swapOnTop}})$ is $0^4 + 1^4 + 2^4 + 3^4 = 98 \gg 33$. So, in both cases only one swap was made between neighboring examples; however, the swap at the top of the list changed the objective dramatically (from 33 to 98) while the swap at the bottom hardly changed the objective at all (from 33 to 34). So, we have now illustrated that the objective function $R_{p,\mathbf{1}}(f)$ concentrates at the top of the list.

The same behavior occurs using different loss functions ℓ. This is summarized in Table 1 for three loss functions: the 0-1 loss which we have just explained, the exponential loss $\ell(r) = e^{-r}$, and the logistic loss $\ell(r) = \log(1 + e^{-r})$. (Note that using natural log for the logistic loss does not give an upper bound on the 0-1 loss, it is off by a multiplicative factor that is irrelevant in experiments.)

$\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7, \mathbf{x}_8$ y: $(-1,+1,-1,+1,-1,-1,+1,+1)$	$R_{4,\mathbf{1}}(f)$	$R_{4,\exp}(f)$	$R_{4,\text{logistic}}(f)$
labels ordered by f_{orig} : $(-1,+1,-1,+1,-1,-1,+1,+1)$	33	17,160.17	430.79
labels ordered by $f_{\text{swapOnBot}}$: $(+1,-1,-1,+1,-1,-1,+1,+1)$	34	72,289.39	670.20
labels ordered by $f_{\text{swapOnTop}}$: $(-1,+1,-1,+1,-1,+1,-1,+1)$	98	130,515.09	1,212.23

Table 1: Values of the objective function $R_{4,\ell}$ for the three slightly different labelings, using the 0-1 loss (column $R_{4,\mathbf{1}}$), exponential loss (column $R_{4,\exp}$), and logistic loss (column $R_{4,\text{logistic}}$). The objective functions change much more in reaction to the swap at the top of the list: the values in the third row (swap on the top) are significantly higher than those in the second row (swap on the bottom).

3.2 A Second Illustration: Reversal of Polarity

Let us assume we want to choose a scoring function f by minimizing our objective $R_{p,\ell}(f)$ over $f \in \mathcal{F}$ where $\mathcal{F}$ has only two functions, $\mathcal{F} = \{f_1, f_2\}$. This is an interesting experiment in which there are only 2 choices available for the function f: the first concentrates on the top of the ranked list, but performs poorly on the rest, whereas the second performs badly on the top of the ranked list, but performs well over all. In fact, the second scoring function f_2 is exactly a negation of the first scoring function f_1. Here are the labels and hypotheses:

labels	$+1$	$+1$	-1	-1	-1	-1	-1	$+1$	$+1$	$+1$	$+1$	$+1$	-1	-1	
f_1 :	(14	13	12	11	10	9	8	7	6	5	4	3	2	1	)/14
f_2 :	$(-14$	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	)/14

Here, f_1 performs well at the top of the list (the two top-scoring examples are positive), but the whole middle of the list is reversed; there are 5 negative examples in a row, and below that 5 positives. On the other hand, f_2 misses the top two examples which have scores -1/14 and -2/14, however, the 10 middle examples are correctly ranked. f_2 has a larger AUC than f_1, but f_1 is better at the top of the list. Now, which of f_1 and f_2 would the misranking objectives from Section 2 prefer? Let us answer this for various $R_{p,\ell}$, for different p and ℓ. Specifically, we will demonstrate that as p becomes larger, $R_{p,\ell}$ prefers the first hypothesis which performs better at the top. Table 2 shows values of $R_{p,\ell}$ for three different loss functions and for various values of p. This table shows that for smaller p, f_2 is preferred. At some value of p, the "polarity" reverses and then f_1 is preferred. So, using steeper price functions means that we are more likely to prefer scoring functions that perform well at the top of the list.

p	$R_{p,\mathbf{1}}(f_1)$	$R_{p,\mathbf{1}}(f_2)$	$\operatorname{argmin}_{f\in\{f_1,f_2\}} R_{p,\mathbf{1}}(f)$	$R_{p,\exp}(f_1)$	$R_{p,\exp}(f_2)$	$\operatorname{argmin}_{f\in\{f_1,f_2\}} R_{p,\exp}(f)$
1	25	24	f_2	50.25	49.80	f_2
2	125	118	f_2	367.39	362.35	f_2
3	625	726	f_1	$2.73*10^3$	$2.70*10^3$	f_2
4	$3.13*10^3$	$4.88*10^3$	f_1	$2.056*10^4$	$2.057*10^4$	f_1
5	$1.56*10^4$	$3.38*10^4$	f_1	$1.56*10^5$	$1.60*10^5$	f_1
6	$7.81*10^4$	$23.56*10^4$	f_1	$1.20*10^6$	$1.28*10^6$	f_1
7	$3.91*10^5$	$16.48*10^5$	f_1	$9.34*10^6$	$10.36*10^6$	f_1
8	$1.95*10^6$	$11.53*10^6$	f_1	$7.29*10^7$	$8.53*10^7$	f_1
9	$9.77*10^6$	$80.71*10^6$	f_1	$5.72*10^8$	$7.13*10^8$	f_1
10	$4.88*10^7$	$56.50*10^7$	f_1	$4.50*10^9$	$6.02*10^9$	f_1

p	$R_{p,\text{logistic}}(f_1)$	$R_{p,\text{logistic}}(f_2)$	$\operatorname{argmin}_{f\in\{f_1,f_2\}} R_{p,\text{logistic}}(f)$
1	34.34	34.09	f_2
2	170.18	167.90	f_2
3	851.09	836.46	f_2
4	$4.29*10^3$	$4.22*10^3$	f_2
5	$2.18*10^4$	$2.15*10^4$	f_2
6	$1.114*10^5$	$1.110*10^5$	f_2
7	$5.72*10^5$	$5.79*10^5$	f_1
8	$2.96*10^6$	$3.05*10^6$	f_1
9	$1.53*10^7$	$1.63*10^7$	f_1
10	$7.98*10^7$	$8.74*10^7$	f_1

Table 2: This table shows that as the price function gets steeper (as p increases), the scoring function f_1 that performs better on the top of the list is preferred. We show the values for each of the objectives $R_{p,\mathbf{1}}$, $R_{p,\exp}$ and $R_{p,\text{logistic}}$ for $p = 1,\ldots,10$ applied to f_1 (first column) and f_2 (second column). The third column shows which of the two scoring functions f_1 or f_2 achieve a lower value of the objective.

3.3 Third Illustration: Contribution of Each Positive-Negative Pair

Consider the following list of labels and function values :

$$\begin{array}{rl} y: & (1\ \ 1\ -1\ \ 1\ \ 1\ -1\ \ 1\ \ 1\ -1\ \ 1\ \ 1\ -1\ \ 1\ \ 1\ -1\ -1\ -1) \\ f: & (20\ 19\ 18\ 17\ 16\ 15\ 14\ 13\ 12\ 11\ 10\ \ 9\ \ 8\ \ 7\ \ 6\ \ 5\ \ 4\ \ 3\ \ 2\ \ 1)/20 \end{array}$$

Figure 1 illustrates the amount that each positive-negative pair contributes to $R_{p,\exp}$ for various values of p. We aim to show that $R_{p,\exp}$ becomes more influenced by the highest scoring negative examples as p is increased. On the vertical axis are the positive examples $i = 1,\ldots,12$ ordered by score, with the highest scoring examples at the bottom. On the horizontal axis are the negative examples $k = 1,\ldots,8$ ordered by score, with the highest scoring examples on the

left. The value of the $(i,k)^{th}$ entry is the contribution of the k^{th} highest scoring negative example, $\left(\sum_{\bar{i}} e^{-(f(\mathbf{x}_{\bar{i}})-f(\tilde{\mathbf{x}}_k))}\right)^p$, multiplied by the proportion attributed to the i^{th} highest scoring positive example, $e^{-(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k))}/\sum_{\bar{i}} e^{-(f(\mathbf{x}_{\bar{i}})-f(\tilde{\mathbf{x}}_k))}$. As we adjust the value of p, one can see that most of the contribution shifts towards the left, or equivalently, towards the highest scoring negative examples.

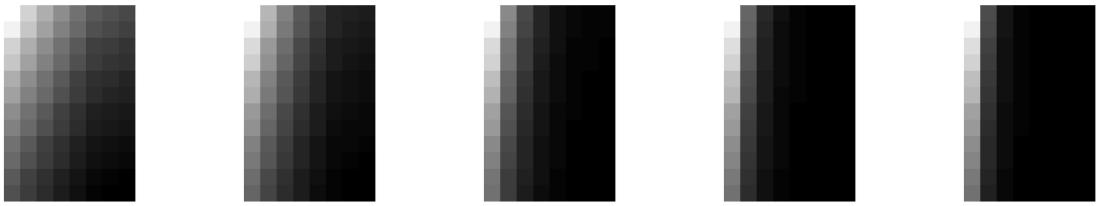

Figure 1: Contribution of each positive-negative pair to the objective $R_{p,\text{exp}}$. Each square represents an i,k pair, where i is an index along the vertical axis, and k is along the horizontal axis. Lighter colors indicate larger contributions to $R_{p,\text{exp}}$. The upper left corner represents the highest (worst) ranked negative and the lowest (worst) ranked positive.

4. A Generalization Bound for $R_{p,\mathbf{1}}$

We present two bounds, where the second has better dependence on p than the first. A preliminary version of the first bound appears in the conference version of this paper (Rudin, 2006). This work is inspired by the works of Koltchinskii and Panchenko (2002), Cucker and Smale (2002), and Bousquet (2003).

Assume that the positive instances $\mathbf{x}_i \in \mathcal{X}$, $i = 1,...,I$ are chosen independently and at random (iid) from a fixed but unknown probability distribution $\mathcal{D}_+$ on $\mathcal{X}$. Assume the negative instances $\tilde{\mathbf{x}}_k \in \mathcal{X}$, $k = 1,...,K$ are chosen iid from $\mathcal{D}_-$. The notation $\mathbf{x} \sim \mathcal{D}$ means $\mathbf{x}$ is chosen randomly according to distribution $\mathcal{D}$. The notation $S_+ \sim \mathcal{D}_+^I$ means each of the I elements of the training set S_+ are chosen independently at random according to $\mathcal{D}_+$. Similarly for $S_- \sim \mathcal{D}_-^K$.

We now define the "true" objective function for the underlying distribution:

$$\begin{aligned} R^{\text{true}}_{p,\mathbf{1}}(f) &:= \left(\mathbb{E}_{\mathbf{x}_-\sim\mathcal{D}_-}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\mathbf{1}_{[f(\mathbf{x}_+)-f(\mathbf{x}_-)\le 0]}\right)^p\right)^{1/p} \\ &= \left\|\mathbb{P}_{\mathbf{x}_+\sim\mathcal{D}_+}\left(f(\mathbf{x}_+)-f(\mathbf{x}_-)\le 0|\mathbf{x}_-\right)\right\|_{L_p(\mathcal{X}_-,\mathcal{D}_-)}. \end{aligned}$$

The empirical loss associated with $R^{\text{true}}_{p,\mathbf{1}}(f)$ is the following:

$$R^{\text{empirical}}_{p,\mathbf{1}}(f) := \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I}\mathbf{1}_{[f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k)\le 0]}\right)^p\right)^{1/p}.$$

Here, for a particular $\tilde{\mathbf{x}}_k$, $R^{\text{empirical}}_{p,\mathbf{1}}(f)$ takes into account the average number of positive examples that have scores below $\tilde{\mathbf{x}}_k$. It is a monotonic function of $R_{p,\mathbf{1}}$. To make this notion more general, let

us consider the average number of positive examples that have scores that are *close to* or below $\tilde{\mathbf{x}}_k$. A more general version of $R^{\text{empirical}}_{p,\mathbf{1}}(f)$ is thus defined as:

$$R^{\text{empirical}}_{p,\mathbf{1},\theta}(f) := \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I}\mathbf{1}_{[f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k)\leq\theta]}\right)^p\right)^{1/p}.$$

This terminology incorporates the "margin" value θ. As before, we suffer some loss whenever positive example $\mathbf{x}_i$ is ranked below negative example $\tilde{\mathbf{x}}_k$, but now we also suffer loss whenever $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_k$ have scores within θ of each other. Note that $R^{\text{empirical}}_{p,\mathbf{1},\theta}$ is an empirical quantity, so it can be measured for any θ. We will state two bounds, proved in Section 10, where the second is tighter than the first. The first bound is easier to understand and is a direct corollary of the second bound.

Theorem 2 *(First Generalization Bound) For all $\varepsilon > 0, p \geq 1$, $\theta > 0$, the probability over random choice of training set, $S_+ \sim \mathcal{D}_+^I, S_- \sim \mathcal{D}_-^K$ that there exists an $f \in \mathcal{F}$ such that*

$$R^{\text{true}}_{p,\mathbf{1}}(f) \geq R^{\textit{empirical}}_{p,\mathbf{1},\theta}(f) + \varepsilon$$

is at most:

$$2\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon\theta}{8}\right)\left(\exp\left[-2\left(\frac{\varepsilon}{4}\right)^{2p}K\right] + \exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right]\right).$$

Here the covering number $\mathcal{N}(\mathcal{F},\varepsilon)$ is defined as the number of ε-sized balls needed to cover $\mathcal{F}$ in L_∞, and it is used here as a complexity measure for $\mathcal{F}$. This expression states that, provided I and K are large, then with high probability, the true error $R^{\text{true}}_{p,\mathbf{1}}(f)$ is not too much more than the empirical error $R^{\text{empirical}}_{p,\mathbf{1},\theta}(f)$.

It is important to note the implications of this bound for scalability. More examples are required for larger p. This is because we are concentrating on a small portion of input space corresponding to the top of the ranked list. If most of the value of $R^{\text{true}}_{p,\mathbf{1}}$ comes from a small portion of input space, it is necessary to have more examples in that part of the space in order to estimate its value with high confidence. The fact that more examples are required for large p can affect performance in practice. A 1-dimensional demonstration of this fact is given at the end of Section 10.

Theorem 2 shows that the dependence on p is important for generalization. The following theorem shows that in most circumstances, we have much better dependence on p. Specifically, the dependence can be shifted from $-\varepsilon^{2p}$ in the exponential to a factor related to $-\varepsilon^2\left(\inf_f R^{\text{true}}_{p,\mathbf{1}}(f)\right)^{2(p-1)}$. The bound becomes much tighter than Theorem 2 when all hypotheses have a large enough true risk, that is, when $\inf_f R^{\text{true}}_{p,\mathbf{1}}(f)$ is large compared to ε.

Theorem 3 *(Second Generalization Bound) For all $\varepsilon > 0, p \geq 1$, $\theta > 0$, the probability over random choice of training set, $S_+ \sim \mathcal{D}_+^I, S_- \sim \mathcal{D}_-^K$ that there exists an $f \in \mathcal{F}$ such that*

$$R^{\text{true}}_{p,\mathbf{1}}(f) \geq R^{\textit{empirical}}_{p,\mathbf{1},\theta}(f) + \varepsilon$$

is at most:

$$2\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon\theta}{8}\right)\left(\exp\left[-2K\max\left\{\frac{\varepsilon^2}{16}\left(R_{p,\min}\right)^{2(p-1)}, \left(\frac{\varepsilon}{4}\right)^{2p}\right\}\right] + \exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right]\right).$$

where $R_{p,\min} := \inf_{f\in\mathcal{F}} R^{\text{true}}_{p,\mathbf{1}}(f)$.

The proof is in Section 10. The dependence on p is now much better than in Theorem 2. It is possible that the bound can be tightened in other ways, for instance, to use a different type of covering number. For instance, one might use the "sloppy covering number" in Rudin and Schapire (2009)'s ranking bound, which is adapted from the classification bound of Schapire et al. (1998).

The purpose of Theorems 2 and 3 is to provide the theoretical justification required for our choice of objective, provided a sufficient number of training examples. Having completed this, let us now write an algorithm for minimizing that objective.

5. A Boosting-Style Algorithm

We now choose a specific form for our objective $R_{g,\ell}$ by choosing ℓ. We have already chosen g to be a power law, $g(r) = r^p$. From now on, ℓ will be the exponential loss $\ell(r) = e^{-r}$. One could just as easily choose another loss; we choose the exponential loss in order to compare with RankBoost. The objective when $p = 1$ is exactly that of RankBoost, whose global objective is $R_{1,\exp}$. Here is the objective function, $R_{p,\exp}$ for $p \geq 1$:

$$R_{p,\exp}(f) := \sum_{k=1}^{K} \left(\sum_{i=1}^{I} e^{-(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k))} \right)^p .$$

The function f is constructed as a linear combination of "weak rankers" or "ranking features," $\{h_j\}_{j=1,\ldots,n}$, with $h_j : \mathcal{X} \to [0,1]$ so that $f = \sum_j \lambda_j h_j$, where $\boldsymbol{\lambda} \in \mathcal{R}^n$. Thus, the hypothesis space $\mathcal{F}$ is the class of convex combinations of weak rankers. Our objective is now $R_{p,\exp}(\boldsymbol{\lambda})$:

$$R_{p,\exp}(\boldsymbol{\lambda}) := \sum_{k=1}^{K} \left(\sum_{i=1}^{I} e^{-\left(\sum_j \lambda_j h_j(\mathbf{x}_i) - \sum_j \lambda_j h_j(\tilde{\mathbf{x}}_k)\right)} \right)^p = \sum_{k=1}^{K} \left(\sum_{i=1}^{I} e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}} \right)^p ,$$

where we have rewritten in terms of a matrix $\mathbf{M}$, which describes how each individual weak ranker j ranks each positive-negative pair $\mathbf{x}_i, \tilde{\mathbf{x}}_k$; this will make notation significantly easier. Define an index set that enumerates all positive-negative pairs $\mathcal{C}_p = \{ik : i \in 1,\ldots,I, k \in 1,\ldots,K\}$ where index ik corresponds to the i^{th} positive example and the k^{th} negative example. Formally,

$$M_{ik,j} := h_j(\mathbf{x}_i) - h_j(\tilde{\mathbf{x}}_k).$$

The size of $\mathbf{M}$ is $|\mathcal{C}_p| \times n$. The notation $(\cdot)_a$ means the a^{th} index of the vector, that is,

$$(\mathbf{M}\boldsymbol{\lambda})_{ik} := \sum_{j=1}^{n} M_{ik,j}\lambda_j = \sum_{j=1}^{n} \lambda_j h_j(\mathbf{x}_i) - \lambda_j h_j(\tilde{\mathbf{x}}_k).$$

The function $R_{p,\exp}(\boldsymbol{\lambda})$ is convex in $\boldsymbol{\lambda}$. This is because $e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}}$ is a convex function of $\boldsymbol{\lambda}$, any sum of convex functions is convex, and a composition of an increasing convex function with a convex function is convex. (Note that $R_{p,\exp}(\boldsymbol{\lambda})$ is convex but not necessarily strictly convex.)

We now derive a boosting-style coordinate descent algorithm for minimizing $R_{p,\exp}$ as a function of $\boldsymbol{\lambda}$. At each iteration of the algorithm, the coefficient vector $\boldsymbol{\lambda}$ is updated. At iteration t, we denote the coefficient vector by $\boldsymbol{\lambda}_t$. There is much background material available on the convergence of similar coordinate descent algorithms (for instance, see Zhang and Yu, 2005). We start with the objective at iteration t:

$$R_{p,\exp}(\boldsymbol{\lambda}_t) := \sum_{k=1}^{K} \left(\sum_{i=1}^{I} e^{(-\mathbf{M}\boldsymbol{\lambda}_t)_{ik}} \right)^p .$$

We then compute the variational derivative along each "direction" and choose weak ranker j_t to have largest absolute variational derivative. The notation $\mathbf{e}_j$ means a vector of 0's with a 1 in the j^{th} entry.

$$j_t \in \operatorname*{argmax}_j \left[-\frac{dR_{p,\exp}(\boldsymbol{\lambda}_t + \alpha \mathbf{e}_j)}{d\alpha}\Big|_{\alpha=0} \right], \text{ where}$$

$$\frac{dR_{p,\exp}(\boldsymbol{\lambda}_t + \alpha \mathbf{e}_j)}{d\alpha}\Big|_{\alpha=0} = p \sum_{k=1}^{K} \left[\left(\sum_{i=1}^{I} e^{(-\mathbf{M}\boldsymbol{\lambda}_t)_{ik}} \right)^{p-1} \left(\sum_{i=1}^{I} -M_{ik,j} e^{-(\mathbf{M}\boldsymbol{\lambda}_t)_{ik}} \right) \right].$$

Define the vector $\mathbf{q}_t$ on pairs i,k as $q_{t,ik} := e^{(-\mathbf{M}\boldsymbol{\lambda}_t)_{ik}}$, and the weight vector $\mathbf{d}_t$ as $d_{t,ik} := q_{t,ik} / \sum_{ik} q_{t,ik}$. Our choice of j_t becomes (ignoring constant factors that do not affect the argmax):

$$\begin{aligned} j_t \quad &\in \quad \operatorname*{argmax}_j \sum_{k=1}^{K} \left[\left(\sum_{i=1}^{I} d_{t,ik} \right)^{p-1} \sum_{i=1}^{I} d_{t,ik} M_{ik,j} \right] \\ &= \quad \operatorname*{argmax}_j \sum_{ik} \tilde{d}_{t,ik} M_{ik,j}, \text{ where } \tilde{d}_{t,ik} = d_{t,ik} \left(\sum_{i'} d_{t,i'k} \right)^{p-1}. \end{aligned}$$

To update the coefficient of weak ranker j_t, we now perform a linesearch for the minimum of $R_{p,\exp}$ along the j_t^{th} direction. The distance to travel in the j_t^{th} direction, denoted α_t, solves $0 = \frac{dR_{p,\exp}(\boldsymbol{\lambda}_t + \alpha \mathbf{e}_{j_t})}{d\alpha}\Big|_{\alpha_t}$. Ignoring division by constants, this equation becomes:

$$0 = \sum_{k=1}^{K} \left[\left(\sum_{i=1}^{I} d_{t,ik} e^{-\alpha_t M_{ik,j_t}} \right)^{p-1} \left(\sum_{i=1}^{I} M_{ik,j_t} d_{t,ik} e^{-\alpha_t M_{ik,j_t}} \right) \right]. \tag{1}$$

The value of α_t can be computed analytically in some cases, for instance, when the weak rankers are binary-valued and $p = 1$ (this is RankBoost). Otherwise, we simply use a linesearch to solve this equation for α_t. To complete the algorithm, we set $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \alpha_t \mathbf{e}_{j_t}$. To avoid having to compute $\mathbf{d}_{t+1}$ directly from $\boldsymbol{\lambda}_t$, we can perform the update by:

$$d_{t+1,ik} = \frac{d_{t,ik} e^{-\alpha_t M_{ik,j_t}}}{z_t} \quad \text{where} \quad z_t := \sum_{ik} d_{t,ik} e^{-\alpha_t M_{ik,j_t}}.$$

The full algorithm is shown in Figure 2. This implementation is not optimized for very large data sets since the size of $\mathbf{M}$ is $|C_p| \times n$. Note that the weak learning part of this algorithm in Step 3(a), when written in this form, is the same as for AdaBoost and RankBoost. Thus, any current implementation of a weak learning algorithm for AdaBoost or RankBoost can be directly used for the P-Norm Push.

6. Uniqueness of the Minimizer

We now show that a function $f = \sum_j \lambda_j h_j$ (or limit of functions) minimizing our objective is unique in some sense. Since $\mathbf{M}$ is not required to be invertible (and often is not), we cannot expect to find a unique vector $\boldsymbol{\lambda}$; one may achieve the identical values of $(\mathbf{M}\boldsymbol{\lambda})_{ik}$ with different choices of $\boldsymbol{\lambda}$. It is also true that elements of $\boldsymbol{\lambda}_t$ may approach $\pm\infty$, and furthermore, elements of $\mathbf{M}\boldsymbol{\lambda}_t$ often approach

1. **Input:** $\{\mathbf{x}_i\}_{i=1,...,I}$ positive examples, $\{\tilde{\mathbf{x}}_k\}_{k=1,...,K}$ negative examples, $\{h_j\}_{j=1,...,n}$ weak classifiers, $t_{\max}$ number of iterations, p power.

2. **Initialize:** $\lambda_{1,j} = 0$ for $j = 1,...,n$, $d_{1,ik} = 1/IK$ for $i = 1,...,I$, $k = 1,...,K$ $M_{ik,j} = h_j(\mathbf{x}_i) - h_j(\tilde{\mathbf{x}}_k)$ for all i,k,j

3. **Loop for** $t = 1,...,t_{\max}$

   (a) $j_t \in \text{argmax}_j \sum_{ik} \tilde{d}_{t,ik} M_{ik,j}$ where $\tilde{d}_{t,ik} = d_{t,ik} \left(\sum_{i'} d_{t,i'k}\right)^{p-1}$

   (b) Perform a linesearch for α_t. That is, find a value α_t that solves (1).

   (c) $\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \alpha_t \mathbf{e}_{j_t}$, where $\mathbf{e}_{j_t}$ is 1 in position j_t and 0 elsewhere.

   (d) $z_t = \sum_{ik} d_{t,ik} e^{-\alpha_t M_{ik,j_t}}$

   (e) $d_{t+1,ik} = d_{t,ik} e^{-\alpha_t M_{ik,j_t}} / z_t$ for $i = 1,...,I$, $k = 1,...,K$

4. **Output:** $\boldsymbol{\lambda}_{t_{\max}}$

Figure 2: Pseudocode for the "P-Norm Push" algorithm.

$+\infty$, so it would seem difficult to prove (or even define) uniqueness. A trick that comes in handy for such situations is to use the closure of the space $Q' := \{\mathbf{q}' \in \mathcal{R}_+^{IK} | q'_{ik} = e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}} \text{ for some } \boldsymbol{\lambda} \in \mathcal{R}^n\}$. The closure of Q' includes the limits where $\mathbf{M}\boldsymbol{\lambda}_t$ becomes infinite, and considers the linear combination of hypotheses $\mathbf{M}\boldsymbol{\lambda}$ rather than $\boldsymbol{\lambda}$ itself, so it does not matter whether $\mathbf{M}$ is invertible. With the help of convex analysis, we will be able to show that our objective function yields a unique minimizer in the closure of Q'. Here is our uniqueness theorem:

Theorem 4 *Define $Q' := \{\mathbf{q}' \in \mathcal{R}_+^{IK} | q'_{ik} = e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}}$ for some $\boldsymbol{\lambda} \in \mathcal{R}^n\}$ and define $\bar{Q}'$ as the closure of Q' in $\mathcal{R}^{IK}$. Then for $p \geq 1$, there is a unique $\mathbf{q}'^* \in \bar{Q}'$ where:*

$$\mathbf{q}'^* = \text{argmin}_{\mathbf{q}' \in \bar{Q}'} \sum_k \left(\sum_i q'_{ik} \right)^p .$$

Our uniqueness proof depends mainly on the theory of convex duality for a class of Bregman distances, as defined by Della Pietra et al. (2002). This proof is inspired by Collins et al. (2002) who have proved uniqueness of this type for AdaBoost. In the case of AdaBoost, the primal optimization problem corresponds to a minimization over relative entropy. Our case is more unusual and the primal is not a common function. The proof of Theorem 4 is located in Section 11.

7. Variations of the Objective and Relationship to Information Retrieval Measures

It is possible to use variations of our basic derivation in Section 2 to derive other specialized objectives. Some of these objectives are similar to current popular quality measurements from information retrieval (IR), such as the "discounted cumulative gain" (DCG) (Järvelin and Kekäläinen, 2000). A basic property of this quality measurement, and additionally the average precision (the mean of precision values), is that it is proportional to a sum over relevant documents (which are the

positive examples in this setting), and uses a discounting factor that decreases according to the rank of a relevant document. The discounting factor here is analogous to the price function. Let us use the framework we have developed to derive new quality measurements with these properties.

Our derivation in Section 2 is designed to push the highly ranked negative examples down. Rearranging this argument, we can also pull the positive examples up, using the "reverse height." The reverse height of positive example i is the number of negative examples ranked above it.

$$\text{Reverse Height}(i) := \sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}.$$

The reverse height is very similar to the rank used in the IR quality measurements. The reverse height only considers the relationship of the positives to the negatives, and disregards the relationship of positives to each other. Precisely, define:

$$\text{Rank}(i) := \sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]} + \sum_{\bar{i}} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\mathbf{x}_{\bar{i}})]} = \text{Reverse Height}(i) + \sum_{\bar{i}} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\mathbf{x}_{\bar{i}})]}.$$

The rank can often be substituted for the reverse height. For discounting factor $g : \mathcal{R}_+ \to \mathcal{R}_+$, consider the variations of our objective:

$$R_{g,\mathbf{1}}^{\text{Reverse Height}}(f) = \sum_i g\left(\text{Reverse Height}(i)\right) = \sum_i g\left(\sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}\right).$$

$$R_{g,\mathbf{1}}^{\text{Rank}}(f) = \sum_i g\left(\text{Rank}(i)\right) = \sum_i g\left(\sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]} + \sum_{\bar{i}} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\mathbf{x}_{\bar{i}})]}\right).$$

Then, one might maximize $R_{g,\mathbf{1}}^{\text{Rank}}$ for various g. The function g should achieve the largest values for the positive examples i that possess the smallest reverse heights or ranks, since those are at the top of the list. It should thus be a decreasing function with steep negative slope near the y-axis. Choosing $g(z) = 1/z$ gives the average value of $1/\text{rank}$. Choosing $g(z) = 1/\ln(1+z)$ gives the discounted cumulative gain:

$$\text{AveR}(f) = \sum_i \frac{1}{\text{Rank}(i)} = \sum_i \frac{1}{\sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]} + \sum_{\bar{i}} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\mathbf{x}_{\bar{i}})]}},$$

$$\text{DCG}(f) = \sum_i \frac{1}{\ln\left(1+\text{Rank}(i)\right)} = \sum_i \frac{1}{\ln\left(1+\sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]} + \sum_{\bar{i}} \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\mathbf{x}_{\bar{i}})]}\right)}.$$

Let us consider the practical implications of minimizing the negation of the DCG. The discounting function $1/\ln(1+z)$ is decreasing, but its negation is not convex so there is no optimization guarantee. This is true even if we incorporate the exponential loss since $-1/\ln(1+e^z)$ is not convex. The same observation holds for the AveR.

It is possible, however, to choose a different discounting factor that allows us to create a convex objective to minimize. Let us choose a discounting factor of $-\ln(1+z)$, which is similar to the discounting factors for the AveR and DCG in that it is decreasing and convex. Figure 3 illustrates these discounting factors. Using this new discounting factor, and using the reverse height rather than the rank (which is an arbitrary choice), we arrive at the following objective:

$$R_{g_{\text{IR}},\mathbf{1}}(f) := \sum_i \ln\left(1 + \sum_k \mathbf{1}_{[f(\mathbf{x}_i) \leq f(\tilde{\mathbf{x}}_k)]}\right),$$

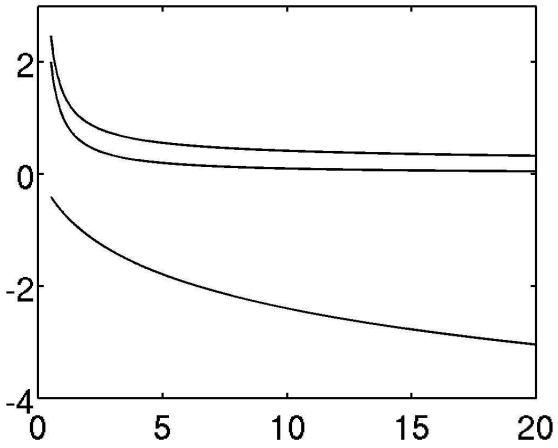

Figure 3: Discounting factor for discounted cumulative gain $1/\ln(1+z)$ (upper curve), discounting factor for the average of the reciprocal of the ranks $1/z$ (middle curve), and new discounting factor $-\ln(1+z)$ (lower curve) versus z.

and bounding the 0-1 loss from above,

$$R_{g_{\mathrm{IR}},\exp}(f) := \sum_i \ln\left(1+\sum_k e^{-(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k))}\right). \qquad \text{“IR Push”} \tag{2}$$

Equation (2) is our version of IR-ranking measures, which we refer to by “IR Push” in Section 8. It is also very similar in essence to the objective for the multilabel problem defined by Dekel et al. (2004). The objective (2) is globally convex. In general, one must be careful when defining discounting factors in order to avoid non-convexity. Figure 4 illustrates the contribution of each positive-negative pair to $R_{g_{\mathrm{IR}},\exp}(f)$ for the set of labels and examples defined in Section 3.3. The slant towards the lower left indicates that this objective is biased towards the top of the list.

Concentrating on the Bottom: Since our objective concentrates at the top of the ranked list, it can just as easily be made to concentrate on the bottom of the ranked list by reversing the positive and negative examples, or equivalently, by using the reverse height with a discounting factor of $-z^p$. In this case, our p-norm objective becomes:

$$R^{\mathrm{Bottom}}_{p,\exp}(f) := \sum_{i=1}^{I}\left(\sum_{k=1}^{K} e^{-(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k))}\right)^p.$$

Here, positive examples that score very badly are heavily penalized. $R^{\mathrm{Bottom}}_{p,\exp}(f)$ is also convex, so it can be easily minimized. Also, one can now write an objective that concentrates on the top and bottom simultaneously such as $R_{p,\exp}(f) + \text{const}\, R^{\mathrm{Bottom}}_{p,\exp}(f)$.

Crucial Pairs Formulation: The bipartite ranking problem is a specific case of the pairwise ranking problem. For the more general problem, the labels are replaced by a “truth function” $\pi : \mathcal{X}\times\mathcal{X} \rightarrow$

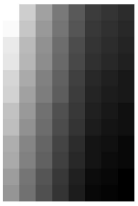

Figure 4: Contribution of each positive-negative pair to the objective $R_{g_{IR},\exp}$. Each square represents an i,k pair, where i is an index along the vertical axis, and k is along the horizontal axis, as described in Section 3.3. Lighter colors indicate larger contribution. The value of the i,k^{th} entry is the contribution of the i^{th} positive example, $\ln\left(1+\sum_k e^{-(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k))}\right)$, multiplied by the proportion of the loss attributed to the k^{th} negative example, $e^{-(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k))}/\sum_{\bar{k}} e^{-(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_{\bar{k}}))}$.

$\{0,1\}$, indicating whether the first element of the pair should be ranked above the second. In this case, one can replace the objective by:

$$R_{g,\ell}^{\text{Crucial Pairs}}(f) := \sum_{k=1}^{m} g\left(\sum_{i=1}^{m} \ell\Big(f(\mathbf{x}_i)-f(\mathbf{x}_k)\Big)\pi(\mathbf{x}_i,\mathbf{x}_k)\right),$$

where the indices i and k now run over all training examples. A slightly more general version of the above formula for $g(z)=z^p$ and the exponential loss was used by Ji et al. (2006) for the natural language processing problem of named entity recognition in Chinese. This algorithm performed quite well, in fact, within the margin of error of the best algorithm, but with a much faster training time. Its performance was substantially better than the support vector machine algorithm tested for this experiment. In Ji et al. (2006)'s setup, the P-Norm Push was used twice; the first time, a low value of p was chosen and a cutoff was made. The algorithm was used again for re-ranking (after some additional processing) with a higher value of p.

8. Experiments

The experiments of Ji et al. (2006) indicate the usefulness of our approach for larger, real-world problems. In this section, we will discuss the performance of the P-Norm Push on some smaller problems, since smaller problems are challenging when it comes to generalization. The choices we have made in Section 5 allow us to compare with RankBoost, which also uses the exponential loss. Furthermore, the choice of g as an adjustable power law allows us to illustrate the effect of the price g on the quality of the solution. Experiments have been performed using the P-Norm Push for $p=1$ (RankBoost), $2,4,8,16$ and 64, and using the IR Push information retrieval objective (2). For the P-Norm Push, the linesearch for α_t was performed using matlab's "fminunc" subroutine. The total number of iterations, $t_{\max}$, was fixed at 100 for all experiments. For the information retrieval objective, "fminunc" was used for the full optimization, which can be done for small experiments. Data were obtained from the UCI machine learning repository (Asuncion and Newman, 2007) and

all features were normalized to $[0,1]$. The three data sets chosen were MAGIC, ionosphere, and housing.

The first experiment uses the UCI MAGIC data set, which contains data from the Major Atmospheric Gamma Imaging Cherenkov Telescope project. The goal is to discriminate the statistical signatures of Monte Carlo simulated "gamma" particles from simulated "hadron" particles. In this problem, there are several relevant points on the ROC curve that determine the quality of the result. These points correspond to different acceptable false positive rates for different experiments, and all are close to the top of the list. There are 19020 examples (12332 gamma and 6688 hadron) and 11 features. Positive examples represent gamma particles and negative examples represent hadron particles. As a sample run, we chose 1000 examples randomly for training and tested on the rest.

Table 3 shows how different algorithms (the columns) performed with respect to different quality measures (the rows) on the MAGIC data. Each column of Table 3 represents a P-Norm Push or IR Push trial. The quality of the results is measured using the AUC (top row, larger values are better), $R_{p,\mathbf{1}}$ for various p (middle rows, smaller values are better), and the DCG and AveR (bottom rows, larger values are better). The best algorithms for each measure are summarized in bold and in the rightmost column. ROC curves and zoomed-in versions of the ROC curves for this sample run are shown in Figure 5. We expect the P-Norm Push for small p to yield the best results for

MAGIC data set

measure	p=1	p=2	p=4	p=8	p=16	p=64	IR	best
AUC	0.8370	**0.8402**	**0.8397**	0.8363	0.8329	0.8288	0.8284	small p
$R_{2,\mathbf{1}}$	6.5515	5.9731	5.5896	**5.4806**	**5.4990**	5.5819	5.5886	medium p
$R_{4,\mathbf{1}}$	4.2134	3.4875	2.8875	2.5638	2.4291	**2.3651**	**2.3582**	IR / large p
$R_{8,\mathbf{1}}$	3.8830	2.9138	2.1091	1.6266	1.3923	**1.2396**	**1.2257**	IR / large p
$R_{16,\mathbf{1}}$	6.8153	4.7208	3.0545	1.9698	1.4494	**1.1096**	**1.0823**	IR / large p
DCG	1.4022	1.4048	1.4066	1.4084	**1.4087**	**1.4087**	**1.4087**	IR / large p
AveR	8.1039	8.5172	8.6860	9.6701	**9.7520**	**9.7679**	**9.7688**	IR / large p

Table 3: Test performance of minimizers of $R_{p,\text{exp}}$ and $R_{g_{IR},\text{exp}}$ on a sample run with the MAGIC data set. Only significant digits are kept (factors of 10 have been removed). The best scores in each row are in bold and the right column summarizes the result by listing which algorithms performed the best with respect to each quality measure.

optimizing AUC, and we expect the large p and IR columns to yield the best results for $R_{p,\mathbf{1}}$ when p is large, and for the DCG and AveR. In other words, the rightmost column ought to say "small p" towards the top, followed by "medium p," and then "IR / large p." This general trend is observed. In this particular trial run, the IR Push and P-Norm Push for $p = 64$ yielded almost identical results, and their ROC curves are almost on top of each other in Figure 5.

The next experiment uses a much smaller data set, namely the UCI ionosphere data set, which has 351 examples (225 positive and 126 negative). These are data collected from a phased array of antennas. The goal is to distinguish "good" radar returns from "bad" radar returns. The good returns represent signals that reflect back towards the antenna, indicating structure in the ionosphere. The features are based on characteristics of the received signal. Out of the 34 features, we choose 5 of them (the last 5 features), which helps to alleviate overfitting, though there is still significant variation in results due to the small size of the data set. We used 3-fold cross-validation, where all

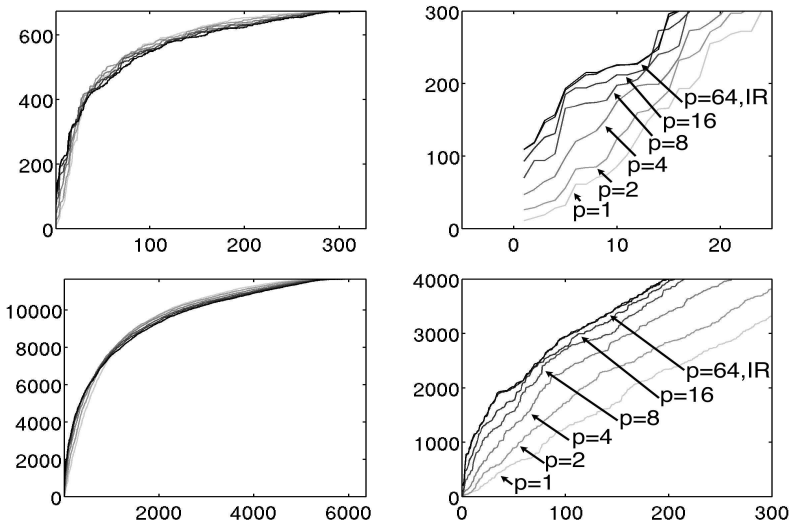

Figure 5: ROC Curves for the P-Norm Push and IR Push on the MAGIC data set. All plots are number of true positives vs. number of false positives. *Upper Left:* Full ROC Curves for training. *Upper Right:* Zoomed-in version of training ROC Curves. *Lower Left:* Full ROC Curves for testing. *Lower Right:* Zoomed-in version of testing ROC Curves.

algorithms were run once on each split, and the mean performance is reported in Table 4. ROC curves from one of the trials is presented in Figure 6. The trend from small to large p is able to be observed, despite variation due to train/test splits.

ionosphere data set

measure	p=1	p=2	p=4	p=8	p=16	p=64	IR	best
AUC	**0.6797**	0.6732	0.6700	0.6612	0.6479	0.6341	0.6409	small p
$R_{2,\mathbf{1}}$	2.1945	2.1931	2.1515	**2.1213**	2.1575	2.1974	2.1811	med/lg p
$R_{4,\mathbf{1}}$	2.0841	1.9891	1.8041	1.5911	1.4327	**1.3104**	1.4046	IR / large p
$R_{8,\mathbf{1}}$	3.7099	3.3459	2.6271	1.8950	1.3861	**1.0823**	1.2979	IR / large p
$R_{16,\mathbf{1}}$	1.7294	1.4558	0.8786	0.4236	0.2437	**0.1884**	0.2272	IR / large p
DCG	13.9197	14.1308	14.3261	14.5902	14.6916	**14.7903**	14.7169	IR / large p
AveR	2.9712	3.1610	3.3041	3.5084	3.5849	**3.6571**	3.6076	IR/ large p

Table 4: Mean test performance of minimizers of $R_{p,\text{exp}}$ and $R_{g_{IR},\text{exp}}$ over 3-fold cross-validation on the ionosphere data set.

We last consider the Boston Housing data set, which has 506 examples (35 positive, 471 negative), 13 features. This data set is skewed; there are significantly fewer positive examples than negative examples. In order to use the housing data set for a bipartite ranking problem, we used the fourth feature (which is binary) as the label y. The fourth feature describes whether a tract bounds the Charles River. Since there is some correlation between this feature and the other features

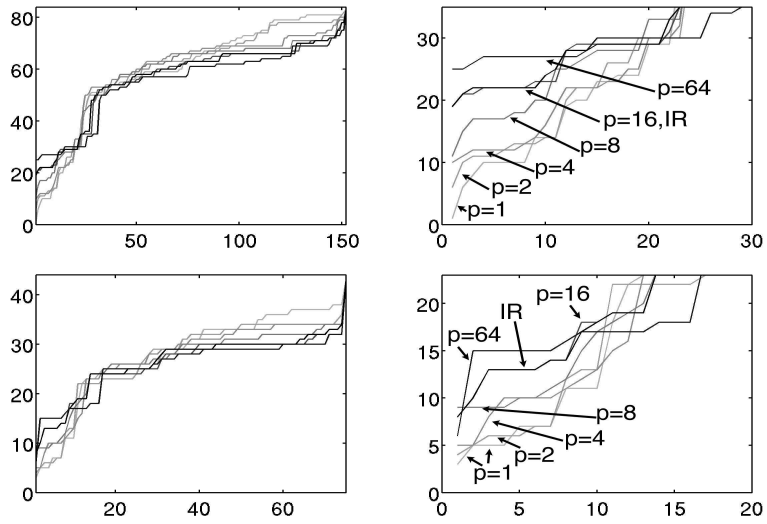

Figure 6: ROC Curves for the P-Norm Push and IR Push on ionosphere data set. All plots are number of true positives vs. number of false positives. *Upper Left:* Full ROC Curves for training. *Upper Right:* Zoomed-in version of training ROC Curves. *Lower Left:* Full ROC Curves for testing. *Lower Right:* Zoomed-in version of testing ROC Curves.

(such as distance to employment centers and tax-rate), it is reasonable for our learning algorithm to predict whether the tract bounds the Charles River based on the other features. We used 3-fold cross-validation (≈ 12 positives in each test set), where all algorithms were run once on each split, and the mean performance is reported in Table 5. ROC curves from one of the trials is presented in Figure 7. The trend from small to large p is again generally observed, despite variation due to data set size.

For all of these experiments, in agreement with our algorithm's derivation, a larger push (p large) causes the algorithm to perform better near the top of the ranked list on the training set. As discussed, this ability to correct the top of the list is not without sacrifice; we do sacrifice the ranks of items farther down on the list and we do reduce the value of the AUC, but we have made this choice on purpose in order to perform better near the top of the list.

9. Discussion and Open Problems

Here we describe interesting directions for future work.

9.1 Producing Dramatic Changes in the ROC curve

An open question is to quantify what properties of a hypothesis space and data set would allow an increase in p to cause a dramatic change in the ROC curve. In Section 8, we have shown cases where the benefits of increasing p are substantial, and in Section 3.2 we have shown that a dramatic

housing data set

measure	p=1	p=2	p=4	p=8	p=16	p=64	IR	best
AUC	**0.7739**	0.7633	0.7532	0.7500	0.7420	0.7330	0.7373	small p
$R_{2,\mathbf{1}}$	**3222**	3406	3665	3799	3818	3759	3847	small p
$R_{4,\mathbf{1}}$	**294078**	**292870**	304457	307135	305498	298611	304915	small/med p
$R_{8,\mathbf{1}}$	3.9056	3.5246	3.3908	3.2953	**3.2479**	3.3173	**3.2346**	IR / large p
$R_{16,\mathbf{1}}$	1.1762	0.9694	0.8337	0.8028	**0.7801**	0.8816	**0.7788**	IR / large p
DCG	3.6095	3.6476	3.6757	3.6858	**3.6977**	3.6671	**3.6931**	IR / large p
AveR	0.5241	0.5644	0.6022	0.6124	**0.6258**	0.6012	**0.6250**	IR / large p

Table 5: Mean test performance of minimizers of $R_{p,\text{exp}}$ and $R_{g_{IR},\text{exp}}$ over 3-fold cross validation with the housing data set. Only significant digits are kept (factors of 10 have been removed). The best scores in each row are in bold and the right column summarizes the result by listing which algorithms performed the best with respect to each quality measure.

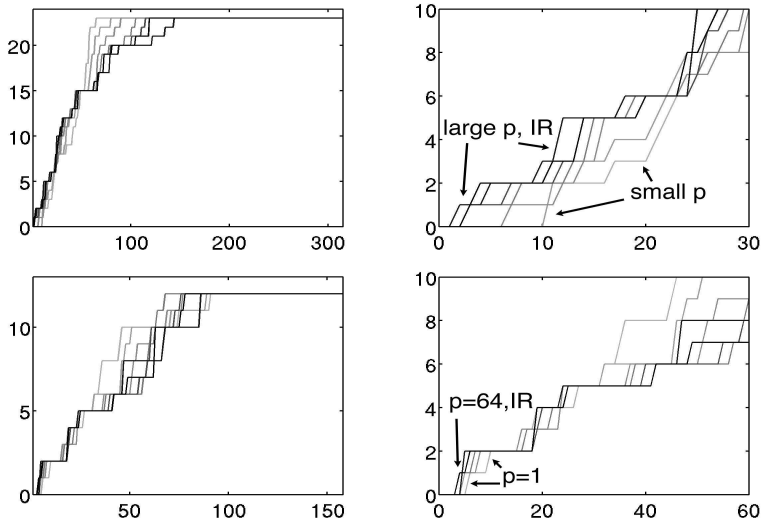

Figure 7: ROC Curves for the P-Norm Push and IR Push on the housing data set. All plots are number of true positives vs. number of false positives. *Upper Left:* Full ROC Curves for training. *Upper Right:* Zoomed-in version of training ROC Curves. *Lower Left:* Full ROC Curves for testing. *Lower Right:* Zoomed-in version of testing ROC Curves.

change is possible, even using an extremely small hypothesis space. However, it is sometimes the case that changes in p do not greatly affect the ROC curve.

A factor involved in this open question involves the flexibility of the hypothesis space with respect to the training set. Given a low capacity hypothesis space in which there is not too much flexibility in the set of solutions that yield good rankings, increasing p will not have much of an

effect. On the other hand, if a learning machine is high capacity, it probably has the flexibility to change the shape of the ROC curve dramatically. However, a high capacity learning machine generally is able to produce a consistent (or nearly consistent) ranking, and again, the choice of p probably does not have much effect. With respect to optimization on the training set, we have found the effect of increasing p to be the most dramatic when the hypothesis space is: limited (so as not to produce an almost consistent ranking), not too limited (features themselves are better than random guesses) and flexible (for instance, allowing some hypotheses to negate in order to produce a better solution as in Section 3.2). If such hypotheses are not available, we believe it is unlikely that any algorithm, whether the P-Norm Push, or any optimization algorithm for information retrieval measures, would be able to achieve a dramatic change in the ROC curve.

9.2 Optimizing $R_{\max}$ Directly

Given that there is no generalization guarantee for the ∞-norm, that is, $R_{\max}$, is it useful to directly minimize $R_{\max}$? This is still a convex optimization problem, and variations of this are done in other contexts, for instance, in the context of label ranking by Shalev-Shwartz and Singer (2006) and Crammer and Singer (2001). One might consider, for instance, optimizing $R_{\max}$ and measuring success on the test set using $R_{p,\mathbf{1}}$ for $p < \infty$.

One answer is provided by the equivalence of norms in finite dimensions. For instance, the value of $R_{\max}$ scales with $R_{p,\mathbf{1}}$, as demonstrated in Theorem 1. So optimizing $R_{\max}$ would still possibly be useful with respect to measuring success on smaller p (though in this case, one could optimize $R_{p,\ell}$).

9.3 Choices for ℓ and g

An important direction for future research is the choice of loss function ℓ and price function g. This framework is flexible in that different choices for ℓ and g can be chosen based on the particular goal, whether it is to optimize the AUC, $R_{p,\mathbf{1}}$ for some p, one of the IR measures suggested, or something totally different. The objective for the IR measures needed a concave price function $\ln(1+z)$, in which case the objective convex was made convex by using the exponential loss, in other words., $\ln(1+e^x)$ is convex. It may be possible to leverage the loss function in other cases, allowing us to consider more varied price functions while still working with an objective that is convex. One appealing possibility is to choose a non-monotonic function for g, which might allow us to concentrate on a specific portion of the ROC Curve; however, it may be difficult to maintain the convexity of the objective through the choice of the loss function.

Now we move on to the proofs.

10. Proof of Theorem 2 and Theorem 3

We define a Lipschitz function $\phi : \mathcal{R} \to \mathcal{R}$ (with Lipschitz constant $\mathrm{Lip}(\phi)$) which will act as our loss function, and gives us the margin. We will eventually use the same piecewise linear definition of ϕ as Koltchinskii and Panchenko (2002), but for now, we require only that ϕ obey $0 \le \phi(z) \le 1 \; \forall z$ and $\phi(z) = 1$ for $z < 0$. Since $\phi(z) \ge \mathbf{1}_{[z \le 0]}$, we can define an upper bound for $R^{\text{true}}_{p,\mathbf{1}}(f)$:

$$R^{\text{true}}_{p,\phi}(f) := \left(\mathbb{E}_{\mathbf{x}_- \sim \mathcal{D}_-} \left(\mathbb{E}_{\mathbf{x}_+ \sim \mathcal{D}_+} \phi\big(f(\mathbf{x}_+) - f(\mathbf{x}_-)\big) \right)^p \right)^{1/p}.$$

We have $R^{\text{true}}_{p,\mathbf{1}}(f) \leq R^{\text{true}}_{p,\phi}(f)$. The empirical error associated with $R^{\text{true}}_{p,\phi}$ is:

$$R^{\text{empirical}}_{p,\phi}(f) := \left(\frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{I} \sum_{i=1}^{I} \phi\big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\big) \right)^p \right)^{1/p}.$$

First, we bound from above the quantity $R^{\text{true}}_{p,\phi}$ by two terms: the empirical error term $R^{\text{empirical}}_{p,\phi}$, and a term characterizing the deviation of $R^{\text{empirical}}_{p,\phi}$ from $R^{\text{true}}_{p,\phi}$ uniformly:

$$\begin{aligned} R^{\text{true}}_{p,\mathbf{1}}(f) &\leq R^{\text{true}}_{p,\phi}(f) = R^{\text{true}}_{p,\phi}(f) - R^{\text{empirical}}_{p,\phi}(f) + R^{\text{empirical}}_{p,\phi}(f) \\ &\leq \sup_{\bar{f} \in \mathcal{F}} \left(R^{\text{true}}_{p,\phi}(\bar{f}) - R^{\text{empirical}}_{p,\phi}(\bar{f}) \right) + R^{\text{empirical}}_{p,\phi}(f). \end{aligned}$$

The proof of Theorem 3 mainly involves an upper bound on the first term. The second term will be upper bounded by $R^{\text{empirical}}_{p,\mathbf{1},\theta}(f)$ by our choice of ϕ. Define $L(f)$ as follows:

$$L(f) := R^{\text{true}}_{p,\phi}(f) - R^{\text{empirical}}_{p,\phi}(f).$$

Let us outline the proof that follows. The goal is to bound $L(f)$ uniformly over $f \in \mathcal{F}$. To do this, we use a covering number argument similar to that of Cucker and Smale (2002). First, we will cover $\mathcal{F}$ by L_∞ disks. We show in Lemma 5 (below) that the value of $L(f)$ within each disk does not change very much provided that the disks are small. We then derive a probabilistic bound on $L(f)$ for any f in Lemma 9, and use this bound on representatives f_r from each disk. A union bound over disks yields the result. The most effort of this proof is devoted to the bound on $L(f)$ in Lemma 9 below, which uses McDiarmid's Inequality. Let us now proceed with the proof.

The following lemma is true for every training set S. It will be used later to show that the value of $L(f)$ does not change much within each L_∞ ball.

Lemma 5 *For any two functions* $f_1, f_2 \in L_\infty(\mathcal{X})$,

$$L(f_1) - L(f_2) \leq 4\text{Lip}(\phi)||f_1 - f_2||_\infty.$$

Proof First, we rearrange the terms:

$$\begin{aligned} L(f_1) - L(f_2) &= R^{\text{true}}_{p,\phi}(f_1) - R^{\text{empirical}}_{p,\phi}(f_1) - R^{\text{true}}_{p,\phi}(f_2) + R^{\text{empirical}}_{p,\phi}(f_2) \\ &= [R^{\text{true}}_{p,\phi}(f_1) - R^{\text{true}}_{p,\phi}(f_2)] - [R^{\text{empirical}}_{p,\phi}(f_1) - R^{\text{empirical}}_{p,\phi}(f_2)]. \end{aligned} \tag{3}$$

We bound from above the second bracketed term of (3),

$$\begin{aligned}
&R^{\text{empirical}}_{p,\phi}(f_1) - R^{\text{empirical}}_{p,\phi}(f_2) \\
&= \left[\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f_1(\mathbf{x}_i) - f_1(\tilde{\mathbf{x}}_k)\Big)\right]^p\right]^{1/p} - \left[\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f_2(\mathbf{x}_i) - f_2(\tilde{\mathbf{x}}_k)\Big)\right]^p\right]^{1/p} \\
&\leq \left[\frac{1}{K}\sum_{k=1}^{K}\left|\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f_1(\mathbf{x}_i) - f_1(\tilde{\mathbf{x}}_k)\Big) - \frac{1}{I}\sum_{i=1}^{I}\phi\Big(f_2(\mathbf{x}_i) - f_2(\tilde{\mathbf{x}}_k)\Big)\right|^p\right]^{1/p} \\
&\leq \left[\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{I}\sum_{i=1}^{I}\left|\phi\Big(f_1(\mathbf{x}_i) - f_1(\tilde{\mathbf{x}}_k)\Big) - \phi\Big(f_2(\mathbf{x}_i) - f_2(\tilde{\mathbf{x}}_k)\Big)\right|\right]^p\right]^{1/p} \\
&\leq \left[\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{I}\sum_{i=1}^{I}\text{Lip}(\phi)\left|f_1(\mathbf{x}_i) - f_1(\tilde{\mathbf{x}}_k) - f_2(\mathbf{x}_i) + f_2(\tilde{\mathbf{x}}_k)\right|\right]^p\right]^{1/p} \\
&\leq \left[\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{I}\sum_{i=1}^{I}\text{Lip}(\phi)2\sup_{\mathbf{x}}\left|f_1(\mathbf{x}) - f_2(\mathbf{x})\right|\right]^p\right]^{1/p} = 2\text{Lip}(\phi)\|f_1 - f_2\|_\infty.
\end{aligned}$$

Here, we have used Minkowski's inequality for $\ell_p(\mathcal{R}^K)$, which is the triangle inequality $\|f-g\|_p \geq \|f\|_p - \|g\|_p$, and the definition of the Lipschitz constant for ϕ. An identical calculation for the first bracketed term of (3), again using Minkowski's inequality yields:

$$R^{\text{true}}_{p,\phi}(f_1) - R^{\text{true}}_{p,\phi}(f_2) \leq 2\text{Lip}(\phi)||f_1 - f_2||_\infty.$$

Combining the two terms yields the statement of the lemma. ■

The following step appears in Cucker and Smale (2002). Let $\ell_\varepsilon := \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{8\text{Lip}(\phi)}\right)$, the covering number of $\mathcal{F}$ by L_∞ disks of radius $\frac{\varepsilon}{8\text{Lip}(\phi)}$. Define $f_1, f_2, ..., f_{\ell_\varepsilon}$ to be the centers of such a cover. In other words, the collection of L_∞ disks B_r centered at f_r and with radius $\frac{\varepsilon}{8\text{Lip}(\phi)}$ is a cover for $\mathcal{F}$. In the proof of the theorem, we will use the center of each disk to act as a representative for the whole disk. So, we must show that we do not lose too much by using f_r as a representative for disk B_r.

Lemma 6 *For all* $\varepsilon > 0$,

$$\mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left\{\sup_{f\in B_r} L(f) \geq \varepsilon\right\} \leq \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left\{L(f_r) \geq \frac{\varepsilon}{2}\right\}.$$

Proof By Lemma 5, for every training set S and for all $f \in B_r$,

$$\sup_{f\in B_r} L(f) - L(f_r) \leq 4\text{Lip}(\phi)\sup_{f\in B_r}||f - f_r||_\infty \leq 4\text{Lip}(\phi)\frac{\varepsilon}{8\text{Lip}(\phi)} = \frac{\varepsilon}{2}$$

Thus,

$$\sup_{f\in B_r} L(f) \geq \varepsilon \implies L(f_r) \geq \frac{\varepsilon}{2}.$$

The statement of the lemma follows directly. ■

Here is an inequality that will be useful in the next proof as the mechanism for incorporating p into the bound.

Lemma 7 *For $0 \leq a, b \leq 1$,*

$$|a^{1/p} - b^{1/p}| \leq \min\left\{ |a-b|a^{(1/p)-1}, |a-b|^{1/p} \right\}.$$

Proof For $p = 1$ there is nothing to prove, so take $p > 1$. We need to show both

$$|a^{1/p} - b^{1/p}| \leq |a-b|a^{(1/p)-1} \tag{4}$$

and

$$|a^{1/p} - b^{1/p}| \leq |a-b|^{1/p}. \tag{5}$$

Let us show (5) first. For $z_1, z_2 \geq 0$, it is true that $z_1^p + z_2^p \leq (z_1 + z_2)^p$ as an immediate consequence of the binomial theorem. When $a \geq b$, substitute $z_1 = (a-b)^{1/p}$, $z_2 = b^{1/p}$. The result follows after simplification. The case $a \leq b$ is completely symmetric so no additional work is needed. To show (4), consider first the case $a \geq b$, so that

$$b^{1/p-1} \geq a^{1/p-1}.$$

Multiplying by b yields $b^{1/p} \geq a^{1/p-1}b$, negating and adding $a^{1/p}$ yields

$$a^{1/p} - b^{1/p} \leq a^{1/p} - a^{1/p-1}b, \text{ so } a^{1/p} - b^{1/p} \leq (a-b)a^{1/p-1}.$$

Exactly the same steps (with reversed inequalities) can be used to show the $b \geq a$ case. ■

The benefit of using the minimum in Lemma 7 is that the first term most often gives a tighter bound. In the case where it does not do so, the second term applies. An illustration of this inequality is provided in Figure 8.

We now incorporate the fact that the training set is chosen randomly. We will use a generalization of Hoeffding's inequality due to McDiarmid, as follows:

Theorem 8 ***(McDiarmid, 1989)*** *Let $X_1, X_2, \ldots X_m$ be independent random variables under distribution D on $\mathcal{X}$. Let $f : \mathcal{X}^m \to \mathcal{R}$ be any function such that:*

$$\sup_{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m, \mathbf{x}_i'} \left| f(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_m) - f(\mathbf{x}_1, \ldots, \mathbf{x}_i', \ldots \mathbf{x}_m) \right| \leq c_i \quad \textit{for} \quad 1 \leq i \leq m.$$

Then for any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}_{X_1, X_2, \ldots, X_m \sim D}\left\{ f(X_1, X_2, \ldots, X_m) - \mathbb{E}\left[f(X_1, X_2, \ldots, X_m) \right] \geq \varepsilon \right\} &\leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2} \right), \\
\mathbb{P}_{X_1, X_2, \ldots, X_m \sim D}\left\{ \mathbb{E}\left[f(X_1, X_2, \ldots, X_m) \right] - f(X_1, X_2, \ldots, X_m) \geq \varepsilon \right\} &\leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2} \right),
\end{aligned}$$

and thus by the union bound,

$$\mathbb{P}_{X_1, X_2, \ldots, X_m \sim D}\left\{ \left| f(X_1, X_2, \ldots, X_m) - \mathbb{E}\left[f(X_1, X_2, \ldots, X_m) \right] \right| \geq \varepsilon \right\} \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^m c_i^2} \right).$$

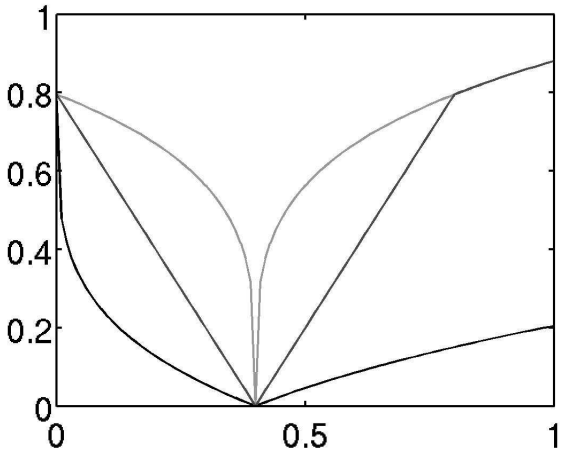

Figure 8: Functions $|a^{1/p} - b^{1/p}|$ (lower curve), $|a-b|^{1/p}$ (upper curve), and $\min(|a-b|a^{(1/p)-1}, |a-b|^{1/p})$ (middle curve) versus b. For this figure $p = 4$ and $a = 0.4$. One can see that in most cases, $|a-b|a^{(1/p)-1}$ is a better approximation to $|a^{1/p} - b^{1/p}|$ than $|a-b|^{1/p}$).

Here is our main probabilistic bound on $L(f)$ for an individual f. It uses McDiarmid's Inequality (Theorem 8) and Lemma 7.

Lemma 9 *For all* $\varepsilon_1 > 0$, *for all* $f \in \mathcal{F}$*:*

$$\begin{aligned}
&\mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(L(f) \geq \varepsilon_1\right) \\
&\quad\leq\quad 2\exp\left[-2K\max\left\{\frac{\varepsilon_1^2}{4}\left(R_{p,\phi}^{\text{true}}(f)\right)^{2(p-1)}, \left(\frac{\varepsilon_1}{2}\right)^{2p}\right\}\right] + 2\exp\left[-\frac{\varepsilon_1^2}{2}I + \ln K\right]. \qquad (6)
\end{aligned}$$

Proof Define

$$R_{p,\phi}^{\mathcal{DS}}(f) := \left(\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p\right)^{1/p}.$$

Now,

$$\begin{aligned}
&\mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\Big(L(f) \geq \varepsilon_1\Big) \\
&\quad=\quad \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\Big(R_{p,\phi}^{\text{true}}(f) - R_{p,\phi}^{\mathcal{DS}}(f) + R_{p,\phi}^{\mathcal{DS}}(f) - R_{p,\phi}^{\text{empirical}}(f) \geq \varepsilon_1\Big) \\
&\quad\leq\quad \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(R_{p,\phi}^{\text{true}}(f) - R_{p,\phi}^{\mathcal{DS}}(f) \geq \frac{\varepsilon_1}{2}\right) + \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(R_{p,\phi}^{\mathcal{DS}}(f) - R_{p,\phi}^{\text{empirical}}(f) \geq \frac{\varepsilon_1}{2}\right) \\
&\quad=:\quad \text{term}_1 + \text{term}_2. \qquad (7)
\end{aligned}$$

We bound term_1 and term_2 of (7) separately.

Bound on term$_1$*:* The following uses Lemma 7 above (translating into notation of the lemma):

$$\begin{aligned}
&R^{\text{true}}_{p,\phi}(f) - R^{\mathcal{DS}}_{p,\phi}(f) \\
&= \left(\mathbb{E}_{\mathbf{x}_-\sim\mathcal{D}_-}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\mathbf{x}_-)\Big)\right)^p\right)^{1/p} \\
&\quad - \left(\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p\right)^{1/p} \\
&= a^{1/p} - b^{1/p} \leq |a^{1/p} - b^{1/p}| \leq \min\left\{|a-b|a^{(1/p)-1}, |a-b|^{1/p}\right\}.
\end{aligned}$$

Thus for all $\varepsilon_1 > 0$,

$$\begin{aligned}
&\mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(R^{\text{true}}_{p,\phi}(f) - R^{\mathcal{DS}}_{p,\phi}(f) \geq \frac{\varepsilon_1}{2}\right) \\
&\leq \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(\min\left\{|a-b|a^{(1/p)-1}, |a-b|^{1/p}\right\} \geq \frac{\varepsilon_1}{2}\right) \\
&= \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(|a-b|a^{1/p-1} \geq \frac{\varepsilon_1}{2} \bigcap |a-b|^{\frac{1}{p}} \geq \frac{\varepsilon_1}{2}\right) \\
&= \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(|a-b| \geq \frac{\varepsilon_1}{2}a^{1-1/p} \bigcap |a-b| \geq \left(\frac{\varepsilon_1}{2}\right)^p\right) \\
&= \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(|a-b| \geq \frac{\varepsilon_1}{2}\left(R^{\text{true}}_{p,\phi}(f)\right)^{p-1} \bigcap |a-b| \geq \left(\frac{\varepsilon_1}{2}\right)^p\right) \\
&= \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(|a-b| \geq \max\left\{\frac{\varepsilon_1}{2}\left(R^{\text{true}}_{p,\phi}(f)\right)^{p-1}, \left(\frac{\varepsilon_1}{2}\right)^p\right\}\right).
\end{aligned}$$

Let

$$\varepsilon_2 := \max\left\{\frac{\varepsilon_1}{2}\left(R^{\text{true}}_{p,\phi}(f)\right)^{p-1}, \left(\frac{\varepsilon_1}{2}\right)^p\right\}.$$

Then,

$$\begin{aligned}
&\mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(R^{\text{true}}_{p,\phi}(f) - R^{\mathcal{DS}}_{p,\phi}(f) \geq \frac{\varepsilon_1}{2}\right) \\
&\leq \mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(\left|\mathbb{E}_{\mathbf{x}_-\sim\mathcal{D}_-}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\mathbf{x}_-)\Big)\right)^p - \right.\right. \\
&\qquad \left.\left.\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p\right| \geq \varepsilon_2\right).
\end{aligned}$$

The largest possible change in $\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p$ due to the replacement of one negative example is $1/K$. By McDiarmid's inequality,

$$\begin{aligned}
&\mathbb{P}_{S_-\sim\mathcal{D}_-^K}\left(R^{\text{true}}_{p,\phi}(f) - R^{\mathcal{DS}}_{p,\phi}(f) \geq \frac{\varepsilon_1}{2}\right) \\
&\leq \exp\left(-\frac{2\varepsilon_2^2}{K\frac{1}{K^2}}\right) = 2\exp(-2K\varepsilon_2^2) \\
&= 2\exp\left(-2K\max\left\{\frac{\varepsilon_1^2}{4}\left(R^{\text{true}}_{p,\phi}(f)\right)^{2(p-1)}, \left(\frac{\varepsilon_1}{2}\right)^{2p}\right\}\right). \qquad (8)
\end{aligned}$$

Bound on term$_2$:

$$\begin{aligned} & R^{\mathcal{DS}}_{p,\phi}(f) - R^{\text{empirical}}_{p,\phi}(f) \\ = \quad & \left(\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\mathbf{x}_-)\Big)\right)^p\right)^{1/p} \\ & - \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p\right)^{1/p}. \end{aligned}$$

Thus for all $\varepsilon_1 > 0$,

$$\begin{aligned} & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(R^{\mathcal{DS}}_{p,\phi}(f) - R^{\text{empirical}}_{p,\phi}(f) \geq \frac{\varepsilon_1}{2}\right) \\ = \quad & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(\frac{1}{K^{1/p}}\left\|\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\cdot)\Big)\right\|_{\ell_p(R^K)}\right. \\ & \qquad \left. - \frac{1}{K^{1/p}}\left\|\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\cdot)\Big)\right\|_{\ell_p(R^K)} \geq \frac{\varepsilon_1}{2}\right) \\ \leq \quad & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(\frac{1}{K^{1/p}}\left\|\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\cdot)\Big) - \frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\cdot)\Big)\right\|_{\ell_p(R^K)} \geq \frac{\varepsilon_1}{2}\right) \\ \leq \quad & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(\left\|\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\cdot)\Big) - \frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\cdot)\Big)\right\|_{\ell_\infty(R^K)} \geq \frac{\varepsilon_1}{2}\right) \\ = \quad & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(\exists k : \left|\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\tilde{\mathbf{x}}_k)\Big) - \frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right| \geq \frac{\varepsilon_1}{2}\right). \end{aligned}$$

We now use McDiarmid's Inequality. The largest possible change in $\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)$ due to the replacement of one positive example is $1/I$. Thus, for all $\tilde{\mathbf{x}}_k$,

$$\begin{aligned} & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(\left|\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\tilde{\mathbf{x}}_k)\Big) - \frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right| \geq \frac{\varepsilon_1}{2}\right) \\ \leq \quad & 2\exp\left[-\frac{2\left(\frac{\varepsilon_1}{2}\right)^2}{I\frac{1}{I^2}}\right] = 2\exp\left[-\frac{\varepsilon_1^2}{2}I\right]. \end{aligned}$$

By the union bound over the K negative examples:

$$\begin{aligned} & \mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(\exists k : \left|\mathbb{E}_{\mathbf{x}_+\sim\mathcal{D}_+}\phi\Big(f(\mathbf{x}_+) - f(\tilde{\mathbf{x}}_k)\Big) - \frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right| \geq \frac{\varepsilon_1}{2}\right) \\ \leq \quad & 2K\exp\left[-\frac{\varepsilon_1^2}{2}I\right] = 2\exp\left[-\frac{\varepsilon_1^2}{2}I + \ln(K)\right], \end{aligned}$$

and thus,

$$\mathbb{P}_{S_+\sim\mathcal{D}_+^I, S_-\sim\mathcal{D}_-^K}\left(R^{\text{true}}_{p,\phi}(f) - R^{\mathcal{DS}}_{p,\phi}(f) \geq \frac{\varepsilon_1}{2}\right) \leq 2\exp\left[-\frac{\varepsilon_1^2}{2}I + \ln K\right].$$

Combining this with (8) and (7) yields the statement of the lemma.

■

Proof *(Of Theorem 2 and Theorem 3)* Since the B_r are a cover for $\mathcal{F}$, it is true that

$$\sup_{f\in\mathcal{F}} L(f) \geq \varepsilon \iff \exists r \leq \ell_\varepsilon \text{ such that } \sup_{f\in B_r} L(f) \geq \varepsilon.$$

First applying the union bound over balls, then applying Lemma 6 we find:

$$\begin{aligned}
&\mathbb{P}_{S_+\sim\mathcal{D}^I, S_-\sim\mathcal{D}^K}\left\{\sup_{f\in\mathcal{F}} L(f) \geq \varepsilon\right\} \\
&\quad\leq \sum_{r=1}^{\ell_\varepsilon} \mathbb{P}_{S_+\sim\mathcal{D}^I, S_-\sim\mathcal{D}^K}\left\{\sup_{f\in B_r} L(f) \geq \varepsilon\right\} \\
&\quad\leq \sum_{r=1}^{\ell_\varepsilon} \mathbb{P}_{S_+\sim\mathcal{D}^I, S_-\sim\mathcal{D}^K}\left\{L(f_r) \geq \varepsilon/2\right\}.
\end{aligned}$$

We bound from above using (6) in order to prove Theorem 3 using $\varepsilon_1 = \varepsilon/2$, also $R^{\text{true}}_{p,\phi}(f_r) \geq R^{\text{true}}_{p,\mathbf{1}}(f_r)$ and additionally $R^{\text{true}}_{p,\mathbf{1}}(f_r) \geq \inf_{f\in\mathcal{F}} R^{\text{true}}_{p,\mathbf{1}}(f)$ for every f_r:

$$\begin{aligned}
&\mathbb{P}_{S_+\sim\mathcal{D}^I, S_-\sim\mathcal{D}^K}\left\{\sup_{f\in\mathcal{F}} L(f) \geq \varepsilon\right\} \\
&\quad\leq \sum_{r=1}^{\ell_\varepsilon} 2\exp\left[-2K\max\left\{\frac{\varepsilon^2}{16}\left(R^{\text{true}}_{p,\phi}(f_r)\right)^{2(p-1)}, \left(\frac{\varepsilon}{4}\right)^{2p}\right\}\right] + 2\exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right] \\
&\quad\leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{8\text{Lip}(\phi)}\right)\left[2\exp\left[-2K\max\left\{\frac{\varepsilon^2}{16}\left(\min_r R^{\text{true}}_{p,\phi}(f_r)\right)^{2(p-1)}, \left(\frac{\varepsilon}{4}\right)^{2p}\right\}\right]\right. \qquad (9) \\
&\qquad\qquad \left.+2\exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right]\right] \\
&\quad\leq \mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{8\text{Lip}(\phi)}\right)\left[2\exp\left[-2K\max\left\{\frac{\varepsilon^2}{16}\left(\inf_{\tilde{f}\in\mathcal{F}} R^{\text{true}}_{p,\mathbf{1}}(\tilde{f})\right)^{2(p-1)}, \left(\frac{\varepsilon}{4}\right)^{2p}\right\}\right]\right. \\
&\qquad\qquad \left.+2\exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right]\right].
\end{aligned}$$

Now we put everything together. The probability that there exists an $f \in \mathcal{F}$ where

$$R^{\text{true}}_{p,\phi}(f) \geq R^{\text{empirical}}_{p,\phi}(f) + \varepsilon$$

is at most

$$\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon}{8\text{Lip}(\phi)}\right)\left[2\exp\left[-2K\max\left\{\frac{\varepsilon^2}{16}\left(R_{p,\min}\right)^{2(p-1)}, \left(\frac{\varepsilon}{4}\right)^{2p}\right\}\right] + 2\exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right]\right],$$

where $R_{p,\min} = \inf_f R^{\text{true}}_{p,\mathbf{1}}(f)$. Let us choose $\phi(z) = 1$ for $z \leq 0$, $\phi(z) = 0$ for $z \geq \theta$, and linear in between, with slope $-1/\theta$. Thus, $\text{Lip}(\phi) = 1/\theta$. Since $\phi(z) \leq 1$ for $z \leq \theta$, we have:

$$\begin{aligned} R^{\text{empirical}}_{p,\phi}(f) &= \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I}\phi\Big(f(\mathbf{x}_i) - f(\tilde{\mathbf{x}}_k)\Big)\right)^p\right)^{1/p} \\ &\leq \left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{I}\sum_{i=1}^{I}\mathbf{1}_{[f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_k)\leq\theta]}\right)^p\right)^{1/p} = R^{\text{empirical}}_{p,\mathbf{1},\theta}(f). \end{aligned}$$

Thus, the probability that there exists an $f \in \mathcal{F}$ where

$$R^{\text{true}}_{p,\mathbf{1}}(f) \geq R^{\text{empirical}}_{p,\mathbf{1},\theta}(f) + \varepsilon$$

is at most

$$\mathcal{N}\left(\mathcal{F}, \frac{\varepsilon\theta}{8}\right)\left[2\exp\left[-2K\max\left\{\frac{\varepsilon^2}{16}\left(R_{p,\min}\right)^{2(p-1)}, \left(\frac{\varepsilon}{4}\right)^{2p}\right\}\right] + 2\exp\left[-\frac{\varepsilon^2}{8}I + \ln K\right]\right].$$

Thus, the theorem has been proved. A tighter bound is obtained if we bound differently at (9): instead of using $R^{\text{true}}_{p,\mathbf{1}}(f_r) \geq \inf_{f\in\mathcal{F}} R^{\text{true}}_{p,\mathbf{1}}(f)$, we could stop at $R^{\text{true}}_{p,\mathbf{1}}(f_r) \geq \min_r R^{\text{true}}_{p,\mathbf{1}}(f_r)$ and then choose the $\{f_r\}_r$ to maximize $\min_r R^{\text{true}}_{p,\mathbf{1}}(f_r)$.

Theorem 2 follows directly from the statement of Theorem 3. ■

10.1 1-Dimensional Illustration

As discussed earlier, since most of the value of $R^{\text{true}}_{p,\mathbf{1}}$ comes from a small portion of the domain, more examples are needed to compensate. Let us give a 1-dimensional illustration where this is the case. Almost half the distribution (proportion $\frac{1}{2} - \frac{\varepsilon}{2}$) consists of negative examples uniformly distributed on $[-1,0]$. Almost half the distribution (proportion $\frac{1}{2} - \frac{\varepsilon}{2}$) are positive examples uniformly distributed on $[0,1]$. An $\varepsilon/2$ proportion of the distribution are positive examples distributed on $[-2,-1]$, and the remaining $\varepsilon/2$ are negative examples on $[1,2]$. Drawing a training set of size m from that distribution, with probability $(1-\varepsilon)^m$, all examples will be drawn from $[-1,1]$, missing the essential part of the distribution. Let the hypothesis space $\mathcal{F}$ consist of one monotonically increasing function, and one monotonically decreasing function. Assuming the test set is large and represents the full distribution, the correct function to minimize $R_{\max}$ on the test set is the decreasing function. However, with high probability $(1-\varepsilon)^m$, the increasing function will be (wrongly) chosen, achieving on the training set, $R_{\max} = 0$, but on the test set, the worst possible value $R_{\max} = I$. Thus, $R_{\max}$ relies heavily on an ε-sized portion of the input space. Contrast this with behavior of the AUC, which is hardly affected by this portion of the input space, and is close to 1 with high probability for both training and testing.

11. Proof of Theorem 4

We will use a theorem of Della Pietra et al. (2002), and we will follow their definitions leading to this theorem. Consider a function $\phi : S \subset \mathcal{R}^{IK} \to [-\infty, \infty]$ (unrelated to the ϕ of the proof of Theorem 3). We will use this function to define a Bregman distance and consider an optimization

problem related to this Bregman distance. The dual of this optimization problem will be almost exactly the same as minimization of $R_{p,\exp}$ due to our choice of ϕ. The theorem of Della Pietra et al. (2002) will then provide a kind of uniqueness of the minimizer. The most difficult part of this theorem is finding the function ϕ and showing that the conditions of the framework are satisfied.

Let us first give the definition of a Bregman distance with respect to function ϕ, and then define the primal and dual optimization problems. The *effective domain* of ϕ, denoted Δ_ϕ, is the set of points where ϕ is finite. The function ϕ is *proper* if there is no $\mathbf{p}$ such that $\phi(\mathbf{p}) = -\infty$ and at least some $\mathbf{p}$ with $\phi(\mathbf{p}) \neq \infty$. (Do not confuse the vector $\mathbf{p} \in \mathcal{R}^{IK}$ with the scalar power p. Entries of $\mathbf{p}$ will always be indexed by p_{ik} to avoid confusion.) A proper function ϕ is *essentially smooth* if it is differentiable on the interior of the domain $\text{int}(\Delta_\phi)$ and if $\lim_\ell |\nabla\phi(\mathbf{p}_\ell)| = +\infty$ (element-wise) whenever $\mathbf{p}_\ell$ is a sequence in $\text{int}(\Delta_\phi)$, converging to a point on the boundary. Assume that the function ϕ is *Legendre*, meaning that it is closed (lower semi-continuous), convex and proper, and additionally that $\text{int}(\Delta_\phi)$ is convex, and ϕ is essentially smooth and strictly convex on $\text{int}(\Delta_\phi)$. The *Bregman Distance* associated with ϕ is $B_\phi : \Delta_\phi \times \text{int}(\Delta_\phi) \to [0,\infty]$ defined as:

$$B_\phi(\mathbf{p},\mathbf{q}) := \phi(\mathbf{p}) - \phi(\mathbf{q}) - \langle \nabla\phi(\mathbf{q}), \mathbf{p} - \mathbf{q}\rangle.$$

Fix a vector $\mathbf{p}_0 \in \Delta_\phi$. The *feasible set* for $\mathbf{p}_0$ with respect to matrix $\mathbf{M} \in \mathcal{R}^{IK\times n}$ is: $\mathcal{P} = \{\mathbf{p} \in \mathcal{R}^{IK} | \mathbf{p}^T\mathbf{M} = \mathbf{p}_0^T\mathbf{M}\}$. This will be the domain of the primal problem. The primal problem is to find, for fixed $\mathbf{q}_0 \in \Delta_\phi$:

$$\text{argmin}_{\mathbf{p}\in\mathcal{P}} B_\phi(\mathbf{p},\mathbf{q}_0). \qquad \textit{(primal problem)}$$

Now we lead up to the definition of the dual problem. The *Legendre-Bregman Conjugate* associated with ϕ is $\ell_\phi : \text{int}(\Delta_\phi) \times \mathcal{R}^{IK} \to \mathcal{R} \cup \{\infty\}$ defined as:

$$\ell_\phi(\mathbf{q},\mathbf{v}) := \sup_{\mathbf{p}\in\Delta_\phi} \Big(\langle \mathbf{v},\mathbf{p}\rangle - B_\phi(\mathbf{p},\mathbf{q}) \Big).$$

Note that for fixed $\mathbf{q}$, the Legendre-Bregman conjugate is exactly the convex conjugate of $B_\phi(\cdot,\mathbf{q})$. The *Legendre-Bregman Projection* is the argument of the sup whenever it is well-defined, namely, $\mathcal{L}_\phi : \text{int}(\Delta_\phi) \times \mathcal{R}^{IK} \to \Delta_\phi$ is defined by:

$$\mathcal{L}_\phi(\mathbf{q},\mathbf{v}) := \text{argmax}_{\mathbf{p}\in\Delta_\phi} \Big(\langle \mathbf{v},\mathbf{p}\rangle - B_\phi(\mathbf{p},\mathbf{q}) \Big),$$

whenever this is well-defined. Della Pietra et al. (2002) have shown that:

$$\mathcal{L}_\phi(\mathbf{q},\mathbf{v}) = (\nabla\phi)^{-1}(\nabla\phi(\mathbf{q}) + \mathbf{v}). \tag{10}$$

The dual problem can also be viewed as a minimization of a Bregman distance. Namely, it can be shown (cf. Proposition 2.7 of Della Pietra et al., 2002) that the dual objective can be written in terms of $\mathcal{L}_\phi(\mathbf{q}_0,\mathbf{v})$:

$$\langle \mathbf{v},\mathbf{p}_0\rangle - \ell_\phi(\mathbf{q}_0,\mathbf{v}) = B_\phi(\mathbf{p}_0,\mathbf{q}_0) - B_\phi\Big(\mathbf{p}_0, \mathcal{L}_\phi(\mathbf{q}_0,\mathbf{v})\Big).$$

Thus, since the first term on the right is not a function of $\mathbf{v}$, the dual problem can be written:

$$\begin{aligned} &\text{argmax}_{\mathbf{v}\in\mathcal{R}^{IK}} B_\phi(\mathbf{p}_0,\mathbf{q}_0) - B_\phi\Big(\mathbf{p}_0, \mathcal{L}_\phi(\mathbf{q}_0,\mathbf{v})\Big) \\ = \quad &\text{argmin}_{\mathbf{v}\in\mathcal{R}^{IK}} B_\phi\Big(\mathbf{p}_0, \mathcal{L}_\phi(\mathbf{q}_0,\mathbf{v})\Big), \qquad \textit{(dual problem)} \end{aligned}$$

where we have assumed in the domain of $\mathbf{v}$ that $\Delta_{\phi^*} = \mathcal{R}^{IK}$ and where ϕ^* is the convex conjugate of ϕ. We will rewrite the domain of the dual problem as the class $\mathcal{Q}$, defined as follows. For the $\mathbf{q}_0 \in \Delta_\phi$ and $\mathbf{M} \in \mathcal{R}^{IK \times n}$ fixed in the primal problem, the *Legendre-Bregman projection family for* $\mathbf{q}_0$ *and* $\mathbf{M}$ is defined by:

$$\mathcal{Q}(\mathbf{q}_0, \mathbf{M}) = \{\mathbf{q} \in \Delta_\phi | \mathbf{q} = \mathcal{L}_\phi(\mathbf{q}_0, -\mathbf{M}\boldsymbol{\lambda}) \text{ for some } \boldsymbol{\lambda} \in \mathcal{R}^n\}.$$

So instead of considering the minimizer with respect to $\mathbf{v}$, we will instead consider the minimizer with respect to $\mathbf{q} \in \mathcal{Q}$. In order to proceed, a few more technical conditions are required, namely:

- A1. ϕ is Legendre.
- A2. $\Delta_{\phi^*} = \mathcal{R}^{IK}$ where ϕ^* is the convex conjugate of ϕ.
- A3. B_ϕ extends to a function $B_\phi : \Delta_\phi \times \Delta_\phi \to [0, \infty]$ such that $B_\phi(\mathbf{p}, \mathbf{q})$ is continuous in $\mathbf{p}$ and $\mathbf{q}$, and satisfies $B_\phi(\mathbf{p}, \mathbf{q}) = 0$ iff $\mathbf{p} = \mathbf{q}$.
- A4. $\mathcal{L}_\phi$ extends to a function $\mathcal{L}_\phi : \Delta_\phi \times \mathcal{R}^{IK} \to \Delta_\phi$ satisfying $\mathcal{L}_\phi : (\mathbf{q}, \mathbf{0}) = \mathbf{q}$, such that $\mathcal{L}_\phi(\mathbf{q}, \mathbf{v})$ and $B_\phi(\mathcal{L}_\phi(\mathbf{q}, \mathbf{v}), \mathbf{q})$ are jointly continuous in $\mathbf{q}$ and $\mathbf{v}$.
- A5. $B_\phi(\mathbf{p}, \cdot)$ is *coercive* for every $\mathbf{p} \in \Delta_\phi \backslash \text{int}(\Delta_\phi)$, where a function $f : S \to [-\infty, \infty]$ is coercive if the level sets $\{\mathbf{q} \in S | f(\mathbf{q}) \leq c\}$ are bounded for every $c \in \mathcal{R}$.

We now state Proposition 3.2 of Della Pietra et al. (2002) which will give us uniqueness within the closure of the set $\mathcal{Q}$. Define $\bar{\mathcal{Q}}$ as the closure of $\mathcal{Q}$ in $\mathcal{R}^{IK}$.

Theorem 10 *(Della Pietra et al., 2002) Let ϕ satisfy A1.-A5. and suppose that $\mathbf{p}_0, \mathbf{q}_0 \in \Delta_\phi$ with $B_\phi(\mathbf{p}_0, \mathbf{q}_0) < \infty$. Then there exists a unique $\mathbf{q}^* \in \Delta_\phi$ satisfying the following four properties:*

1. $\mathbf{q}^* \in \mathcal{P} \bigcap \bar{\mathcal{Q}}$
2. $B_\phi(\mathbf{p}, \mathbf{q}) = B_\phi(\mathbf{p}, \mathbf{q}^*) + B_\phi(\mathbf{q}^*, \mathbf{q})$ *for any* $\mathbf{p} \in \mathcal{P}$ *and* $\mathbf{q} \in \bar{\mathcal{Q}}$
3. $\mathbf{q}^* = \text{argmin}_{\mathbf{p} \in \mathcal{P}} B_\phi(\mathbf{p}, \mathbf{q}_0)$ *(primal problem)*
4. $\mathbf{q}^* = \text{argmin}_{\mathbf{q} \in \bar{\mathcal{Q}}} B_\phi(\mathbf{p}_0, \mathbf{q})$ *(dual problem)*

Moreover, any one of these four properties determines $\mathbf{q}^*$ uniquely.

If we can prove that our objective function fits into this framework, we can use part (4) of this theorem to provide uniqueness in the closure of the set $\mathcal{Q}$, which will be related to our set $\mathcal{Q}'$. Let us now do exactly this.

Consider the following function $\phi : \mathcal{R}^{IK}_{\gtrdot 0} \to [-\infty, \infty]$:

$$\phi(\mathbf{q}) := \sum_{ik} q_{ik} \gamma(q_{ik}, \mathbf{q}), \text{ where } \gamma(q_{ik}, \mathbf{q}) := \ln \left(\frac{q_{ik}}{p^{1/p} (\sum_{i'} q_{i'k})^{(p-1)/p}} \right).$$

We extend the definition to $\mathcal{R}^{IK}_+$ by the conventions $0 \ln 0 = 0$ and $q_{ik}\gamma(q_{ik}, \mathbf{q}) = 0$ whenever $q_{ik} = 0$ for all i. Thus, Δ_ϕ is now $\mathcal{R}^{IK}_+$. The boundary in our case is where q_{ik} equals 0 for one or more ik pairs. We must now show that ϕ is Legendre.

Lemma 11 *ϕ is strictly convex in* $\mathrm{int}(\Delta_\phi)$*, where Δ_ϕ are vectors in $\mathcal{R}_+^{IK}$ with strictly positive entries.*

The proof is in the Appendix.

Lemma 12 *ϕ is Legendre.*

Proof ϕ is proper since there is no $\mathbf{q}$ such that $\phi(\mathbf{q}) = -\infty$. In order to achieve this, the term inside the logarithm must be exactly zero. When that happens, $q_{ik} = 0$, and by our convention, $q_{ik}\gamma(q_{ik}, \mathbf{q}) = 0$, thus the entire *ik* term is zero rather than $-\infty$. It can be verified that ϕ is lower semi-continuous. Also, $\mathrm{int}(\Delta_\phi) = \mathcal{R}_{\gg 0}^{IK}$ which is convex. We have already shown that ϕ is strictly convex on $\mathrm{int}(\Delta_\phi)$ in Lemma 11, and by our definition of ϕ on the boundary, it is convex on Δ_ϕ. We now show that ϕ is essentially smooth with respect to the boundary. Consider the following calculation for the gradient of ϕ in $\mathrm{int}(\Delta_\phi)$:

$$(\nabla\phi(\mathbf{q}))_{ik} = \frac{\partial\phi(\mathbf{q})}{\partial q_{ik}} = \frac{1}{p} + \ln\left(\frac{q_{ik}}{p^{1/p}\left(\sum_{i'} q_{i'k}\right)^{(p-1)/p}}\right) = \frac{1}{p} + \gamma(q_{ik}, \mathbf{q}), \tag{11}$$

since $\gamma(q_{ik}, \mathbf{q}) \to -\infty$ as $q_{ik} \to 0$, ϕ is essentially smooth. All the conditions have now been checked. ■

Also, we require the following:

Lemma 13 *Conditions A1.-A5. are obeyed.*

The proof of this lemma is in the Appendix.

Proof *(Of Theorem 4)* Let us compute the quantities above for our function ϕ, namely we would like to find the space Q and the dual objective $B_\phi(\mathbf{p}_0, \mathbf{q})$. Using (11) it can be shown that:

$$\left((\nabla\phi)^{-1}(\mathbf{z})\right)_{ik} = pe^{(z_{ik}-1/p)}\left(\sum_{i'} e^{(z_{i'k}-1/p)}\right)^{p-1}.$$

We now wish to compute $\mathcal{L}_\phi$. First, let us compute a term that appears often:

$e^{z_{ik}-1/p}$ where $z_{ik} = (\nabla\phi(\mathbf{q}) + \mathbf{v})_{ik} = \dfrac{1}{p} + \gamma(q_{ik}, \mathbf{q}) + v_{ik}$ can be rewritten:

$$e^{z_{ik}-1/p} \quad = \quad \exp\left[\frac{1}{p} - \frac{1}{p} + \gamma(q_{ik}, \mathbf{q}) + v_{ik}\right] = e^{v_{ik}}e^{\gamma(q_{ik}, \mathbf{q})}.$$

Thus from (10),

$$\begin{aligned}
\mathcal{L}_\phi(\mathbf{q}, \mathbf{v})_{ik} &= pe^{v_{ik}}e^{\gamma(q_{ik}, \mathbf{q})}\left(\sum_{i'} e^{v_{i'k}}e^{\gamma(q_{i'k}, \mathbf{q})}\right)^{p-1} \\
&= pe^{v_{ik}}\frac{q_{ik}}{p^{1/p}(\sum_{i'} q_{i'k})^{(p-1)/p}}\left(\frac{\sum_{i'} e^{v_{i'k}}q_{i'k}}{p^{1/p}(\sum_{i'} q_{i'k})^{(p-1)/p}}\right)^{p-1} \\
&= e^{v_{ik}}q_{ik}\left(\sum_{i'} e^{v_{i'k}}q_{i'k}\right)^{(p-1)}\frac{1}{(\sum_{i'} q_{i'k})^{(p-1)}}.
\end{aligned} \tag{12}$$

In our case, we choose $\mathbf{q}_0$ to be constant, $q_{0ik} = q_0$ for all i,k. We can now obtain Q:

$$Q(\mathbf{q}_0, \mathbf{M}) = \left\{ \mathbf{q} \middle| \mathbf{q} = e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}} \left(\sum_{i'} e^{-(\mathbf{M}\boldsymbol{\lambda})_{i'k}} \right)^{(p-1)} \frac{q_0}{I^{(p-1)}} \text{ for some } \boldsymbol{\lambda} \in \mathcal{R}^n \right\}.$$

In order to make the last fraction become 1, we choose $q_0 = I^{(p-1)}$. We now need only to define $\mathbf{p}_0$ in order to define the dual problem. In our case, we choose $\mathbf{p}_0 = \mathbf{0}$ so that the dual objective function is $B_\phi(\mathbf{0}, \mathbf{q})$. Let us choose $\mathbf{q} \in Q$, that is, $\mathbf{q}_{ik} = e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}} \left(\sum_{i'} e^{-(\mathbf{M}\boldsymbol{\lambda})_{i'k}} \right)^{(p-1)}$ and substitute using (11) and the definitions of ϕ and B_ϕ:

$$\begin{aligned}
B_\phi(\mathbf{0}, \mathbf{q}) &= \phi(\mathbf{0}) - \phi(\mathbf{q}) - \langle \nabla\phi(\mathbf{q}), \mathbf{0} - \mathbf{q} \rangle \\
&= -\phi(\mathbf{q}) + \mathbf{q} \cdot \nabla\phi(\mathbf{q}) \\
&= -\phi(\mathbf{q}) + \frac{1}{p} \sum_{ik} q_{ik} + \phi(\mathbf{q}) \\
&= \frac{1}{p} \sum_k \left(\sum_i e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}} \right) \left(\sum_{i'} e^{-(\mathbf{M}\boldsymbol{\lambda})_{i'k}} \right)^{(p-1)} \\
&= \frac{1}{p} \sum_k \left(\sum_i e^{-(\mathbf{M}\boldsymbol{\lambda})_{ik}} \right)^p = \frac{1}{p} R_{p,\exp}(\boldsymbol{\lambda}).
\end{aligned}$$

Thus, we have arrived at exactly the objective function for our algorithm. In other words, the function ϕ was carefully chosen so that the dual objective would be exactly as we wished, modulo the constant factor $1/p$ which does not affect minimization.

Part 4 of Theorem 10 tells us that the objective function of our algorithm has a unique minimizer in $\bar{Q}$ as long as A1.-A5. are obeyed, which holds from Lemma 13. It remains only to show that a vector in $\bar{Q}$ yields a unique vector in $\bar{Q}'$. Consider a sequence of vectors in Q defined element-wise by $q_{\ell,ik} = e^{-(\mathbf{M}\boldsymbol{\lambda})_{\ell,ik}} \left(\sum_{i'} e^{-(\mathbf{M}\boldsymbol{\lambda})_{\ell,i'k}} \right)^{p-1}$ such that $q_\ell \to \bar{q}$ as $\ell \to \infty$. Then consider the sequence defined by:

$$\frac{q_{\ell,ik}}{\left(\sum_{i'} q_{\ell,i'k}\right)^{(p-1)/p}} = e^{-(\mathbf{M}\boldsymbol{\lambda})_{\ell,ik}}.$$

By definition of Q', each vector in this sequence is in Q'. This sequence converges pointwise to $\frac{\bar{q}_{ik}}{\left(\sum_{i'} \bar{q}_{i'k}\right)^{(p-1)/p}} \in \bar{Q}'$, or if $\bar{q}_{ik} = 0$, then the ik^{th} component of the sequence converges to 0. Since we are in a finite dimensional space, namely $\mathcal{R}^{IK}$, pointwise convergence is sufficient. ■

It was unnecessary to state the primary objective $B_\phi(\mathbf{p}, \mathbf{q}_0)$ explicitly to prove the theorem, however, we state it in order to compare with the relative entropy case where $p = 1$. Recall that $\mathbf{q}_0$ is the constant vector with entries I^{p-1}. Thus, $(\nabla\phi(\mathbf{q}_0))_{ik} = \frac{1}{p} + \gamma(q_0, \mathbf{q}_0) = \frac{1}{p} + \ln(q_0 / [p^{1/p} (I q_0)^{(p-1)/p}]) = \frac{1}{p}(1 - \ln p)$ for all ik.

$$\begin{aligned}
B_\phi(\mathbf{p}, \mathbf{q}_0) &= \phi(\mathbf{p}) - \phi(\mathbf{q}_0) - \langle \nabla\phi(\mathbf{q}_0), \mathbf{p} - \mathbf{q}_0 \rangle \\
&= \phi(\mathbf{p}) - \langle \nabla\phi(\mathbf{q}_0), \mathbf{p} \rangle + \frac{1}{p} I^p K = \phi(\mathbf{p}) - (\nabla\phi(\mathbf{q}_0))_{ik} \sum_{ik} p_{ik} + \frac{1}{p} I^p K \\
&= \sum_{ik} p_{ik} \ln \left[\frac{p_{ik}}{p^{1/p} \left(\sum_{i'} p_{i'k}\right)^{(p-1)/p}} \right] - \frac{1}{p}(1 - \ln p) \sum_{ik} p_{ik} + \frac{1}{p} I^p K.
\end{aligned}$$

For $p = 1$ this reduces exactly to the relative entropy case.

One interesting note is how to find a function ϕ to suit such a problem; when we introduced it, we gave no indication of the techniques required to find such a function. In this case, we discovered the function ϕ again via convex duality. We knew the desired dual problem was precisely our objective $R_{p,\exp}$, thus, we were able to recover the primal problem by convex conjugation. The double dual in this case is the objective itself. From there, the function ϕ was obtained by analogy with the relative entropy case.

12. Conclusions

We have provided a method for constructing a ranked list where correctness at the top of the list is most important. Our main contribution is a general set of convex objective functions determined by a loss ℓ and price function g. A boosting-style algorithm based on a specific family of these objectives is derived. We have demonstrated the effect of a number of different price functions, and it is clear, both theoretically and empirically, that a steeper price function concentrates harder at the top of the list.

Acknowledgments

Thanks to the anonymous reviewers regarding Theorem 3, the experiments, and parts of the discussion, and especially for finding the numerous errors that come about from single authorship. Thanks to Rob Schapire for his advice, particularly with Lemma 11, thanks to Adrian Banner for careful proofreading, and thanks to Sinan Güntürk, particularly regarding Lemma 7. Also thanks to Eero Simoncelli and Dave Waltz. This work was supported by an NSF postdoctoral research fellowship under grant DBI-0434636 at New York University.

Appendix A.

We provide proofs of Lemma 11 and Lemma 13.

Proof (Of Lemma 11) First, rewrite ϕ:

$$\phi(\mathbf{q}) = \sum_k \left[\left(\sum_i q_{ik} \ln q_{ik} \right) - (\ln p^{1/p}) \left(\sum_i q_{ik} \right) - \frac{p-1}{p} \left(\sum_i q_{ik} \right) \ln \left(\sum_i q_{ik} \right) \right].$$

The middle term is linear so it does not affect convexity of the sum. It is sufficient to prove convexity of the following function, since ϕ would then be a sum (over k) of convex functions. Define $f : \mathcal{R}^I_{\geqslant 0} \to \mathcal{R}$ as follows, for $\mathbf{q} \in \mathcal{R}^I_+$:

$$f(\mathbf{q}) := \left(\sum_i q_i \ln q_i \right) + \frac{1-p}{p} \left(\sum_i q_i \right) \ln \left(\sum_i q_i \right).$$

Thus, the Hessian is:

$$\frac{\partial f(\mathbf{q})}{\partial q_\ell \partial q_i} = \frac{1}{q_i} \delta_{i=\ell} + \frac{1-p}{p} \frac{1}{\sum_{i'} q_{i'}}.$$

To show that the Hessian is positive definite, we show that $\mathbf{w}^{\mathbf{T}}\mathbf{H}\mathbf{w} > 0$ whenever $\mathbf{w} \neq \mathbf{0}$.

$$\begin{aligned}\sum_{i\ell} w_i w_\ell \frac{\partial f}{\partial q_\ell \partial q_i} &= \sum_i w_i^2 \frac{1}{q_i} + \frac{1-p}{p}\left(\sum_i w_i\right)^2 \frac{1}{\sum_i q_i} \\ &= \left(\frac{1}{\sum_i q_i}\right)\left[\left(\sum_i w_i^2 \frac{1}{q_i}\right)\left(\sum_i q_i\right) + \left(\frac{1}{p}-1\right)\left(\sum_i w_i\right)^2\right].\end{aligned}$$

Now, consider the Cauchy-Schwarz inequality, used in the following way:

$$\left(\sum_i w_i\right)^2 = \left\langle \frac{\mathbf{w}}{\sqrt{\mathbf{q}}}, \sqrt{\mathbf{q}}\right\rangle^2 \leq \left\|\frac{\mathbf{w}}{\sqrt{\mathbf{q}}}\right\|_2^2 \|\sqrt{\mathbf{q}}\|_2^2 = \left(\sum_i \frac{w_i^2}{q_i}\right)\left(\sum_i q_i\right).$$

Substituting back,

$$\begin{aligned}\sum_{i\ell} w_i w_\ell \frac{\partial f}{\partial q_\ell \partial q_i} &\geq \left(\frac{1}{\sum_i q_i}\right)\left[\left(\sum_i w_i^2 \frac{1}{q_i}\right)\left(\sum_i q_i\right) + \frac{1}{p}\left(\sum_i w_i\right)^2 - \left(\sum_i w_i^2 \frac{1}{q_i}\right)\left(\sum_i q_i\right)\right] \\ &= \left(\frac{1}{\sum_i q_i}\right)\frac{1}{p}\left(\sum_i w_i\right)^2.\end{aligned}$$

Recall that equality in Cauchy-Schwarz is only achieved when vectors are dependent, that is, for some $\alpha \in \mathcal{R}$, $w_i = \alpha q_i$ for all i. Since the elements of $\mathbf{q}$ are all strictly positive, if $w_i = \alpha q_i$, then at the same time we cannot have $\sum_i w_i = 0$. Thus, when equality holds in Cauchy-Schwarz, then $(\sum_i w_i)^2 > 0$ unless $\mathbf{w} = \mathbf{0}$. Thus, whether the Cauchy-Schwarz inequality is strict or not, we have:

$$\sum_{i\ell} w_i w_\ell \frac{\partial f}{\partial q_\ell \partial q_i} > 0 \text{ whenever } \mathbf{w} \neq \mathbf{0}.$$

Thus, ϕ is strictly convex. ■

Proof *(Of Lemma 13)* Condition A1. was proven in Lemma 12. To show A2., note that:

$$(\phi^*(\mathbf{v}))_{ik} = p e^{(v_{ik}-1/p)}\left(\sum_{i'} e^{(v_{i'k}-1/p)}\right)^{p-1}.$$

Thus, $\Delta_{\phi^*} = \mathcal{R}^{IK}$.

For A3., let us simplify using (11), where this calculation is valid for $\mathbf{p}, \mathbf{q} \in \Delta_\phi \times \text{int}(\Delta_\phi)$:

$$\begin{aligned}B_\phi(\mathbf{p},\mathbf{q}) &= \phi(\mathbf{p}) - \phi(\mathbf{q}) - \nabla\phi(\mathbf{q})\cdot(\mathbf{p}-\mathbf{q}) \\ &= \sum_{ik} p_{ik}\gamma(p_{ik},\mathbf{p}) - \phi(\mathbf{q}) - \frac{1}{p}\sum_{ki}(p_{ik}-q_{ik}) - \sum_{ik} p_{ik}\gamma(q_{ik},\mathbf{q}) + \phi(\mathbf{q}) \\ &= \sum_{ik} p_{ik}\Big(\gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q})\Big) - \frac{1}{p}\sum_{ik}(p_{ik}-q_{ik}). \qquad (13)\end{aligned}$$

Now we consider the boundary. If for some ik pair, $p_{ik} = 0$ then let $p_{ik}(\gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q})) = 0$ for all $\mathbf{q}$. If for some ik pair, $p_{ik} \neq 0$ and additionally $q_{ik} = 0$ we define $B_\phi(\mathbf{p},\mathbf{q}) = \infty$. This completes our definition of B_ϕ on the boundary of $\mathcal{R}_+^{IK}$. Let us prove that $B_\phi(\mathbf{p},\mathbf{q}) = 0$ implies $\mathbf{p} = \mathbf{q}$. Considering the interior, B_ϕ can only be 0 at a minimum since it is non-negative. A necessary condition for B_ϕ to be at a minimum is for $\partial B_\phi(\mathbf{p},\mathbf{q})/\partial p_{ik} = 0$ for all ik:

$$\forall ik \quad 0 = \frac{\partial B_\phi(\mathbf{p},\mathbf{q})}{\partial p_{ik}} = 1 - \frac{p-1}{p} + \gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q}) - \frac{1}{p} \Rightarrow \forall ik \quad \gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q}) = 0.$$

It is true that $\gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q}) = 0$ for pair ik implies that $p_{ik} = q_{ik}$. To see this, note that one can determine p_{ik} directly from the $\gamma(p_{ik},\mathbf{p})$ values as follows. Set $z_{ik} := p^{1/p}\exp(\gamma(p_{ik},\mathbf{p}))$. Now,

$$z_{ik}\left(\sum_{i'} z_{i'k}\right)^{p-1} = p_{ik}.$$

Hence, $\forall ik, \gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q}) = 0$ implies that $\mathbf{p} = \mathbf{q}$. Consider now the boundary. If for any ik, $p_{ik} \neq 0$ and $q_{ik} = 0$ then $B_\phi(\mathbf{p},\mathbf{q}) = \infty \neq 0$. So, if $B_\phi(\mathbf{p},\mathbf{q}) = 0$, then whenever $q_{ik} = 0$ we must have $p_{ik} = 0$. On the other hand, if $p_{ik} = 0$, there will be a contribution to $B_\phi(\mathbf{p},\mathbf{q})$ of $\frac{1}{p}q_{ik}$, implying that q_{ik} must be 0 in order for $B_\phi(\mathbf{p},\mathbf{q}) = 0$. Thus, A3. holds.

We now show A4. Let us define the boundary values for $\mathcal{L}_\phi$. If for some k we have $\sum_{i'} q_{i'k} = 0$, then let $(\mathcal{L}_\phi(\mathbf{q},\mathbf{v}))_{ik} = 0$ for all i. Otherwise, (12) can be used as written. Thus, we always have $\mathcal{L}_\phi(\mathbf{q},\mathbf{0}) = \mathbf{q}$, and $\mathcal{L}_\phi(\mathbf{q},\mathbf{v})$ is jointly continuous in $\mathbf{q}$ and $\mathbf{v}$. Now consider $B_\phi(\mathcal{L}_\phi(\mathbf{q},\mathbf{v}),\mathbf{q})$. Let us simplify this expression in the interior, starting from (12) and (13) and using the notation $\{\mathcal{L}\}_{ik}$ for the vector $\{\mathcal{L}_\phi(\mathbf{q},\mathbf{v})\}_{ik}$.

$$\begin{aligned}
B_\phi(\mathcal{L},\mathbf{q}) &= \sum_{ik}\mathcal{L}_{ik}\Big(\gamma(\mathcal{L}_{ik},\mathcal{L}) - \gamma(q_{ik},\mathbf{q})\Big) - \frac{1}{p}\sum_{ik}(\mathcal{L}_{ik} - q_{ik}) \\
&= \sum_{ik}\mathcal{L}_{ik}\ln\left(\frac{\mathcal{L}_{ik}}{p^{1/p}(\sum_{i'}\mathcal{L}_{i'k})^{(p-1)/p}}\frac{p^{1/p}(\sum_{i'}q_{i'k})^{(p-1)/p}}{q_{ik}}\right) - \frac{1}{p}\sum_{ik}(\mathcal{L}_{ik} - q_{ik}) \\
&= \sum_{ik}\mathcal{L}_{ik}\ln\left[\frac{e^{v_{ik}}q_{ik}(\sum_{i'}e^{v_{i'k}}q_{i'k})^{p-1}\left(\frac{1}{\sum_{i'}q_{i'k}}\right)^{p-1}}{\left[(\sum_{i'}e^{v_{i'k}}q_{i'k})^p\left(\frac{1}{\sum_i q_{i'k}}\right)^{p-1}\right]^{(p-1)/p}}\frac{(\sum_{i'}q_{i'k})^{(p-1)/p}}{q_{ik}}\right] \\
&\qquad -\frac{1}{p}\sum_{ik}(\mathcal{L}_{ik} - q_{ik}) \\
&= \sum_{ik}\mathcal{L}_{ik}v_{ik} - \frac{1}{p}\sum_{ik}(\mathcal{L}_{ik} - q_{ik}).
\end{aligned}$$

Thus, since $\mathcal{L}_\phi$ is jointly continuous in $\mathbf{q}$ and $\mathbf{v}$, B_ϕ is jointly continuous in $\mathbf{q}$ and $\mathbf{v}$.

For A5., we need to show that $B_\phi(\mathbf{p},\cdot)$ is coercive, meaning that the level set $\{\mathbf{q} \in \Delta_\phi : B_\phi(\mathbf{p},\mathbf{q}) \leq c\}$ is bounded, with $\mathbf{p} \in \Delta_\phi \backslash \mathrm{int}(\Delta_\phi)$ which are vectors in $\mathcal{R}_+^{IK}$ with at least one entry that is 0. Recall that we use the convention $0\ln 0 = 0$. Consider from (13), using the fact that for any ik pair,

$\ln q_{ik}^{(p-1)/p} \leq \ln \left(\sum_i q_{ik}\right)^{(p-1)/p}$:

$$\begin{aligned} B_\phi(\mathbf{p},\mathbf{q}) &= \sum_{ik} p_{ik}\Big(\gamma(p_{ik},\mathbf{p}) - \gamma(q_{ik},\mathbf{q})\Big) - \frac{1}{p}\sum_{ik}(p_{ik}-q_{ik}) \\ &= \sum_{ik} -p_{ik}\gamma(q_{ik},\mathbf{q}) + \frac{1}{p}q_{ik} + \text{function}(\mathbf{p}) \\ &\geq \sum_{ik} -p_{ik}\ln q_{ik} + p_{ik}\ln\left(q_{ik}^{(p-1)/p}\right) + \frac{1}{p}q_{ik} + \text{function}(\mathbf{p}) \\ &= \frac{1}{p}\sum_{ik} -p_{ik}\ln q_{ik} + q_{ik} + \text{function}(\mathbf{p}). \end{aligned}$$

Since logarithms grow slowly, one can choose a q_{ik} large enough so that this sum exceeds any fixed constant c, regardless of the values of the other q_{ik}'s. Thus, the set $\{\mathbf{q} \in \Delta_\phi : B_\phi(\mathbf{p},\mathbf{q}) \leq c\}$ is bounded. We are done checking the conditions. ■

References

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

Arthur Asuncion and David J. Newman. UCI machine learning repository, 2007. URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

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

Ulf Brefeld and Tobias Scheffer. AUC maximizing support vector learning. In *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning*, 2005.

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

Stéphan Clemençon and Nicolas Vayatis. Empirical performance maximization for linear rank statistics. In *Advances in Neural Information Processing Systems 22*, 2008.

Stéphan Clemençon, Gabor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1/2/3), 2002.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

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

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49, 2002.

Ofer Dekel, Christopher Manning, and Yoram Singer. Log-linear models for label ranking. In *Advances in Neural Information Processing Systems 16*, 2004.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Duality and auxiliary functions for Bregman distances. Technical Report CMU-CS-01-109R, School of Computer Science, Carnegie Mellon University, 2002.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

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.

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

Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *SIGIR '00: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, New York, NY, USA, 2000. ACM.

Heng Ji, Cynthia Rudin, and Ralph Grishman. Re-ranking algorithms for name tagging. In *HLT/NAACL workshop on computationally hard problems and joint interference in speech and language processing*, 2006.

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

Quoc Le and Alex Smola. Direct optimization of ranking measures. arXiv:0704.3359v1, November 2007.

Sofus A. Macskassy, Foster Provost, and Saharon Rosset. Pointwise ROC confidence bounds: An empirical evaluation. In *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning*, 2005.

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

Michael C. Mozer, Robert Dodier, Michael D. Colagrosso, Csar Guerra-salcedo, and Richard Wolniewicz. Prodding the ROC curve: Constrained optimization of classifier performance. In *Advances in Neural Information Processing Systems 14*, pages 1409–1415, 2002.

Alain Rakotomamonjy. Optimizing AUC with support vector machine (SVM). In *Proceedings of European Conference on Artificial Intelligence Workshop on ROC Curve and AI, Valencia, Spain*, 2004.

Cynthia Rudin. Ranking with a p-norm push. In *Proceedings of the Ninteenth Annual Conference on Learning Theory*, pages 589–604, 2006.

Cynthia Rudin and Robert E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research*, 10:2193–2232, October 2009.

Cynthia Rudin, Corinna Cortes, Mehryar Mohri, and Robert E. Schapire. Margin-based ranking meets boosting in the middle. In Peter Auer and Ron Meir, editors, *Proceedings of the Eighteenth Annual Conference on Learning Theory*, pages 63–78. Springer, 2005.

Cynthia Rudin, Rebecca Passonneau, Axinia Radeva, Haimonti Dutta, Steve Ierome, and Delfina Isaac. Predicting manhole events in Manhattan : A case study in extended knowledge discovery. Accepted for publication to Machine Learning, 2009.

Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

Shai Shalev-Shwartz and Yoram Singer. Efficient learning of label ranking by soft projections onto polyhedra. *Journal of Machine Learning Research*, 7:1567–1599, December 2006.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, Sept 2005.

Nicolas Usunier, Massih-Reza Amini, and Patrick Gallinari. A data-dependent generalisation error bound for the AUC. In *Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning*, 2005.

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

Tong Zhang and Bin Yu. Boosting with early stopping - convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579, 2005.

Zhaohui Zheng, Hongyuan Zha, Tong Zhang, Olivier Chapelle, Keke Chen, and Gordon Sun. A general boosting method and its application to learning ranking functions for web search. In *Advances in Neural Information Processing Systems 19*, 2007.