Surrogate Functions for Maximizing Precision at the Top

Purushottam Kar
Microsoft Research India
t-purkar@microsoft.com

Harikrishna Narasimhan[*]
Indian Institute of Science
harikrishna@csa.iisc.ernet.in

Prateek Jain
Microsoft Research India
prajain@microsoft.com

Abstract

The problem of maximizing precision at the top of a ranked list, often dubbed Precision@k (prec@k), finds relevance in myriad learning applications such as ranking, multi-label classification, and learning with severe label imbalance. However, despite its popularity, there exist significant gaps in our understanding of this problem and its associated performance measure.

The most notable of these is the lack of a convex upper bounding surrogate for prec@k. We also lack scalable perceptron and stochastic gradient descent algorithms for optimizing this performance measure. In this paper we make key contributions in these directions. At the heart of our results is a family of truly upper bounding surrogates for prec@k. These surrogates are motivated in a principled manner and enjoy attractive properties such as consistency to prec@k under various natural margin/noise conditions.

These surrogates are then used to design a class of novel perceptron algorithms for optimizing prec@k with provable mistake bounds. We also devise scalable stochastic gradient descent style methods for this problem with provable convergence bounds. Our proofs rely on novel uniform convergence bounds which require an in-depth analysis of the structural properties of prec@k and its surrogates. We conclude with experimental results comparing our algorithms with state-of-the-art cutting plane and stochastic gradient algorithms for maximizing prec@k.

1 Introduction

Ranking a given set of points or labels according to their relevance forms the core of several real-life learning systems. For instance, in classification problems with a rare-class as is the case in spam/anomaly detection, the goal is to rank the given emails/events according to their likelihood of being from the rare-class (spam/anomaly). Similarly, in multi-label classification problems, the goal is to rank the labels according to their likelihood of being relevant to a data point Tsoumakas and Katakis [2007].

The ranking of items at the top is of utmost importance in these applications and several performance measures, such as Precision@k, Average Precision and NDCG have been designed to promote accuracy at top of ranked lists. Of these, the Precision@k (prec@k) measure is especially popular in a variety of domains. Informally, prec@k counts the number of relevant items in the top-k positions of a ranked list and is widely used in domains such as binary classification Joachims [2005], multi-label classification Prabhu and Varma [2014] and ranking Le and Smola [2007].

Given its popularity, prec@k has received attention from algorithmic, as well as learning theoretic perspectives. However, there remain specific deficiencies in our understanding of this performance measure. In fact, to the best of our knowledge, there is only one known convex surrogate function for prec@k, namely, the *struct-SVM* surrogate due to Joachims [2005] which, as we reveal in this work, is not an upper bound on prec@k in general, and need not recover an optimal ranking even in strictly separable settings.

Our aim in this paper is to develop efficient algorithms for optimizing prec@k for ranking problems with binary relevance levels. Since the intractability of binary classification in the agnostic setting Guruswami and Raghavendra [2009] extends to prec@k, our goal would be to exploit natural notions of *benign-ness* usually observed in natural distributions to overcome such intractability results.

[*] Work done while H.N. was an intern at Microsoft Research India, Bangalore.

1.1 Our Contributions

We make several contributions in this paper that both, give deeper insight into the prec@k performance measure, as well as provide scalable techniques for optimizing it.

Precision@k margin: motivated by the success of margin-based frameworks in classification settings, we develop a family of margin conditions appropriate for the prec@k problem. Recall that the prec@k performance measure counts the number of relevant items at the top k positions of a ranked list. The simplest of our margin notions, that we call the *weak* (k, γ)*-margin*, is said to be present if a privileged set of k relevant items can be separated from all irrelevant items by a margin of γ. This is the least restrictive margin condition that allows for a perfect ranking w.r.t prec@k. Notably, it is much less restrictive than the binary classification notion of margin which requires all relevant items to be separable from all irrelevant items by a certain margin. We also propose two other notions of margin suited to our perceptron algorithms.

Surrogate functions for prec@k: we design a family of three novel surrogates for the prec@k performance measure. Our surrogates satisfy two key properties. Firstly they always upper bound the prec@k performance measure so that optimizing them promotes better performance w.r.t prec@k. Secondly, these surrogates satisfy *conditional consistency* in that they are consistent w.r.t. prec@k under some noise condition. We show that there exists a one-one relationship between the three prec@k margin conditions mentioned earlier and these three surrogates so that each surrogate is consistent w.r.t. prec@k under one of the margin conditions. Moreover, our discussion reveals that the three surrogates, as well as the three margin conditions, lie in a concise hierarchy.

Perceptron and SGD algorithms: using insights gained from the previous analyses, we design two perceptron-style algorithms for optimizing prec@k. Our algorithms can be shown to be a natural extension of the classical perceptron algorithm for binary classification Rosenblatt [1958]. Indeed, akin to the classical perceptron, both our algorithms enjoy mistake bounds that reduce to crisp convergence bounds under the margin conditions mentioned earlier. We also design a mini-batch-style stochastic gradient descent algorithm for optimizing prec@k.

Learning theory: in order to prove convergence bounds for the SGD algorithm, and online-to-batch conversion bounds for our perceptron algorithms, we further study prec@k and its surrogates and prove uniform convergence bounds for the same. These are novel results and require an in-depth analysis into the involved structure of the prec@k performance measure and its surrogates. However, with these results in hand, we are able to establish crisp convergence bounds for the SGD algorithm, as well as generalization bounds for our perceptron algorithms.

Paper Organization: Section 2 presents the problem formulation and sets up the notation. Section 3 introduces three novel surrogates and margin conditions for prec@k and reveals the interplay between these with respect to consistency to prec@k. Section 4 presents two perceptron algorithms for prec@k and their mistake bounds, as well as a mini-batch SGD-based algorithm. Section 5 discusses uniform convergence bounds for our surrogates and their application to convergence and online-to-batch conversion bounds for our the perceptron and SGD-style algorithms. We conclude with empirical results in Section 6.

1.2 Related Work

There has been much work in the last decade in designing algorithms for bipartite ranking problems. While the earlier methods for this problem, such as RankSVM, focused on optimizing pair-wise ranking accuracy Herbrich et al. [2000], Joachims [2002], Freund et al. [2003], Burges et al. [2005], of late, there has been enormous interest in performance measures that promote good ranking performance at the top portion of the ranked list, and in ranking methods that directly optimize these measures Clémençon and Vayatis [2007], Rudin [2009], Agarwal [2011], Boyd et al. [2012], Narasimhan and Agarwal [2013a,b], Li et al. [2014].

In this work, we focus on one such evaluation measure – Precision@k, which is widely used in practice. The only prior algorithms that we are aware of that directly optimize this performance measure are a structural SVM based cutting plane method due to Joachims [2005], and an efficient stochastic implementation of the same due to Kar et al. [2014]. However, as pointed out earlier, the convex surrogate used in these methods is not well-suited for prec@k.

It is also important to note that the bipartite ranking setting considered in this work is different from other popular forms of ranking such as subset or list-wise ranking settings, which arise in several information retrieval applications, where again there has been much work in optimizing performance measures that emphasize on accuracy at the top (e.g. NDCG) Valizadegan et al. [2009], Cao et al. [2007], Yue et al. [2007], Le and Smola [2007], Chakrabarti et al. [2008], Yun et al. [2014]. There has also been some recent work on perceptron style ranking methods for list-wise ranking problems Chaudhuri and Tewari [2014], but these methods are tailored to optimize the NDCG and MAP measures, which are different from the prec@k measure that we consider here. Other less related works include online ranking algorithms for optimizing ranking measures in an adversarial setting with limited feedback Chaudhuri and Tewari [2015].

2 Problem Formulation and Notation

We will be presented with a set of labeled points $(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{0,1\}$. We shall use $\mathbf{X}$ to denote the entire dataset, $\mathbf{X}_+$ and $\mathbf{X}_-$ to denote the set of positive and negatively (null) labeled points, and $\mathbf{y} \in \{0,1\}^n$ to denote the label vector. $\mathbf{z} = (\mathbf{x}, y)$ shall denote a labeled data point. Our results readily extend to multi-label and ranking settings but for sake of simplicity, we focus only on bipartite ranking problems, where the goal is to rank (a subset of) positive examples above the negative ones.

Given n labeled data points $\mathbf{z}_1, \ldots, \mathbf{z}_n$ and a scoring function $s : \mathcal{X} \rightarrow \mathbb{R}$, let $\sigma_s \in S_n$ be the permutation that sorts points according to the scores given by s i.e. $s(\mathbf{x}_{\sigma_s(i)}) \geq s(\mathbf{x}_{\sigma_s(j)})$ for $i \leq j$. The Precision@k measure for this scoring function can then be expressed as:

$$\text{prec@k}(s;\ \mathbf{z}_1, \ldots, \mathbf{z}_n) = \sum_{i=1}^{k} (1 - \mathbf{y}_{\sigma_s(i)}). \tag{1}$$

Note that the above is a "loss" version of the performance measure which penalizes any top-k ranked data points that have a null label. For simplicity, we will use the abbreviated notation $\text{prec@k}(s) := \text{prec@k}(s;\ \mathbf{z}_1, \ldots, \mathbf{z}_n)$. We will also use the shorthand $s_i = s(\mathbf{x}_i)$. For any label vectors $\mathbf{y}', \mathbf{y}'' \in \{0,1\}^n$, we define

$$\begin{aligned} \Delta(\mathbf{y}', \mathbf{y}'') &= \sum_{i=1}^{n} (1 - \mathbf{y}'_i)\mathbf{y}''_i, \\ K(\mathbf{y}', \mathbf{y}'') &= \sum_{i=1}^{n} \mathbf{y}'_i \mathbf{y}''_i. \end{aligned} \tag{2}$$

Let $n_+(\mathbf{y}') = K(\mathbf{y}', \mathbf{y}') = \|\mathbf{y}'\|_1$ denote the number of positives in the label vector $\mathbf{y}'$ and $n_+ = n_+(\mathbf{y})$ denote the number of actual positives. Let $\mathbf{y}^{(s,k)}$ be the label vector that assigns the label 1 only to the top k ranked items according to the scoring function s. That is, $\mathbf{y}_i^{(s,k)} = 1$ if if $\sigma_s^{-1}(i) \leq k$ and 0 otherwise. It is easy to verify that for any scoring function s, $\Delta(\mathbf{y}, \mathbf{y}^{(s,k)}) = \text{prec@k}(s)$.

3 A Family of Novel Surrogates for prec@k

As prec@k is a non-convex loss function that is hard to optimize directly, it is natural to seek surrogate functions that act as a good proxy for prec@k. There will be two properties that we shall desire of such a surrogate:

1. **Upper Bounding Property**: the surrogate should *upper bound* the prec@k loss function, so that minimizing the surrogate promotes small prec@k loss.
2. **Conditional Consistency**: under some regularity assumptions, optimizing the surrogate should yield an optimal solution for prec@k as well.

Motivated by the above requirements, we develop a family of surrogates which upper bound the prec@k loss function and are consistent to it under certain margin/noise conditions. We note that the results of Calauzènes et al. [2012] that negate the possibility of consistent convex surrogates for ranking performance measures do not apply to our results since they are neither stated for prec@k, nor do they negate the possibility of conditional consistency.

It is notable that the seminal work of Joachims [2005] did propose a convex surrogate for prec@k, that we refer to as $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$. However, as the discussion below shows, this surrogate is not even an upper bound on prec@k let alone be consistent to it. Understanding the reasons for the failure of this surrogate would be crucial in designing our own.

3.1 The Curious Case of $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$

The $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ surrogate is a part of a broad class of surrogates called *struct-SVM* surrogates that are designed for structured output prediction problems that can have exponentially large output spaces Joachims [2005]. Given a set of n labeled data points, $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ is defined as

$$\max_{\substack{\hat{\mathbf{y}} \in \{0,1\}^n \\ \|\hat{\mathbf{y}}\|_1 = k}} \left\{ \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \mathbf{y}_i)\, s_i \right\}. \tag{3}$$

The above surrogate penalizes a scoring function if there exists a set of k points with large scores (i.e. the second term is large) which are actually negatives (i.e. the first term is large). However, since the candidate labeling $\hat{\mathbf{y}}$ is restricted to labeling just k points as positive whereas the true label vector $\mathbf{y}$ has n_+ positives, in cases where $n_+ > k$, a non-optimal candidate labeling $\hat{\mathbf{y}}$ can exploit the remaining $n_+ - k$ labels to hide the high scoring negative points, thus confusing the surrogate function. This indicates that this surrogate may not be an upper bound to prec@k. We refer the reader to Appendix A for an explicit example where, not only does this surrogate not upper bound prec@k, but more importantly, minimizing $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ does not produce a model that is optimal for prec@k, even in separable settings where all positives points are separated from negatives by a margin.

In the sequel, we shall propose three surrogates, all of which are consistent with prec@k under various noise/margin conditions. The surrogates, as well as the noise conditions, will be shown to form a hierarchy.

3.2 The Ramp Surrogate $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$

The key to maximizing prec@k in a bipartite ranking setting is to select a subset of k relevant items and rank them at the top k positions. This can happen iff the *top ranked* k relevant items are not outranked by any irrelevant item. Thus, a surrogate must penalize a scoring function that assigns scores to irrelevant items that are higher than those of the top ranked relevant items. Our *ramp* surrogate $\ell^{\text{ramp}}_{\text{prec@k}}(s)$ implicitly encodes this strategy:

$$\max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i \right\} - \underbrace{\max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i}_{(P)}. \tag{4}$$

The term (P) contains the sum of scores of the k highest scoring positives. Note that $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$ is similar to the "ramp" losses for binary classification Do et al. [2008]. We now show that $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$ is indeed an upper bounding surrogate for prec@k.

Claim 1. *For any $k \leq n_+$ and scoring function s, we have $\ell^{\text{ramp}}_{\text{prec@k}}(s) \geq \text{prec@k}(s)$. Moreover, if $\ell^{\text{ramp}}_{\text{prec@k}}(s) \leq \xi$ for a given scoring function s, then there necessarily exists a set $S \subset [n]$ of size at most k such that for all $\|\hat{\mathbf{y}}\|_1 = k$, we have $\sum_{i \in S} s_i \geq \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i + \Delta(\mathbf{y}, \hat{\mathbf{y}}) - \xi$.*

Proofs for this section are deferred to Appendix B. We can show that this surrogate is conditionally consistent as well. To do so, we introduce the notion of *weak (k, γ)-margin.*

Definition 2 (*Weak* (k, γ)-margin)**.** *A set of n labeled data points satisfies the* weak (k, γ)*-margin condition if for some scoring function s and set $S_+ \subseteq \mathbf{X}_+$ of size k,*

$$\min_{i \in S_+} s_i - \max_{j : \mathbf{y}_j = 0} s_j \geq \gamma.$$

Moreover, we say that the function s realizes *this margin. We abbreviate the* weak $(k, 1)$*-margin condition as simply the* weak *k-margin condition.*

Clearly, a dataset has a *weak* (k,γ)-margin iff there exist some k positive points that substantially outrank all negatives. Note that this notion of margin is strictly weaker than the standard notion of margin for binary classification as it allows all but those k positives to be completely mingled with the negatives. Moreover, this seems to be one of the most natural notions of margin for prec@k. The following lemma establishes that $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$ is indeed consistent w.r.t. prec@k under the *weak* k-margin condition.

Claim 3. *For any scoring function s that realizes the* weak *k-margin over a dataset, $\ell^{ramp}_{\text{prec@k}}(s) = \text{prec@k}(s) = 0$.*

This suggests that $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$ is not only a tight surrogate, but tight at the optimal scoring function, i.e. $\text{prec@k}(s) = 0$; this along with upper bounding property implies consistency. However, it is also a non-convex function due to the term (P). To obtain convex surrogates, we perform relaxations on this term by first rewriting it as follows:

$$(P) = \sum_{i=1}^{n} \mathbf{y}_i s_i - \underbrace{\min_{\substack{\tilde{\mathbf{y}} \preceq \mathbf{y} \\ \|\tilde{\mathbf{y}}\|_1 = n_+ - k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i}_{(Q)}, \tag{5}$$

where $\tilde{\mathbf{y}} \preceq \mathbf{y}$ implies that $\mathbf{y}_i = 0 \Rightarrow \tilde{\mathbf{y}}_i = 0$. Thus, to convexify the surrogate $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$, we need to design a convex upper bound on (Q). Notice that the term (Q) contains the sum of the scores of the $n_+ - k$ lowest ranked positive data points. This can be readily upper bounded in several ways which give us different surrogate functions.

3.3 The Max Surrogate $\ell^{\textbf{max}}_{\textbf{prec@k}}(\cdot)$

An immediate convex upper bound on (Q) is obtained by replacing the sum of scores of the $n_+ - k$ lowest ranked positives with those of the highest ranked ones as follows: $(Q) \leq \max_{\substack{\tilde{\mathbf{y}} \preceq (1-\hat{\mathbf{y}})\cdot\mathbf{y} \\ \|\tilde{\mathbf{y}}\|_1 = n_+ - k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i$, which gives us the $\ell^{\text{max}}_{\text{prec@k}}(s)$ surrogate defined below:

$$\max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y},\hat{\mathbf{y}}) + \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \max_{\substack{\tilde{\mathbf{y}} \preceq (1-\hat{\mathbf{y}})\cdot\mathbf{y} \\ \|\tilde{\mathbf{y}}\|_1 = n_+ - k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \right\}. \tag{6}$$

The above surrogate, being a point-wise maximum over convex functions, is convex, as well as an upper bound on $\text{prec@k}(s)$ since it upper bounds $\ell^{\text{ramp}}_{\text{prec@k}}(s)$. This surrogate can also be shown to be consistent w.r.t. prec@k under the *strong* γ-margin condition defined below for $\gamma = 1$.

Definition 4 (*Strong* γ-margin)**.** *A set of n labeled data points satisfies the γ-*strong *margin condition if for some scoring function s,* $\min_{i:\mathbf{y}_i=1} s_i - \max_{j:\mathbf{y}_j=0} s_j \geq \gamma$.

We notice that the strong margin condition is actually the standard notion of binary classification margin and hence much stronger than the *weak* (k,γ)-margin condition. It also does not incorporate any elements of the prec@k problem. This leads us to look for tighter convex relaxations to the term (Q) that we do below.

3.4 The Avg Surrogate $\ell^{\textbf{avg}}_{\textbf{prec@k}}(\cdot)$

A tighter upper bound on (Q) can be obtained by replacing (Q) by the average score of the false negatives. Define $C(\hat{\mathbf{y}}) = \frac{n_+ - K(\mathbf{y},\hat{\mathbf{y}})}{n_+ - k}$ and consider the relaxation $(Q) \leq \frac{1}{C(\hat{\mathbf{y}})} \sum_{i=1}^{n} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i$. Combining this with (4), we get a new convex surrogate $\ell^{\text{avg}}_{\text{prec@k}}(s)$ defined as:

$$\max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y},\hat{\mathbf{y}}) + \sum_{i=1}^{n} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \frac{1}{C(\hat{\mathbf{y}})} \sum_{i=1}^{n} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \right\}. \tag{7}$$

We refer the reader to Appendix B.4 for a proof that $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ is an upper bounding surrogate. It is notable that for $k = n_+$ (i.e. for the PRBEP measure), the surrogate $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ recovers Joachims' original surrogate $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$. To establish conditional consistency of this surrogate, consider the following notion of margin:

$$\begin{array}{cccccc} \text{prec@k}(s) \le & \ell^{\text{ramp}}_{\text{prec@k}}(s) & \le & \ell^{\text{avg}}_{\text{prec@k}}(s) & \le & \ell^{\text{max}}_{\text{prec@k}}(s) \\ & \Uparrow & & \Uparrow & & \Uparrow \\ & \text{weak } (k,\gamma) & \supseteq & (k,\gamma) & \supseteq & \text{strong } \gamma \\ & \text{margin} & & \text{margin} & & \text{margin} \end{array}$$

Figure 1: A hierarchy among the three surrogates for prec@k and the corresponding margin conditions for conditional consistency.

Definition 5 ((k, γ)-margin)**.** *A set of n labeled data points satisfies the (k, γ)-margin condition if for some scoring function s, we have, for all sets $S_+ \subseteq \mathbf{X}_+$ of size $n_+ - k + 1$,*

$$\frac{1}{n_+ - k + 1} \sum_{i \in S_+} s_i - \max_{j : \mathbf{y}_j = 0} s_j \ge \gamma.$$

Moreover, we say that the function s realizes *this margin. We abbreviate the $(k, 1)$-margin condition as simply the k-margin condition.*

We can now establish the consistency of $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ under the k-margin condition. See Appendix B.5 for a proof.

Claim 6. *For any scoring function s that realizes the k-margin over a dataset, $\ell^{avg}_{\text{prec@k}}(s) = \text{prec@k}(s) = 0$.*

We note that the (k, γ)-margin condition is strictly weaker than the *strong* γ-margin condition (Definition 4) since it still allows a non negligible fraction of the positive points to be assigned a lower score than those assigned to negatives. On the other hand, the (k, γ)-margin condition is strictly stronger than the *weak* (k, γ)-margin condition (Definition 2). The weak k-margin condition only requires one set of k-positives to be separated from the negatives, whereas the above margin condition at least requires the average of *all* positives to be separated from the negatives.

As Figure 1 demonstrates, the three surrogates presented above, as well as their corresponding margin conditions, fall in a neat hierarchy. We will now use these surrogates to formulate two perceptron algorithms with mistake bounds with respect to these margin conditions.

4 Perceptron & SGD Algorithms for prec@k

We now present perceptron-style algorithms for maximizing the prec@k performance measure in bipartite ranking settings. Our algorithms work with a stream of binary labeled points and process them in *mini-batches* of a predetermined size b. Mini-batch methods have recently gained popularity and have been used to optimize ranking loss functions such as $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ as well Kar et al. [2014]. It is useful to note that the requirement for mini-batches goes away in ranking and multi-label classification settings, for our algorithms can be applied to individual data points in those settings (e.g. individual queries in ranking settings).

At every time instant t, our algorithms receive a batch of b points $\mathbf{X}_t = \left[\mathbf{x}_t^1, \ldots, \mathbf{x}_t^b\right]$ and rank these points using the existing model. Let Δ_t denote the prec@k loss (equation 1) at time t. If $\Delta_t = 0$ i.e. all top k ranks are occupied by positive points, then the model is not updated. Otherwise, the model is updated using the false positives and negatives. For sake of simplicity, we will only look at linear models in this paper. Depending on the kind of updates we make, we get two variants of the perceptron rule for prec@k.

Our first algorithm, PERCEPTRON@K-AVG, updates the model using a combination of all the false positives and negatives (see Algorithm 1). The effect of the update is a very natural one – it explicitly boosts the scores of the positive points that failed to reach the top ranks, and attenuates the scores of the negative points that got very high scores. It is interesting to note that in the limiting case of $k = 1$ and unit batch length (i.e. $b = 1$), the PERCEPTRON@K-AVG update reduces to that of the standard perceptron algorithm Rosenblatt [1958], Minsky and Papert [1988] for the choice $\hat{\mathbf{y}}_t = \text{sign}(s_t)$. Thus, our algorithm can be seen as a natural extension of the classical perceptron algorithm.

Algorithm 1 PERCEPTRON@K-AVG

```
Input: Batch length $b$
$\mathbf{w}^0 \leftarrow \mathbf{0}, t \leftarrow 0$
while stream not exhausted do
    $t \leftarrow t + 1$
    Receive $b$ data points $\mathbf{X}_t = \left[\mathbf{x}_t^1, \ldots, \mathbf{x}_t^b\right], \mathbf{y}_t \in \{0,1\}^b$
    Calculate $s_t = \mathbf{w}^{t-1}\mathbf{X}_t$ and let $\hat{\mathbf{y}}_t = \mathbf{y}^{(s_t,k)}$
    $\Delta_t \leftarrow \Delta(\mathbf{y}_t, \hat{\mathbf{y}}_t)$
    if $\Delta_t = 0$ then
        $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1}$
    else
        $D_t \leftarrow \frac{\Delta_t}{\|\mathbf{y}_t\|_1 - K(\mathbf{y}_t, \hat{\mathbf{y}}_t)}$
        $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \sum_{i \in [b]} (1 - \mathbf{y}_i)\hat{\mathbf{y}}_i \cdot \mathbf{x}_t^i$          {false positives}
        $\mathbf{w}^t \leftarrow \mathbf{w}^t + D_t \cdot \sum_{i \in [b]} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i \cdot \mathbf{x}_t^i$          {false negatives}
    end if
end while
return $\mathbf{w}^t$
```

Algorithm 2 PERCEPTRON@K-MAX

```
10:     $S_t \leftarrow \text{FN}(s, \Delta_t)$
11:     $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \sum_{i \in [b]} (1 - \mathbf{y}_i)\hat{\mathbf{y}}_i \cdot \mathbf{x}_t^i$          {false positives}
12:     $\mathbf{w}^t \leftarrow \mathbf{w}^t + \sum_{i \in S_t} \mathbf{x}_t^i$          {top ranked false negatives}
```

The next lemma establishes that, similar to the classical perceptron Novikoff [1962], PERCEPTRON@K-AVG also enjoys a mistake bound. Our mistake bound is stated in the most general agnostic setting with the hinge loss function replaced with our surrogate $\ell^{\text{avg}}_{\text{prec@k}}(s)$. All proofs in this section are deferred to Appendix C.

Theorem 7. *Suppose $\left\|\mathbf{x}_t^i\right\| \leq R$ for all t, i. Let $\Delta_T^C = \sum_{t=1}^T \Delta_t$ be the cumulative mistake value observed when Algorithm 1 is executed for T batches. Also, for any $\mathbf{w}$, let $\hat{\mathcal{L}}_T^{avg}(\mathbf{w}) = \sum_{t=1}^T \ell^{avg}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)$. Then we have*

$$\Delta_T^C \leq \min_{\mathbf{w}} \left(\|\mathbf{w}\| \cdot R \cdot \sqrt{4k} + \sqrt{\hat{\mathcal{L}}_T^{avg}(\mathbf{w})} \right)^2.$$

Similar to the classical perceptron mistake bound Novikoff [1962], the above bound can also be reduced to a simpler convergence bound in *separable settings*.

Corollary 8. *Suppose a unit norm $\mathbf{w}^*$ exists such that the scoring function $s : \mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}^*$ realizes the (k, γ)-margin condition for all the batches, then Algorithm 1 guarantees the mistake bound: $\Delta_T^C \leq \frac{4kR^2}{\gamma^2}$.*

The above result assures that, as datasets become "easier" in the sense that their (k, γ)-margin becomes larger, PERCEPTRON@K-AVG will converge to an optimal hyperplane at a faster rate. It is important to note there that the (k, γ)-margin condition is strictly weaker than the standard classification margin condition. Hence for several datasets, PERCEPTRON@K-AVG might be able to find a perfect ranking while at the same time, it might be impossible for standard binary classification techniques to find any reasonable classifier in poly-time Guruswami and Raghavendra [2009].

We note that PERCEPTRON@K-AVG performs updates with all the false negatives in the mini-batches. This raises the question as to whether sparser updates are possible as such updates would be slightly faster as well as, in high dimensional settings, ensure that the model is sparser. To this end we design the PERCEPTRON@K-MAX algorithm (Algorithm 2). PERCEPTRON@K-MAX differs from PERCEPTRON@K-AVG in that it performs updates using only a few of the *top ranked* false negatives. More specifically, for any scoring function s and $m > 0$, define:

$$\text{FN}(s, m) = \underset{S \subset \mathbf{X}_t^+, |S| = m}{\arg\max} \sum_{i \in S} \left(1 - \mathbf{y}_i^{(s,k)}\right) \mathbf{y}_i s_i$$

Algorithm 3 SGD@K-AVG

```
Input: Batch length $b$, step lengths $\eta_t$, feasible set $\mathcal{W}$
Output: A model $\bar{\mathbf{w}} \in \mathcal{W}$
$\mathbf{w}^0 \leftarrow \mathbf{0}, t \leftarrow 0$
while stream not exhausted do
    $t \leftarrow t + 1$
    Receive $b$ data points $\mathbf{X}_t = \left[\mathbf{x}_t^1, \ldots, \mathbf{x}_t^b\right], \mathbf{y}_t \in \{0,1\}^b$
    Set $\mathbf{g}_t \in \partial_{\mathbf{w}} \ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}_{t-1}; \mathbf{X}_t, \mathbf{y}_t)$        {See Algorithm 4}
    $\mathbf{w}_t \leftarrow \Pi_{\mathcal{W}}\left[\mathbf{w}_{t-1} - \eta_t \cdot \mathbf{g}_t\right]$        {project onto set $\mathcal{W}$}
end while
return $\bar{\mathbf{w}} = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{w}_\tau$
```

Algorithm 4 Subgradient calculation for $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$

```
Input: A model $\mathbf{w}_{\text{in}}$, $n$ data points $\mathbf{X}, \mathbf{y}$, parameter $k$
Output: A subgradient $\mathbf{g} \in \partial_{\mathbf{w}} \ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}_{\text{in}}; \mathbf{X}, \mathbf{y})$
Sort pos. and neg. points separately in dec. order of scores assigned by $\mathbf{w}_{\text{in}}$ i.e. $s_1^+ \geq \ldots \geq s_{n_+}^+$ and $s_1^- \geq \ldots \geq s_{n_-}^-$
for $k' = 0 \rightarrow k$ do
    $D_{k'} \leftarrow \frac{k-k'}{n_+ - k'}$
    $\Delta_{k'} \leftarrow k - k' - D_{k'} \sum_{i=k'+1}^{n_+} s_i^+ + \sum_{i=1}^{k-k'} s_i^-$
    $\mathbf{g}_{k'} \leftarrow \sum_{i=1}^{k-k'} \mathbf{x}_i^- - D_{k'} \sum_{i=k'+1}^{n_+} \mathbf{x}_i^+$
end for
$k^* \leftarrow \arg\max_{k'} \Delta_{k'}$
return $\mathbf{g}_{k^*}$
```

as the set of the m top ranked false negatives. PERCEPTRON@K-MAX makes updates only for false positives in the set $\text{FN}(s, \Delta_t)$. Note that Δ_t can significantly smaller than the total number of false negatives if $k \ll n_+$. PERCEPTRON@K-MAX also enjoys a mistake bound but with respect to the $\ell^{\max}_{\text{prec@k}}(\cdot)$ surrogate.

Theorem 9. *Suppose* $\left\|\mathbf{x}_t^i\right\| \leq R$ *for all* t, i. *Let* $\Delta_T^C = \sum_{t=1}^{T} \Delta_t$ *be the cumulative observed mistake value when Algorithm 2 is executed for* T *batches. Also, for any* $\mathbf{w}$, *let* $\hat{\mathcal{L}}_T^{max}(\mathbf{w}) = \sum_{t=1}^{T} \ell^{max}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)$. *Then we have*

$$\Delta_T^C \leq \min_{\mathbf{w}} \left(\|\mathbf{w}\| \cdot R \cdot \sqrt{4k} + \sqrt{\hat{\mathcal{L}}_T^{max}(\mathbf{w})}\right)^2.$$

Similar to PERCEPTRON@K-AVG, we can give a simplified mistake bound in situations where the separability condition specified by Definition 4 is satisfied.

Corollary 10. *Suppose a unit norm* $\mathbf{w}^*$ *exists such that the scoring function* $s : \mathbf{x} \mapsto \mathbf{x}^\top \mathbf{w}^*$ *realizes the* strong γ*-margin condition for all the batches, then Algorithm 2 guarantees the mistake bound:* $\Delta_T^C \leq \frac{4kR^2}{\gamma^2}$.

As the *strong* γ-margin condition is exactly the same as the standard notion of margin for binary classification, the above bound is no stronger than the one for the classical perceptron. However, in practice, we observe that PERCEPTRON@K-MAX at times outperforms even PERCEPTRON@K-AVG, even though the latter has a tighter mistake bound. This suggests that our analysis of PERCEPTRON@K-MAX might not be optimal and fails to exploit latent structures that might be present in the data.

Stochastic Gradient Descent for Optimizing prec@k.

We now extend our algorithmic repertoire to include a stochastic gradient descent (SGD) algorithm for the prec@k performance measure. SGD methods are known to be very successful at optimizing large-scale empirical risk minimization (ERM) problems as they require only a few passes over the data to achieve optimal statistical accuracy.

However, SGD methods typically require access to cheap gradient estimates which are difficult to obtain for non-additive performance measures such as prec@k. This has been noticed before by several previous works Kar et al.

[2014], Narasimhan et al. [2015] who propose to use mini-batch methods to overcome this problem Kar et al. [2014]. By combining the $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ surrogate with mini-batch-style processing, we design SGD@K-AVG (Algorithm 3), a scalable SGD algorithm for optimizing prec@k. The algorithm uses mini-batches to update the current model using gradient descent steps. The subgradient calculation for this surrogate turns out to be non-trivial and is detailed in Algorithm 4.

The task of analyzing this algorithm is made non-trivial by the fact that the gradient estimates available to SGD@K-AVG via Algorithm 4 are far from being unbiased. The luxury of having unbiased gradient estimates is crucially exploited by standard SGD analyses but unfortunately, unavailable to us. To overcome this hurdle, we propose a uniform convergence based proof that, in some sense, bounds the bias in the gradient estimates.

In the following section, we present this, and many other generalization and online-to-batch conversion bounds with applications to our perceptron and SGD algorithms.

5 Generalization Bounds

In this section, we discuss novel uniform convergence (UC) bounds for our proposed surrogates. We will use these UC bounds along with the mistake bounds in Theorems 7 and 9 to prove two key results – 1) online-to-batch conversion bounds for the PERCEPTRON@K-AVG and PERCEPTRON@K-MAX algorithms and, 2) a convergence guarantee for the SGD@K-AVG algorithm.

To better present our generalization and convergence bounds, we use normalized versions of prec@k and the surrogates. To do so we write $k = \kappa \cdot n_+$ for some $\kappa \in (0, 1]$ and define, for any scoring function s, its prec@κ loss as:

$$\text{prec@}\kappa(s;\ \mathbf{z}_1, \ldots, \mathbf{z}_n) = \frac{1}{\kappa n_+}\Delta(\mathbf{y}, \mathbf{y}^{(s,\kappa n_+)}).$$

We will also normalize the surrogate functions $\ell^{\text{ramp}}_{\text{prec@}\kappa}(\cdot)$, $\ell^{\text{max}}_{\text{prec@}\kappa}(\cdot)$, and $\ell^{\text{avg}}_{\text{prec@}\kappa}(\cdot)$ by dividing by $k = \kappa \cdot n_+$.

Definition 11 (Uniform Convergence)**.** *A performance measure $\Psi : \mathcal{W} \times (\mathcal{X} \times \{0, 1\})^n \mapsto \mathbb{R}_+$ exhibits uniform convergence with respect to a set of predictors $\mathcal{W}$ if for some $\alpha(b, \delta) = poly\left(\frac{1}{b}, \log\frac{1}{\delta}\right)$, for a sample $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_b$ of size b chosen i.i.d. (or uniformly without replacement) from an arbitrary population $\mathbf{z}_1, \ldots, \mathbf{z}_n$, we have w.p. $1 - \delta$,*

$$\sup_{\mathbf{w}\in\mathcal{W}} |\Psi(\mathbf{w};\ \mathbf{z}_1, \ldots, \mathbf{z}_n) - \Psi(\mathbf{w};\ \hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_b)| \le \alpha(b, \delta)$$

We now state our UC bounds for prec@κ and its surrogates. We refer the reader to Appendix D for proofs.

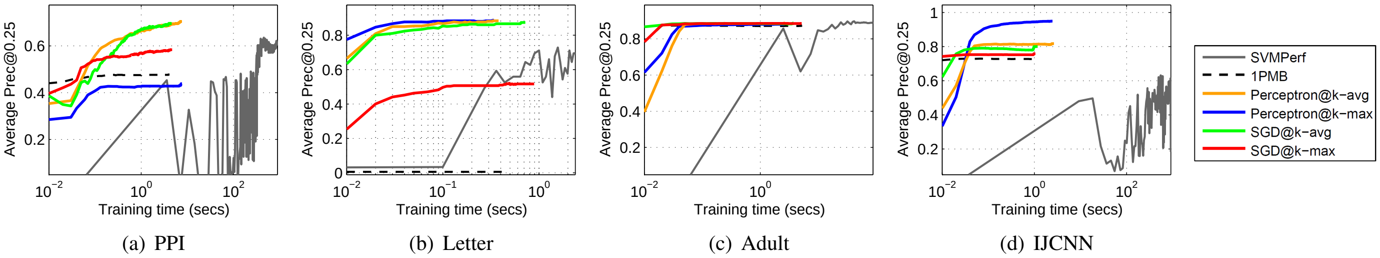
Theorem 12. *The loss function* prec@$\kappa(\cdot)$*, as well as the surrogates $\ell^{ramp}_{\text{prec@}\kappa}(\cdot)$, $\ell^{avg}_{\text{prec@}\kappa}(\cdot)$ and $\ell^{max}_{\text{prec@}\kappa}(\cdot)$, all exhibit uniform convergence at the rate $\alpha(b, \delta) = \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right)$.*

Recently, Kar et al. [2014] also established a similar result for the $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ surrogate. However, a very different proof technique is required to establish similar results for $\ell^{\text{max}}_{\text{prec@}\kappa}(\cdot)$ and $\ell^{\text{avg}}_{\text{prec@}\kappa}(\cdot)$, partly necessitated by the terms in these surrogates which depend, in a complicated manner, on the positives predicted by the candidate labeling $\hat{\mathbf{y}}$. Nevertheless, the above results allow us to establish strong online-to-batch conversion bounds for PERCEPTRON@K-AVG and PERCEPTRON@K-MAX, as well as convergence rates for the SGD@K-AVG method. In the following we shall assume that our data streams are composed of points chosen i.i.d. (or u.w.r.) from some fixed population $\mathcal{Z}$.

Theorem 13. *Suppose an algorithm, when fed a random stream of data points, in T batches of length b each, generates an ensemble of models $\mathbf{w}_1, \ldots, \mathbf{w}_T$ which together suffer a cumulative mistake value of Δ^C_T. Then, with probability at least $1 - \delta$, we have*

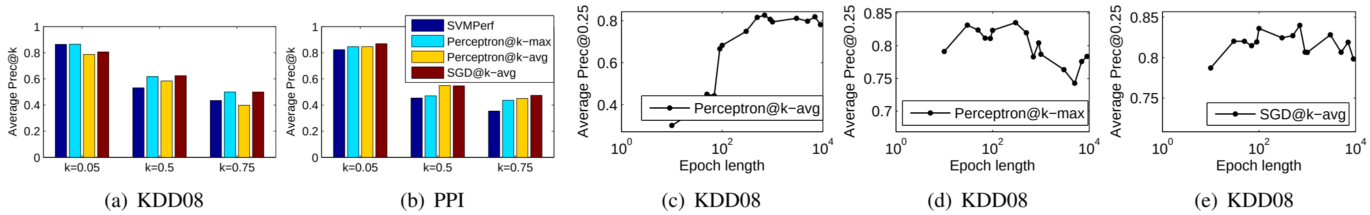
$$\frac{1}{T}\sum_{t=1}^{T} \text{prec@}\kappa(\mathbf{w}^t; \mathcal{Z}) \le \frac{\Delta^C_T}{bT} + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{T}{\delta}}\right).$$

The proof of this theorem follows from Theorem 12 which guarantees that w.p. $1 - \delta$, prec@$\kappa(\mathbf{w}^t; \mathcal{Z}) \le \Delta_t/b + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right)$ for all t. Combining this with the mistake bound from Theorem 7 ensures the following generalization guarantee for the ensemble generated by Algorithm 1.



(a) PPI (b) Letter (c) Adult (d) IJCNN

Figure 2: A comparison of the proposed perceptron and SGD based methods with baseline methods (SVMPerf and **1PMB**) on prec@0.25 maximization tasks. PERCEPTRON@K-AVG and SGD@K-AVG (both based on $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$) are the most consistent methods across tasks.



(a) KDD08 (b) PPI (c) KDD08 (d) KDD08 (e) KDD08

Figure 3: (a), (b): A comparison of different methods on optimizing prec@κ for different values of κ. (c), (d), (e): The performance of the proposed perceptron and SGD methods on prec@0.25 maximization tasks with varying batch lengths b.

Corollary 14. *Let $\mathbf{w}^1, \ldots, \mathbf{w}^T$ be the ensemble of classifiers returned by the* PERCEPTRON@K-AVG *algorithm on a random stream of data points and batch length b. Then, with probability at least $1-\delta$, for any $\mathbf{w}^*$ we have*

$$\frac{1}{T}\sum_{t=1}^{T} \text{prec@}\kappa(\mathbf{w}^t; \mathcal{Z}) \leq \left(\sqrt{\ell^{avg}_{\text{prec@}\kappa}(\mathbf{w}^*; \mathcal{Z})} + C\right)^2,$$

where $C = \mathcal{O}\left(\|\mathbf{w}^*\| R\sqrt{\frac{1}{T}} + \sqrt[4]{\frac{1}{b}\log\frac{T}{\delta}}\right)$.

A similar statement holds for the PERCEPTRON@K-MAX algorithm with respect to the $\ell^{\max}_{\text{prec@}\kappa}(\cdot)$ surrogate as well. Using the results from Theorem 12, we can also establish the convergence rate of the SGD@K-AVG algorithm.

Theorem 15. *Let $\bar{\mathbf{w}}$ be the model returned by Algorithm 3 when executed on a stream with T batches of length b. Then with probability at least $1-\delta$, for any $\mathbf{w}^* \in \mathcal{W}$, we have*

$$\ell^{avg}_{\text{prec@}\kappa}(\bar{\mathbf{w}}; \mathcal{Z}) \leq \ell^{avg}_{\text{prec@}\kappa}(\mathbf{w}^*; \mathcal{Z}) + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{T}{\delta}}\right) + \mathcal{O}\left(\sqrt{\frac{1}{T}}\right)$$

The proof of this Theorem can be found in Appendix E.

6 Experiments

We shall now evaluate our methods on several benchmark datasets for binary classification problems with a rare-class.

Datasets: We evaluated our methods on 7 publicly available benchmark datasets: a) PPI, b) KDD Cup 2008, c) Letter, d) Adult, e) IJCNN, f) Covertype, and g) Cod-RNA. All datasets exhibit moderate to severe label imbalance with the KDD Cup dataset having just 0.61% positives.

Methods: We compared both perceptron algorithms, SGD@K-AVG, as well as an SGD solver for the $\ell^{\max}_{\text{prec@k}}(\cdot)$ surrogate, with the cutting plane-based SVMPerf solver of Joachims [2005]. We also compare against stochastic

1PMB solver of Kar et al. [2014]. The perceptron and SGD methods were given a maximum of 25 passes over the data with a batch length of 500. All methods were implemented in C. We used 70% of the data for training and the rest for testing. All results are averaged over 5 random train-test splits.

Our experiments reveal three interesting insights into the problem of prec@k maximization – 1) using tighter surrogates for optimization routines is indeed beneficial, 2) the presence of a stochastic solver cannot always compensate for the use of a suboptimal surrogate, and 3) mini-batch techniques, applied with perceptron or SGD-style methods, can offer rapid convergence to accurate models.

We first timed all the methods on prec@κ maximization tasks for $\kappa = 0.25$ on various datasets (see Figure 2). Of all the methods, the cutting plane method (SVMPerf) was found to be the most expensive computationally. On the other hand, the perceptron and stochastic gradient methods, which make frequent but cheap updates, were much faster at identifying accurate solutions.

We also observed that PERCEPTRON@K-AVG and SGD@K-AVG, which are based on the tight $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ surrogate, were the most consistent at converging to accurate solutions whereas PERCEPTRON@K-MAX and SGD@K-MAX, which are based on the loose $\ell^{\max}_{\text{prec@k}}(\cdot)$ surrogate, showed large deviations in performance across tasks. Also, **1PMB** and SVMPerf, which are based on the non upper-bounding $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ surrogate, were frequently found to converge to suboptimal solutions.

The effect of working with a tight surrogate is also clear from Figure 3 (a), (b) where the algorithms working with our novel surrogates were found to consistently outperform the SVMPerf method which works with the $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ surrogate. For these experiments, SVMPerf was allowed a runtime of up to $50\times$ of what was given to our methods after which it was terminated.

Finally, to establish the stability of our algorithms, we ran, both the perceptron, as well as the SGD algorithms with varying batch lengths (see Figure 3 (c)-(e)). We found the algorithms to be relatively stable to the setting of the batch length. To put things in perspective, all methods registered a relative variation of less than 5% in accuracies across batch lengths spanning an order of magnitude or more. We present additional experimental results in Appendix F.

Acknowledgments

HN thanks support from a Google India PhD Fellowship.

References

S. Agarwal. The Infinite Push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *11th SIAM International Conference on Data Mining (SDM)*, pages 839–850, 2011.

Stphane Boucheron, Gbor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.

Stephen Boyd, Corinna Cortes, Mehryar Mohri, and Ana Radovanovic. Accuracy at the top. In *26th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 953–961, 2012.

C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *22nd International Conference on Machine Learning (ICML)*, pages 89–96, 2005.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (Non-)existence of Convex, Calibrated Surrogate Losses for Ranking. In *26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *24th International Conference on Machine learning (ICML)*, pages 129–136. ACM, 2007.

Soumen Chakrabarti, Rajiv Khanna, Uma Sawant, and Chiru Bhattacharyya. Structured Learning for Non-Smooth Ranking Losses. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

Sougata Chaudhuri and Ambuj Tewari. Perceptron-like algorithms and generalization bounds for learning to rank. *CoRR*, abs/1405.0591, 2014.

Sougata Chaudhuri and Ambuj Tewari. Online ranking with top-1 feedback. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

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

Chuong B. Do, Quoc Le, Choon Hui Teo, Olivier Chapelle, and Alex Smola. Tighter Bounds for Structured Estimation. In *22nd Annual Conference on Neural Information Processing Systems (NIPS)*, 2008.

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

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM J. Comput.*, 39 (2):742–765, 2009.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

T. Joachims. Optimizing search engines using clickthrough data. In *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

Thorsten Joachims. A Support Vector Method for Multivariate Performance Measures. In *22nd International Conference on Machine Learning (ICML)*, 2005.

Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. Online and stochastic gradient methods for non-decomposable loss functions. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 694–702, 2014.

Quoc V. Le and Alexander J. Smola. Direct optimization of ranking measures. *arXiv preprint arXiv:0704.3359*, 2007.

Nan Li, Rong Jin, and Zhi-Hua Zhou. Top rank optimization in linear time. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1502–1510, 2014.

Marvin Lee Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1988. ISBN 0262631113.

Harikrishna Narasimhan and Shivani Agarwal. A Structural SVM Based Approach for Optimizing Partial AUC. In *30th International Conference on Machine Learning (ICML)*, 2013a.

Harikrishna Narasimhan and Shivani Agarwal. $\text{SVM}_{\text{pAUC}}^{\text{tight}}$: A New Support Vector Method for Optimizing Partial AUC Based on a Tight Convex Upper Bound. In *19th ACM SIGKDD Conference on Knowledge, Discovery and Data Mining (KDD)*, 2013b.

Harikrishna Narasimhan, Purushottam Kar, and Prateek Jain. Optimizing Non-decomposable Performance Measures: A Tale of Two Classes. In *32nd International Conference on Machine Learning (ICML)*, 2015.

A.B.J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622, 1962.

Yashoteja Prabhu and Manik Varma. Fastxml: a fast, accurate and stable tree-classifier for extreme multi-label learning. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*, pages 263–272, 2014.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

C. Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *Journal of Machine Learning Research*, 10:2233–2271, 2009.

Grigorios Tsoumakas and Ioannis Katakis. Multi-Label Classification: An Overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. Learning to rank by optimizing NDCG measure. In *26th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1883–1891, 2009.

Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 271–278, 2007.

Hyokun Yun, Parameswaran Raman, and S Vishwanathan. Ranking via robust binary classification. In *28th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2582–2590, 2014.

Tong Zhang. Covering Number Bounds of Certain Regularized Linear Function Classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

A Structural SVM Surrogate for prec@k

The structural SVM surrogate for prec@k for a set of n points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \in (\mathbb{R}^d \times \{0,1\})^n$ and model $w \in \mathbb{R}^d$ can be written as $\ell^{\text{struct}}_{\text{prec@k}}(w)$:

$$\max_{\substack{\widehat{\mathbf{y}} \in \{0,1\}^n \\ \|\widehat{\mathbf{y}}\|_1 = k}} \left\{ 1 + \sum_{i=1}^{n} \widehat{\mathbf{y}}_i \left(\frac{1}{n} \mathbf{w}^\top \mathbf{x}_i - \frac{1}{k} \mathbf{y}_i \right) \; - \; \frac{1}{n} \sum_{i=1}^{n} \mathbf{y}_i \mathbf{w}^\top \mathbf{x}_i \right\}.$$

We shall now give a simple setting where this surrogate produces a suboptimal model.

Consider a set of 6 points in $\mathbb{R} \times \{0,1\}$: $\{(-1,1), (-1,1), (-2,1), (-3,0), (-3,0), (-3,0)\}$, and suppose we are interested in Prec@1. Note that the optimum model that maximizes prec@1 on these points has a positive sign. We will now show that the model $w^* \in \mathbb{R}$ that maximizes the above structural SVM surrogate on these points has a negative sign. On the contrary, let us assume that w^* has a positive sign, and arrive at a contradiction; we shall consider the following two cases:

(i) $w^* > \frac{3}{2}$. It can be verified that

$$\begin{aligned} \ell^{\text{struct}}_{\text{prec@k}}(w^*) &= 1 + \left(\frac{1}{6}(-w^*) - 1 \right) - \frac{1}{6}(-w^* + -w^* + -2w^*) \\ &= \frac{1}{2} w^* \end{aligned}$$

On the other hand, for the model $w' = -w^*$, we have

$$\begin{aligned} \ell^{\text{struct}}_{\text{prec@k}}(w') &= 1 + \left(\frac{1}{6}(-3w') - 0 \right) - \frac{1}{6}(-w' + -w' + -2w') \\ &= 1 + \left(\frac{1}{6}(3w^*) - 0 \right) - \frac{1}{6}(w^* + w^* + 2w^*) \\ &= 1 - \frac{1}{6} w^* \; < \; \ell^{\text{struct}}_{\text{prec@k}}(w^*), \end{aligned}$$

where the last step follows from $w^* > \frac{3}{2}$; clearly, w^* is not optimal for the structural SVM surrogate, and hence a contradiction.

(i) $w^* \leq \frac{3}{2}$. Here we have

$$\begin{aligned} \ell^{\text{struct}}_{\text{prec@k}}(w^*) &= 1 + \left(\frac{1}{6}(-3w^*) - 0 \right) - \frac{1}{6}(-w^* + -w^* + -2w^*) \\ &= 1 + \frac{1}{6} w^*. \end{aligned}$$

For $w' = -w^*$,

$$\begin{aligned} \ell^{\text{struct}}_{\text{prec@k}}(w') &= 1 + \left(\frac{1}{6}(-3w') - 0 \right) - \frac{1}{6}(-w' + -w' + -2w') \\ &= 1 + \left(\frac{1}{6}(3w^*) - 0 \right) - \frac{1}{6}(w^* + w^* + 2w^*) \\ &= 1 - \frac{1}{6} w^* \; < \; \ell^{\text{struct}}_{\text{prec@k}}(w^*). \end{aligned}$$

Here again, we have a contradiction. Notice that this surrogate can take negative values (when $w < -6$ for example) whereas prec@k is a positive valued function. This clearly indicates that this surrogate cannot upper bound prec@k. More specifically, notice that for $w < 0$, we have $\text{prec@k}(w) = 1$, however, the above analysis demonstrates cases when $\ell^{\text{struct}}_{\text{prec@k}}(w) < 1$ which gives an explicit example that this surrogate is not even an upper bounding surrogate.

B Proofs of Claims from Section 3

B.1 Proof of Claim 1

Claim 1. *For any $k \leq n_+$ and scoring function s, we have*

$$\ell^{\textit{ramp}}_{\text{prec@k}}(s) \geq \text{prec@k}(s).$$

Moreover, if for some scoring function s, we have $\ell^{\textit{ramp}}_{\text{prec@k}}(s) \leq \xi$, then there necessarily exists a set $S \subset [n]$ of size at most k such that for all $\|\hat{\mathbf{y}}\| = k$, we have

$$\sum_{i \in S} s_i \geq \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i + \Delta(\mathbf{y}, \hat{\mathbf{y}}) - \xi.$$

Proof. Let $\hat{\mathbf{y}} = \mathbf{y}^{(s,k)}$ so that we have $\Delta(\mathbf{y}, \hat{\mathbf{y}}) = \text{prec@k}(s)$. Then we have

$$\begin{aligned}
\ell^{\text{ramp}}_{\text{prec@k}}(s) &= \max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i \right\} - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \\
&\geq \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \\
&= \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \max_{\|\tilde{\mathbf{y}}\|_1 = k} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \\
&\geq \Delta(\mathbf{y}, \hat{\mathbf{y}}),
\end{aligned}$$

where the third step follows from the definition of $\hat{\mathbf{y}}$. This proves the first claim. For the second claim, suppose for some scoring function s, we have $\ell^{\text{ramp}}_{\text{prec@k}}(s) \leq \xi$. Then if we consider S^* to be the set of k-highest ranked positive points, then we have

$$\sum_{i \in S^*} s_i = \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \geq \max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i \right\} - \xi \geq \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i + \Delta(\mathbf{y}, \hat{\mathbf{y}}) - \xi,$$

which proves the claim. □

B.2 Proof of Claim 3

Claim 3. *For any scoring function s that realizes the* weak *k-margin over a dataset we have,*

$$\ell^{\textit{ramp}}_{\text{prec@k}}(s) = \text{prec@k}(s) = 0.$$

Proof. Consider a scoring function s that satisfies the weak k-margin condition and any $\hat{\mathbf{y}}$ such that $\|\hat{\mathbf{y}}\|_1 = k$. Based on the prec@k accuracy of $\hat{\mathbf{y}}$, we have the following two cases

Case 1 ($K(\mathbf{y}, \hat{\mathbf{y}}) = k$): In this case we have

$$\Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i = 0 + \sum_{i=1}^{n} \hat{\mathbf{y}}_i s_i - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \leq 0,$$

where the first step follows since $K(\mathbf{y}, \hat{\mathbf{y}}) = k$ and the second step follows since $\|\hat{\mathbf{y}}\|_1 = k$, as well as $K(\mathbf{y}, \hat{\mathbf{y}}) = k$.

Case 2 ($K(\mathbf{y}, \hat{\mathbf{y}}) = k' < k$): In this case let S^* be the set of k top ranked positive points according to the scoring function s. Also let S_1^* be the set of $k' (= K(\mathbf{y}, \hat{\mathbf{y}}))$ top ranked positives and let $S_2^* = S^* \backslash S_1^*$. Then we have

$$\begin{aligned}
\Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^n \hat{\mathbf{y}}_i s_i - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^n \tilde{\mathbf{y}}_i s_i &= \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \underbrace{\sum_{i=1}^n \hat{\mathbf{y}}_i \mathbf{y}_i s_i}_{(A)} + \sum_{i=1}^n \hat{\mathbf{y}}_i (1 - \mathbf{y}_i) s_i - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^n \tilde{\mathbf{y}}_i s_i \\
&\leq \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i \in S_1^*} s_i + \underbrace{\sum_{i=1}^n \hat{\mathbf{y}}_i (1 - \mathbf{y}_i) s_i}_{(B)} - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^n \tilde{\mathbf{y}}_i s_i \\
&\leq \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i \in S_1^*} s_i + \sum_{i \in S_2^*} s_i - (k - k') - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^n \tilde{\mathbf{y}}_i s_i \\
&= k - k' + \sum_{i \in S^*} s_i - (k - k') - \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^n \tilde{\mathbf{y}}_i s_i \\
&= 0,
\end{aligned}$$

where the second step follows since the term (A) consists of k' true positives the third step follows since the term (B) contains $k - k'$ false positives i.e. negatives and the k-margin condition, the fourth step follows since $\Delta(\mathbf{y}, \hat{\mathbf{y}}) = k - K(\mathbf{y}, \hat{\mathbf{y}})$ and the fifth step follows since by the definition of the set S^*, we have

$$\sum_{i \in S^*} s_i = \max_{\substack{\|\tilde{\mathbf{y}}\|_1 = k \\ K(\mathbf{y}, \tilde{\mathbf{y}}) = k}} \sum_{i=1}^n \tilde{\mathbf{y}}_i s_i.$$

In both cases, we have shown the surrogate to be non-positive. Since the performance measure prec@k cannot take negative values, this, along with the upper bounding property implies that prec@k$(s) = 0$ as well. This finishes the proof. □

B.3 A Useful Supplementary Lemma

Lemma 16. *Given a set of n real numbers $x_1 \ldots x_n$ and any two integers $k \leq k' \leq n$, we have*

$$\min_{|S| = k} \frac{1}{k} \sum_{i \in S} x_i \leq \min_{|S'| = k'} \frac{1}{k'} \sum_{j \in S'} x_j$$

Proof. The above is obviously true if $k = k'$ so we assume that $k' > k$. Without loss of generality assume that the set is ordered in ascending order i.e. $x_1 \leq x_2 \leq \ldots \leq x_n$. Thus, the above statement is equivalent to showing that

$$\frac{1}{k} \sum_{i=1}^k x_i \leq \frac{1}{k'} \sum_{j=1}^{k'} x_j \Leftrightarrow \left(\frac{1}{k} - \frac{1}{k'} \right) \sum_{i=1}^k x_i \leq \frac{1}{k'} \sum_{j=k+1}^{k'} x_j \Leftrightarrow \frac{1}{k} \sum_{i=1}^k x_i \leq \frac{1}{k' - k} \sum_{j=k+1}^{k'} x_j,$$

where the last inequality is true since $k - k' > 0$ and the left hand side is the average of numbers which are all smaller than the numbers whose average forms the right hand side. This proves the lemma. □

B.4 Proof of the Upper-bounding Property for the $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ Surrogate

Claim 17. *For any $k \leq n_+$ and scoring function s, we have*

$$\ell^{avg}_{\text{prec@k}}(s) \geq \text{prec@k}(s).$$

Moreover, for linear scoring functions i.e. $s(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i$ for $\mathbf{w} \in \mathcal{W}$, the surrogate $\ell^{avg}_{\text{prec@k}}(\mathbf{w})$ is convex in $\mathbf{w}$.

Proof. We use the fact observed before that for any scoring function, we have $\Delta(\mathbf{y}, \mathbf{y}^{(s,k)}) = \text{prec@k}(s)$. We start off by showing the second part of the claim. Recall the definition of the surrogate $\ell^{\text{avg}}_{\text{prec@k}}(s)$

$$\ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}) = \max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \mathbf{y}_i) \cdot \mathbf{w}^\top \mathbf{x}_i + \frac{1}{C(\hat{\mathbf{y}})} \sum_{i=1}^{n} (1 - \hat{\mathbf{y}}_i) \mathbf{y}_i \cdot \mathbf{w}^\top \mathbf{x}_i \right\}$$

For sake of simplicity, for any $\hat{\mathbf{y}} \in \{0,1\}^n$, define

$$\Delta(s, \hat{\mathbf{y}}) = \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} s_i (\hat{\mathbf{y}}_i - \mathbf{y}_i) + \frac{1}{C(\hat{\mathbf{y}})} \sum_{i=1}^{n} (1 - \hat{\mathbf{y}}_i) \mathbf{y}_i s_i.$$

The convexity of $\ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w})$ follows from the observation that the inner term in the maximization is linear (hence convex) in $\mathbf{w}$ and the $\max$ function is convex and increasing. We now move on to prove the first part. For sake of convenience $\tilde{\mathbf{y}} = \mathbf{y}^{(s,k)}$. Note that $\|\tilde{\mathbf{y}}\|_1 = k$ by definition. This gives us

$$\begin{aligned}
\ell^{\text{avg}}_{\text{prec@k}}(s) &= \max_{\|\hat{\mathbf{y}}\|_1 = k} \Delta(s, \hat{\mathbf{y}}) \geq \Delta(s, \tilde{\mathbf{y}}) \\
&= \Delta(\mathbf{y}, \tilde{\mathbf{y}}) + \sum_{i=1}^{n} s_i (\tilde{\mathbf{y}}_i - \mathbf{y}_i) + \frac{1}{C(\tilde{\mathbf{y}})} \sum_{i=1}^{n} (1 - \tilde{\mathbf{y}}_i) \mathbf{y}_i s_i \\
&= \Delta(\mathbf{y}, \tilde{\mathbf{y}}) + \sum_{i=1}^{n} s_i (\tilde{\mathbf{y}}_i (1 - \mathbf{y}_i) - \mathbf{y}_i (1 - \tilde{\mathbf{y}}_i)) + \frac{n_+ - k}{n_+ - K(\mathbf{y}, \tilde{\mathbf{y}})} \sum_{i=1}^{n} (1 - \tilde{\mathbf{y}}_i) \mathbf{y}_i s_i \\
&= \Delta(\mathbf{y}, \tilde{\mathbf{y}}) + \underbrace{\sum_{i=1}^{n} \tilde{\mathbf{y}}_i (1 - \mathbf{y}_i) s_i}_{(A)} - \underbrace{\frac{k - K(\mathbf{y}, \tilde{\mathbf{y}})}{n_+ - K(\mathbf{y}, \tilde{\mathbf{y}})} \sum_{i=1}^{n} (1 - \tilde{\mathbf{y}}_i) \mathbf{y}_i s_i}_{(B)}.
\end{aligned}$$

Now define $m = \min_{\substack{\tilde{\mathbf{y}}_i = 1 \\ \mathbf{y}_i = 0}} s_i$ and $M = \max_{\substack{\tilde{\mathbf{y}}_i = 0 \\ \mathbf{y}_i = 1}} s_i$. This gives us

$$(A) = \sum_{i=1}^{n} \tilde{\mathbf{y}}_i (1 - \mathbf{y}_i) s_i \geq m \sum_{i=1}^{n} \tilde{\mathbf{y}}_i (1 - \mathbf{y}_i) = \Delta(\mathbf{y}, \tilde{\mathbf{y}}) \cdot m,$$

and

$$(B) = \frac{k - K(\mathbf{y}, \tilde{\mathbf{y}})}{n_+ - K(\mathbf{y}, \tilde{\mathbf{y}})} \sum_{i=1}^{n} (1 - \tilde{\mathbf{y}}_i) \mathbf{y}_i s_i \leq \frac{k - K(\mathbf{y}, \tilde{\mathbf{y}})}{n_+ - K(\mathbf{y}, \tilde{\mathbf{y}})} \sum_{i=1}^{n} (1 - \tilde{\mathbf{y}}_i) \mathbf{y}_i M = (k - K(\mathbf{y}, \tilde{\mathbf{y}})) \cdot M = \Delta(\mathbf{y}, \tilde{\mathbf{y}}) \cdot M.$$

However, by definition of $\tilde{\mathbf{y}} = \mathbf{y}^{(s,k)}$, we have

$$m \geq \min_{\tilde{\mathbf{y}} = 1} s_i \geq \max_{\tilde{\mathbf{y}} = 0} s_i \geq M.$$

Thus we have

$$\ell^{\text{avg}}_{\text{prec@k}}(s) \geq \Delta(\mathbf{y}, \tilde{\mathbf{y}}) + (A) - (B) \geq \Delta(\mathbf{y}, \tilde{\mathbf{y}})(1 + m - M) \geq \Delta(\mathbf{y}, \tilde{\mathbf{y}}) = \text{prec@k}(s)$$ □

B.5 Proof of Claim 6

Claim 6. *For any scoring function s that realizes the k-margin over a dataset we have,*

$$\ell^{\mathit{avg}}_{\text{prec@k}}(s) = \text{prec@k}(s) = 0.$$

Proof. We shall prove that for any $\hat{\mathbf{y}}$ such that $\|\hat{\mathbf{y}}\|_1 = k$, under the k-margin condition, we have $\Delta(s, \hat{\mathbf{y}}) = 0$. This will show us that $\ell^{\text{avg}}_{\text{prec@k}}(s) = \max_{\|\hat{\mathbf{y}}\|_1 = k} \Delta(s, \hat{\mathbf{y}}) = 0$. Using Claim 17 and the fact that $\text{prec@k}(s) \geq 0$ will then prove the claimed result. We will analyze two cases in order to do this

Case 1 ($K(\mathbf{y}, \hat{\mathbf{y}}) = k$): In this case the labeling $\hat{\mathbf{y}}$ is able to identify k relevant points correctly and thus we have $C(\hat{\mathbf{y}}) = 1$ and we have

$$\Delta(s, \hat{\mathbf{y}}) = \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \sum_{i=1}^{n}(1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i$$

Now, since $K(\mathbf{y}, \hat{\mathbf{y}}) = k$, we have $\Delta(\mathbf{y}, \hat{\mathbf{y}}) = 0$ which means for all i such that $\hat{\mathbf{y}}_i = 1$, we also have $\mathbf{y}_i = 1$. Thus, we have $\hat{\mathbf{y}}_i = \hat{\mathbf{y}}_i \mathbf{y}_i$. Thus,

$$\Delta(s, \hat{\mathbf{y}}) = 0 + \sum_{i=1}^{n} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i\mathbf{y}_i)s_i = \sum_{i=1}^{n} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)s_i = 0$$

Case 2 ($K(\mathbf{y}, \hat{\mathbf{y}}) = k' < k$): In this case, $\hat{\mathbf{y}}$ contains false positives. Thus we have

$$\begin{aligned}
\Delta(s, \hat{\mathbf{y}}) &= \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \frac{n_+ - k}{n_+ - k'} \sum_{i=1}^{n}(1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} \hat{\mathbf{y}}_i(1 - \mathbf{y}_i)s_i - \frac{k - k'}{n_+ - k'} \sum_{i=1}^{n} \mathbf{y}_i(1 - \hat{\mathbf{y}}_i)s_i \\
&= (k - k')\left(\underbrace{\frac{1}{k - k'}\Delta(\mathbf{y}, \hat{\mathbf{y}})}_{(A)} + \underbrace{\frac{1}{k - k'}\sum_{i=1}^{n} \hat{\mathbf{y}}_i(1 - \mathbf{y}_i)s_i}_{(B)} - \underbrace{\frac{1}{n_+ - k'}\sum_{i=1}^{n} \mathbf{y}_i(1 - \hat{\mathbf{y}}_i)s_i}_{(C)} \right)
\end{aligned}$$

Now we have, by definition, $(A) = 1$. We also have

$$(B) = \frac{1}{k - k'}\sum_{i=1}^{n} \hat{\mathbf{y}}_i(1 - \mathbf{y}_i)s_i \leq \max_{j : \mathbf{y}_j = 0} s_j,$$

as well as

$$\begin{aligned}
(C) &= \frac{1}{n_+ - k'}\sum_{i=1}^{n} \mathbf{y}_i(1 - \hat{\mathbf{y}}_i)s_i \\
&\geq \min_{\substack{S_+ \subseteq \mathbf{X}_+ \\ |S_+| = n_+ - k'}} \frac{1}{n_+ - k'} \sum_{i \in S_+} \mathbf{y}_i(1 - \hat{\mathbf{y}}_i)s_i \\
&\geq \min_{\substack{S_+ \subseteq \mathbf{X}_+ \\ |S_+| = n_+ - k + 1}} \frac{1}{n_+ - k + 1} \sum_{i \in S_+} \mathbf{y}_i(1 - \hat{\mathbf{y}}_i)s_i,
\end{aligned}$$

where the last step follows from Lemma 16 and the fact that $k' \leq k - 1$ in this case analysis. Then we have

$$\Delta(s, \hat{\mathbf{y}}) = (k - k')((A) + (B) - (C)) \leq (k - k')\left(1 + \max_{j : \mathbf{y}_j = 0} s_j - \min_{\substack{S_+ \subseteq \mathbf{X}_+ \\ |S_+| = n_+ - k + 1}} \frac{1}{n_+ - k + 1}\sum_{i \in S_+} \mathbf{y}_i(1 - \hat{\mathbf{y}}_i)s_i\right) \leq 0$$

where the last step follows because s realizes the k-margin. Having exhausted all cases, we establish the claim. □

C Proofs from Section 4

C.1 Proof of Theorem 7

Theorem 7. *Suppose $\left\|\mathbf{x}_t^i\right\| \leq R$ for all t, i. Let $\Delta_T^C = \sum_{t=1}^T \Delta_t$ be the cumulative observed mistake values when Algorithm 1 is run. Also, for any predictor $\mathbf{w}$, let $\hat{\mathcal{L}}_T(\mathbf{w}) = \sum_{t=1}^T \ell_{\text{prec@k}}^{avg}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)$. Then we have*

$$\Delta_T^C \leq \min_{\mathbf{w}} \left(\|\mathbf{w}\| \cdot R \cdot \sqrt{4k} + \sqrt{\hat{\mathcal{L}}_T(\mathbf{w})} \right)^2.$$

Proof. We will prove the theorem using two lemmata that we state below.

Lemma 18. *For any time step t, we have*

$$\|\mathbf{w}_t\|^2 \leq \|\mathbf{w}_{t-1}\|^2 + 4kR^2 \Delta_t$$

Lemma 19. *For any fixed $\mathbf{w} \in \mathcal{W}$, define $P_t := \langle \mathbf{w}_t, \mathbf{w} \rangle$. Then we have*

$$P_t \geq P_{t-1} + \Delta_t - \ell_{\text{prec@k}}^{avg}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t).$$

Using Lemmata 18 and 19, we can establish the mistake bound as follows. A repeated application of Lemma 19 tells us that

$$P_T \geq \sum_{t=1}^T \Delta_t - \sum_{t=1}^T \ell_{\text{prec@k}}^{\text{avg}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t) = \Delta_t^C - \hat{\mathcal{L}}_T(\mathbf{w}).$$

In case the right hand side is negative, we already have the result with us. In case it is positive, we can now analyze further using the Cauchy-Schwartz inequality, and a repeated application of Lemma 18. Starting from the above we have

$$\begin{aligned} \Delta_T^C &\leq P_T + \hat{\mathcal{L}}_T(\mathbf{w}) \\ &= \langle \mathbf{w}_T, \mathbf{w} \rangle + \hat{\mathcal{L}}_T(\mathbf{w}) \\ &\leq \|\mathbf{w}_T\| \, \|\mathbf{w}\| + \hat{\mathcal{L}}_T(\mathbf{w}) \\ &\leq \|\mathbf{w}\| \sqrt{4kR^2 \cdot \Delta_T^C} + \hat{\mathcal{L}}_T(\mathbf{w}), \end{aligned}$$

which gives us the desired result upon solving the quadratic inequality. We now prove the lemmata below. Note that in the following discussion, we have, for sake of brevity, used the notation $\hat{\mathbf{y}} = \hat{\mathbf{y}}_t = \mathbf{y}^{(\mathbf{w}_{t-1}, k)}$.

Proof of Lemma 18. For time steps where $\Delta_t = 0$, the result obviously holds since $\mathbf{w}_t = \mathbf{w}_{t-1}$. For analyzing other time steps, let $\mathbf{v}_t = D_t \cdot \sum_{i \in [b]} (1 - \hat{\mathbf{y}}_i) \mathbf{y}_i \cdot \mathbf{x}_t^i - \sum_{i \in [b]} (1 - \mathbf{y}_i) \hat{\mathbf{y}}_i \cdot \mathbf{x}_t^i$ so that $\mathbf{w}_t = \mathbf{w}_{t-1} + \mathbf{v}_t$. This gives us

$$\|\mathbf{w}_t\|^2 = \|\mathbf{w}_{t-1}\|^2 + 2 \langle \mathbf{w}_{t-1}, \mathbf{v}_t \rangle + \|\mathbf{v}_t\|^2.$$

Let $s_i = \mathbf{w}_{t-1}^\top \mathbf{x}_t^i$. Then we have

$$\begin{aligned} \langle \mathbf{w}_{t-1}, \mathbf{v}_t \rangle &= D_t \cdot \sum_{i \in [b]} (1 - \hat{\mathbf{y}}_i) \mathbf{y}_i s_i - \sum_{i \in [b]} (1 - \mathbf{y}_i) \hat{\mathbf{y}}_i s_i \\ &= \Delta_t \left(\underbrace{\frac{1}{\|\mathbf{y}_t\|_1 - K(\mathbf{y}_t, \hat{\mathbf{y}}_t)} \sum_{i \in [b]} (1 - \hat{\mathbf{y}}_i) \mathbf{y}_i s_i}_{(A)} - \underbrace{\frac{1}{\Delta_t} \sum_{i \in [b]} (1 - \mathbf{y}_i) \hat{\mathbf{y}}_i s_i}_{(B)} \right) \end{aligned}$$

More specifically, we use the fact that the inequality $(x - l)^2 \leq cx$ has a solution $x \leq (\sqrt{l} + \sqrt{c})^2$ whenever $x, l, c \geq 0$ and $x \geq l$.

$$\leq \quad 0,$$

where the last step follows since (A) is the average of scores given to the false negatives and (B) is the average of scores given to the false positives and by the definition of $\hat{\mathbf{y}}_t$, since false negatives are assigned scores less than false positives, we have $(A) \leq (B)$. We also have

$$\begin{aligned}
\|\mathbf{v}_t\|^2 &= \Delta_t^2 \left\| \frac{1}{\|\mathbf{y}_t\|_1 - K(\mathbf{y}_t, \hat{\mathbf{y}}_t)} \cdot \sum_{i\in[b]} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i \cdot \mathbf{x}_t^i - \frac{1}{\Delta_t} \sum_{i\in[b]} (1-\mathbf{y}_i)\hat{\mathbf{y}}_i \cdot \mathbf{x}_t^i \right\|^2 \\
&\leq 4\Delta_t^2 R^2 \leq 4kR^2\Delta_t,
\end{aligned}$$

since $\Delta_t \leq k$. Combining the two gives us the desired result. □

Proof of Lemma 19. We prove the result using two cases. For sake of convenience, we will refer to $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t$ as $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively.

Case 1 ($\Delta_t = 0$): In this case $P_t = P_{t-1}$ since the model is not updated. However, since $\ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}) \geq \text{prec@k}(\mathbf{w}) \geq 0$ for all $\mathbf{w} \in \mathcal{W}$ (by Claim 17), we still get

$$P_t \geq P_{t-1} - \ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t),$$

as required.

Case 2 ($\Delta_t > 0$): In this case we use the update to $\mathbf{w}_{t-1}$ to evaluate the update to P_{t-1}. For sake of convenience, let us use the notation $s_i = \mathbf{w}^\top \mathbf{x}_t^i$. Also note that in Algorithm 1, $D_t = 1 - \frac{1}{C(\hat{\mathbf{y}})}$.

$$\begin{aligned}
P_t &= P_{t-1} - \sum_{i\in[b]} (1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i + D_t \cdot \sum_{i\in[b]} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \sum_{i\in[b]} (1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i + \left(1 - \frac{1}{C(\hat{\mathbf{y}})}\right) \sum_{i\in[b]} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \underbrace{\left(\sum_{i\in[b]} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \frac{1}{C(\hat{\mathbf{y}})} \sum_{i\in[b]} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \right)}_{(Q)} \\
&\geq P_{t-1} + \Delta_t - \ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t),
\end{aligned}$$

where the last step follows from the definition of $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ which gives us

$$\begin{aligned}
\Delta_t + (Q) &= \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i\in[b]} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \frac{1}{C(\hat{\mathbf{y}})} \sum_{i\in[b]} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&\leq \max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i\in[b]} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \frac{1}{C(\hat{\mathbf{y}})} \sum_{i\in[b]} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \right\} \\
&= \ell^{\text{avg}}_{\text{prec@k}}(s) = \ell^{\text{avg}}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)
\end{aligned}$$

□

This concludes the proof of the mistake bound. □

C.2 Proof of Theorem 9

Theorem 9. *Suppose $\|\mathbf{x}_t^i\| \leq R$ for all t, i. Let $\Delta_T^C = \sum_{t=1}^T \Delta_t$ be the cumulative observed mistake values when Algorithm 2 is run. Also, for any predictor $\mathbf{w}$, let $\hat{\mathcal{L}}_T^{max}(\mathbf{w}) = \sum_{t=1}^T \ell^{max}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)$. Then we have*

$$\Delta_T^C \leq \min_{\mathbf{w}} \left(\|\mathbf{w}\| \cdot R \cdot \sqrt{4k} + \sqrt{\hat{\mathcal{L}}_T^{max}(\mathbf{w})} \right)^2.$$

Proof. As before, we will prove this theorem in two parts. Lemma 18 will continue to hold in this case as well. However, we will need a modified form of Lemma 19 that we prove below. As before, we will use the notation $\hat{\mathbf{y}} = \hat{\mathbf{y}}_t = \mathbf{y}^{(\mathbf{w}_{t-1},k)}$.

Lemma 20. *For any fixed* $\mathbf{w} \in \mathcal{W}$*, define* $P_t := \langle \mathbf{w}_t, \mathbf{w} \rangle$*. Then we have*

$$P_t \geq P_{t-1} + \Delta_t - \ell^{max}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t).$$

Using Lemmata 18 and 20, the theorem follows as before. All that remains now is to prove Lemma 20.

Proof of Lemma 20. We prove the result using two cases as before. For sake of convenience, we will refer to $\mathbf{y}_t$ and $\hat{\mathbf{y}}_t$ as $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively.

Case 1 ($\Delta_t = 0$): In this case $P_t = P_{t-1}$ since the model is not updated. However, since $\ell^{\text{max}}_{\text{prec@k}}(\mathbf{w}) \geq \text{prec@k}(\mathbf{w}) \geq 0$ for all $\mathbf{w} \in \mathcal{W}$ (by Claim 1), we still get

$$P_t \geq P_{t-1} - \ell^{\text{max}}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t),$$

as required.

Case 2 ($\Delta_t > 0$): In this case we use the update to $\mathbf{w}_{t-1}$ to evaluate the update to P_{t-1}. For sake of convenience, let us use the notation $s_i = \mathbf{w}^\top \mathbf{x}_t^i$. Also note that the set $S_t := \text{FN}(\mathbf{w}^{t-1}, \Delta_t)$ contains the false negatives in the top Δ_t positions as ranked by $\mathbf{w}^{t-1}$.

$$\begin{aligned}
P_t &= P_{t-1} - \sum_{i \in [b]} (1 - \mathbf{y}_i)\hat{\mathbf{y}}_i s_i + \sum_{i \in S_t} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \sum_{i \in [b]} (1 - \mathbf{y}_i)\hat{\mathbf{y}}_i s_i - \sum_{i \in [b]} \mathbf{y}_i\hat{\mathbf{y}}_i s_i + \sum_{i \in [b]} \mathbf{y}_i\hat{\mathbf{y}}_i s_i + \sum_{i \in S_t} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \sum_{i \in [b]} \hat{\mathbf{y}}_i s_i + \sum_{i \in [b]} \mathbf{y}_i\hat{\mathbf{y}}_i s_i + \sum_{i \in S_t} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \left(\sum_{i \in [b]} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \sum_{i \in [b]} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i - \sum_{i \in S_t} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i s_i \right) \\
&\geq P_{t-1} - \underbrace{\left(\sum_{i \in [b]} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \max_{\substack{\tilde{\mathbf{y}} \preceq (1-\hat{\mathbf{y}})\cdot\mathbf{y} \\ \|\tilde{\mathbf{y}}\|_1 = n_+ - k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \right)}_{(Q)} \\
&\geq P_{t-1} + \Delta_t - \ell^{\text{max}}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t),
\end{aligned}$$

where the last step follows from the definition of $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ which gives us

$$\begin{aligned}
\Delta_t + (Q) &= \Delta_t + \sum_{i \in [b]} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \max_{\substack{\tilde{\mathbf{y}} \preceq (1-\hat{\mathbf{y}})\cdot\mathbf{y} \\ \|\tilde{\mathbf{y}}\|_1 = n_+ - k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \\
&\leq \max_{\|\hat{\mathbf{y}}\|_1 = k} \left\{ \Delta_t + \sum_{i \in [b]} (\hat{\mathbf{y}}_i - \mathbf{y}_i) s_i + \max_{\substack{\tilde{\mathbf{y}} \preceq (1-\hat{\mathbf{y}})\cdot\mathbf{y} \\ \|\tilde{\mathbf{y}}\|_1 = n_+ - k}} \sum_{i=1}^{n} \tilde{\mathbf{y}}_i s_i \right\} \\
&= \ell^{\text{max}}_{\text{prec@k}}(s) = \ell^{\text{max}}_{\text{prec@k}}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)
\end{aligned}$$

□

This conclude the proof of the theorem. □

D Proof of Theorem 12

Our proof of Theorem 12 crucially utilizes the following two lemmas that helps in exploiting the structure in our surrogate functions. The first basic lemma states that the pointwise supremum of a set of Lipschitz functions is also Lipschitz.

Lemma 21. *Let $f_1, \ldots, f_m$ be m real valued functions $f_i : \mathbb{R}^n \to \mathbb{R}$ such that every f_i is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm. Then the function*

$$g(\mathbf{v}) = \max_{i \in [m]} \; f_i(\mathbf{v})$$

is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm too.

The second lemma establishes the convergence of additive estimates over the top of ranked lists. The abstract nature of the result would allow us to apply it to a wide variety of situations and would be crucial to our analyses.

Lemma 22. *Let $\mathcal{V}$ be a universe with a total order $\succeq$ established on it and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a population of n items arranged in decreasing order. Let $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_b$ be a sample chosen i.i.d. (or without replacement) from the population and arranged in decreasing order as well. Then for any fixed $h : \mathcal{V} \to [-1, 1]$ and $\kappa \in (0, 1]$, we have, with probability at least $1 - \delta$ over the choice of the samples,*

$$\left| \frac{1}{\lceil \kappa n \rceil} \sum_{i=1}^{\lceil \kappa n \rceil} h(\mathbf{v}_i) - \frac{1}{\lceil \kappa b \rceil} \sum_{i=1}^{\lceil \kappa b \rceil} h(\hat{\mathbf{v}}_i) \right| \leq 4 \sqrt{\frac{\log \frac{2}{\delta}}{\kappa b}}$$

Theorem 12. *The performance measure* $\text{prec@}\kappa(\cdot)$*, as well as the surrogates $\ell^{ramp}_{\text{prec@}\kappa}(\cdot)$, $\ell^{avg}_{\text{prec@}\kappa}(\cdot)$ and $\ell^{max}_{\text{prec@}\kappa}(\cdot)$, all exhibit uniform convergence at the rate $\alpha(b, \delta) = \mathcal{O}\left(\sqrt{\frac{1}{b} \log \frac{1}{\delta}}\right)$.*

We will prove the four parts of the theorem in three separate subsections below. We shall consider a population $\mathbf{z}_1, \ldots, \mathbf{z}_n$ and a sample of size b $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_b$ chosen uniformly at random with (i.e. i.i.d.) or without replacement. We shall let p and $\hat{p}$ denote the fraction of positives in the population and the sample respectively. In the following, we shall reserve the notation $\hat{\mathbf{y}}$ for the label vector in the sample and shall use the notation $\tilde{\mathbf{y}}$ to denote candidate labellings in the definition of the surrogate.

D.1 A Uniform Convergence Bound for the prec@$\kappa(\cdot)$ Performance Measure

We note that a point-wise convergence result for prec@$\kappa(\cdot)$ follows simply from Lemma 22. To see this, given a population $\mathbf{z}_1, \ldots, \mathbf{z})n$ and a fixed model $\mathbf{w} \in \mathcal{W}$, construct a parallel population using the transformation $\mathbf{v}_i \leftarrow (\mathbf{w}^\top \mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^2$. We order these tuples according to their first component, i.e. along the scores and use $h(\mathbf{v}_i) = 1 - \mathbf{y}_i$. Let the population be arranged such that $\mathbf{v}_1 \succeq \mathbf{v}_2 \succeq \ldots$. Then this gives us

$$\sum_{i=1}^{k} h(\mathbf{v}_i) = \sum_{i=1}^{k} (1 - \mathbf{y}_i) = \text{prec@k}(\mathbf{y}, \mathbf{y}^{(\mathbf{w},k)}) = \text{prec@k}(\mathbf{w}).$$

Thus, the application of Lemma 22 gives us the following result

Lemma 23. *For any fixed model $\mathbf{w} \in \mathcal{W}$, with probability at least $1 - \delta$ over the choice of b samples, we have*

$$|\text{prec@}\kappa(\mathbf{w}; \mathbf{z}_1, \ldots, \mathbf{z}_n) - \text{prec@}\kappa(\mathbf{w}; \hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_b)| \leq \mathcal{O}\left(\sqrt{\frac{1}{b} \log \frac{1}{\delta}}\right).$$

To prove the uniform convergence result, we will, in some sense, require a uniform version of Lemma 22. To do so we fix some notation. For any fixed $\kappa > 0$, and for any $\mathbf{w} \in \mathcal{W}$, we will define $v_{\mathbf{w}}$ as the largest real number v such that

$$\sum_{i=1}^{n} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x}_i \geq v\right] = \kappa p n$$

Similarly, we will define $\hat{v}_{\mathbf{w}}$ as the largest real number v such that

$$\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \hat{\mathbf{x}}_i \geq v\right] = \kappa\hat{p}b$$

Using this notation we can redefine prec@$\kappa(\cdot)$ on the population, as well as the sample, as

$$\text{prec@}\kappa(\mathbf{w}; \mathbf{z}_1, \ldots, \mathbf{z}_n) := \frac{1}{\kappa pn}\sum_{i=1}^{n} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\mathbf{y}_i = 0\right]$$

$$\text{prec@}\kappa(\mathbf{w}; \hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_b) := \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq \hat{v}_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right]$$

We can now write

$$\begin{aligned}
&\sup_{\mathbf{w}\in\mathcal{W}} \left|\text{prec@}\kappa(\mathbf{w}; \mathbf{z}_1, \ldots, \mathbf{z}_n) - \text{prec@}\kappa(\mathbf{w}; \hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_b)\right| \\
&= \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa pn}\sum_{i=1}^{n} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\mathbf{y}_i = 0\right] - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq \hat{v}_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right]\right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa pn}\sum_{i=1}^{n} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\mathbf{y}_i = 0\right] - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right]\right| \\
&\quad + \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right] - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq \hat{v}_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right]\right| \\
&\leq \underbrace{\sup_{\mathbf{w}\in\mathcal{W}, t\in\mathbb{R}} \left|\frac{1}{\kappa pn}\sum_{i=1}^{n} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq t\right] \cdot \mathbb{I}\left[\mathbf{y}_i = 0\right] - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq t\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right]\right|}_{(A)} \\
&\quad + \underbrace{\sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right] - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq \hat{v}_{\mathbf{w}}\right] \cdot \mathbb{I}\left[\hat{\mathbf{y}}_i = 0\right]\right|}_{(B)}
\end{aligned}$$

Now, using a standard VC-dimension based uniform convergence argument over the class of thresholded classifiers, we get the following result: with probability at least $1 - \delta$

$$(A) \leq \mathcal{O}\left(\sqrt{\frac{1}{b}\left(\log\frac{1}{\delta} + d_{\text{VC}}(\mathcal{W}) \cdot \log b\right)}\right) = \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right),$$

where $d_{\text{VC}}(\mathcal{W})$ is the VC-dimension of the set of classifiers $\mathcal{W}$. Moving on to bound the second term, we can use an argument similar to the one used to prove Lemma 22 to show that

$$\begin{aligned}
(B) &\leq \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq \hat{v}_{\mathbf{w}}\right]\right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] - \kappa\right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{W}} \left|\frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right] - \frac{1}{\kappa pn}\sum_{i=1}^{n} \mathbb{I}\left[\mathbf{w}^\top \mathbf{x} \geq v_{\mathbf{w}}\right]\right|
\end{aligned}$$

$$\leq \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right),$$

where the last step follows from a standard VC-dimension based uniform convergence argument as before. This establishes the following uniform convergence result for the prec@k$(\cdot)$ performance measure

Theorem 24. *We have, with probability at least* $1-\delta$ *over the choice of* b *samples,*

$$\sup_{\mathbf{w}\in\mathcal{W}} \left|\text{prec@}\kappa(\mathbf{w};\mathbf{z}_1,\ldots,\mathbf{z}_n) - \text{prec@}\kappa(\mathbf{w};\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)\right| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right).$$

D.2 A Uniform Convergence Bound for the $\ell^{\textbf{ramp}}_{\textbf{prec@}\kappa}(\cdot)$ Surrogate

We first recall the form of the (normalized) surrogate below - note that this is a non-convex surrogate. Also recall that $k = \kappa \cdot n_+(\mathbf{y})$.

$$\ell^{\text{ramp}}_{\text{prec@}\kappa}(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n) = \underbrace{\max_{\|\tilde{\mathbf{y}}\|_1=k}\left\{\frac{\Delta(\mathbf{y},\tilde{\mathbf{y}})}{k} + \frac{1}{k}\sum_{i=1}^n \tilde{\mathbf{y}}_i\mathbf{w}^\top\mathbf{x}_i\right\}}_{\Psi_1(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n)} - \underbrace{\max_{\substack{\|\tilde{\mathbf{y}}\|_1=k\\ K(\mathbf{y},\tilde{\mathbf{y}})=k}} \frac{1}{k}\sum_{i=1}^n \tilde{\mathbf{y}}_i\mathbf{w}^\top\mathbf{x}_i}_{\Psi_2(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n)}$$

We will now show that both the functions $\Psi_1(\cdot)$, as well as $\Psi_2(\cdot)$, exhibit uniform convergence. This shall suffice to prove that $\ell^{\text{ramp}}_{\text{prec@}\kappa}(\cdot)$ exhibits uniform convergence. To do so we shall show that the two functions exhibit pointwise convergence and that they are Lipschitz. This will allow a standard L_∞ covering number argument Zhang [2002] to give us the required uniform convergence results.

D.2.1 A Uniform Convergence Result for $\Psi_1(\cdot)$

We have

$$\Psi_1(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n) = \max_{\|\tilde{\mathbf{y}}\|_1=\kappa pn}\left\{\frac{1}{\kappa pn}\sum_{i=1}^n \tilde{\mathbf{y}}_i(\mathbf{w}^\top\mathbf{x}_i - \mathbf{y}_i)\right\}+1$$

$$\Psi_1(\mathbf{w};\ \hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b) = \max_{\|\tilde{\mathbf{y}}\|_1=\kappa\hat{p}b}\left\{\frac{1}{\kappa\hat{p}b}\sum_{i=1}^b \tilde{\mathbf{y}}_i(\mathbf{w}^\top\hat{\mathbf{x}}_i - \hat{\mathbf{y}}_i)\right\}+1$$

An application of Corollary 29 indicates that $\Psi_1(\cdot)$ is Lipschitz i.e.

$$|\Psi_1(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n) - \Psi_1(\mathbf{w}';\ \mathbf{z}_1,\ldots,\mathbf{z}_n)| \leq \mathcal{O}\left(\|\mathbf{w}-\mathbf{w}'\|_2\right).$$

Thus, all that remains is to prove pointwise convergence. We decompose the error as follows

$$|\Psi_1(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n) - \Psi_1(\mathbf{w};\ \hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)| \leq \underbrace{\left|\Psi_1(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n) - \max_{\|\tilde{\mathbf{y}}\|_1=\kappa pb}\left\{\frac{1}{\kappa pb}\sum_{i=1}^b \tilde{\mathbf{y}}_i(\mathbf{w}^\top\hat{\mathbf{x}}_i - \hat{\mathbf{y}}_i)\right\}+1\right|}_{(A)}$$

$$+\underbrace{\left|\max_{\|\tilde{\mathbf{y}}\|_1=\kappa pb}\left\{\frac{1}{\kappa pb}\sum_{i=1}^b \tilde{\mathbf{y}}_i(\mathbf{w}^\top\hat{\mathbf{x}}_i - \hat{\mathbf{y}}_i)\right\}+1-\Psi_1(\mathbf{w};\ \hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)\right|}_{(B)}$$

An application of Lemma 22 using $\mathbf{v}_i = \mathbf{w}^\top\hat{\mathbf{x}}_i - \hat{\mathbf{y}}_i$ and $h(\cdot)$ as the identity function shows us that

$$(A) \leq \mathcal{O}\left(\frac{1}{\kappa p}\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right).$$

To bound the residual term (B), notice that an application of the Hoeffding's inequality tells us that with probability at least $1-\delta$

$$|p-\hat{p}| \leq \sqrt{\frac{1}{2b}\log\frac{2}{\delta}},$$

which lets us bound the residual as follows. Assume, for sake of simplicity, that the sample data points have been ordered in decreasing order of the quantity $\mathbf{w}^\top\hat{\mathbf{x}}_i - \mathbf{y}_i$ as well as that $\left|\mathbf{w}^\top\mathbf{x}\right| \leq 1$ for all $\mathbf{x}$.

$$\begin{aligned}
(B) &= \left|\max_{\|\tilde{\mathbf{y}}\|_1=\kappa pb}\left\{\frac{1}{\kappa pb}\sum_{i=1}^{b}\tilde{\mathbf{y}}_i(\mathbf{w}^\top\hat{\mathbf{x}}_i-\hat{\mathbf{y}}_i)\right\} - \max_{\|\tilde{\mathbf{y}}\|_1=\kappa\hat{p}b}\left\{\frac{1}{\kappa\hat{p}b}\sum_{i=1}^{b}\tilde{\mathbf{y}}_i(\mathbf{w}^\top\hat{\mathbf{x}}_i-\hat{\mathbf{y}}_i)\right\}\right| \\
&= \left|\frac{1}{\kappa pb}\sum_{i=1}^{\kappa pb}(\mathbf{w}^\top\hat{\mathbf{x}}_i-\hat{\mathbf{y}}_i) - \frac{1}{\kappa\hat{p}b}\sum_{i=1}^{\kappa\hat{p}b}(\mathbf{w}^\top\hat{\mathbf{x}}_i-\hat{\mathbf{y}}_i)\right| \\
&\leq \left|\sum_{i=1}^{\kappa\min\{p,\hat{p}\}b}\left(\frac{1}{\kappa pb}-\frac{1}{\kappa\hat{p}b}\right)(\mathbf{w}^\top\hat{\mathbf{x}}_i-\hat{\mathbf{y}}_i)\right| + \left|\frac{1}{\kappa\max\{p,\hat{p}\}b}\sum_{i=\kappa\min\{p,\hat{p}\}b+1}^{\kappa\max\{p,\hat{p}\}b}(\mathbf{w}^\top\hat{\mathbf{x}}_i-\hat{\mathbf{y}}_i)\right| \\
&\leq \frac{2}{\kappa b}\left|\frac{p-\hat{p}}{p\hat{p}}\right|\cdot\kappa\min\{p,\hat{p}\}b + \frac{2}{\kappa\max\{p,\hat{p}\}b}\cdot\kappa|p-\hat{p}|b \\
&= 2|p-\hat{p}|\cdot\left(\frac{\min\{p,\hat{p}\}}{p\hat{p}}+\frac{1}{\max\{p,\hat{p}\}}\right) \\
&\leq \sqrt{\frac{1}{2b}\log\frac{2}{\delta}}\cdot\frac{2}{\max\{p,\hat{p}\}} \leq \frac{2}{p}\sqrt{\frac{1}{2b}\log\frac{2}{\delta}}
\end{aligned}$$

This establishes that for any fixed $\mathbf{w}\in\mathcal{W}$, with probability at least $1-\delta$, we have

$$|\Psi_1(\mathbf{w};\ \mathbf{z}_1,\ldots,\mathbf{z}_n) - \Psi_1(\mathbf{w};\ \hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)| \leq \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right)$$

which concludes the uniform convergence proof.

D.2.2 A Uniform Convergence Result for $\Psi_2(\cdot)$

The proof follows similarly here with a direct application of Corollary 29 showing us that $\Psi_2(\cdot)$ is Lipschitz and an application of Lemma 22 along with the observation that $|p-\hat{p}| \leq \sqrt{\frac{1}{2b}\log\frac{2}{\delta}}$ similar to the discussion used above concluding the point-wise convergence proof.

The above two part argument establishes the following uniform convergence result for the $\ell^{\text{ramp}}_{\text{prec}@\kappa}(\cdot)$ performance measure

Theorem 25. *We have, with probability at least* $1-\delta$ *over the choice of* b *samples,*

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\ell^{\textit{ramp}}_{\text{prec}@\kappa}(\mathbf{w};\mathbf{z}_1,\ldots,\mathbf{z}_n) - \ell^{\textit{ramp}}_{\text{prec}@\kappa}(\mathbf{w};\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)\right| \leq \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right).$$

D.3 A Uniform Convergence Bound for the $\ell^{\textbf{avg}}_{\textbf{prec}@\kappa}(\cdot)$ Surrogate

This will be the most involved of the four bounds, given the intricate nature of the surrogate. We will prove this result using a series of partial results which we state below. As before, for any $\mathbf{w}\in\mathcal{W}$ and any $\tilde{\mathbf{y}}$, we define

$$\Delta(\mathbf{w},\tilde{\mathbf{y}}) := \frac{1}{\kappa pn}\left(\Delta(\mathbf{y},\tilde{\mathbf{y}}) + \sum_{i=1}^{n}(\tilde{\mathbf{y}}_i-\mathbf{y}_i)\mathbf{w}^\top\mathbf{x}_i + \frac{1}{C(\tilde{\mathbf{y}})}\sum_{i=1}^{n}(1-\tilde{\mathbf{y}}_i)\mathbf{y}_i\mathbf{w}^\top\mathbf{x}_i\right)$$

$$\hat{\Delta}(\mathbf{w},\tilde{\mathbf{y}}) := \frac{1}{\kappa\hat{p}b}\left(\Delta(\hat{\mathbf{y}},\tilde{\mathbf{y}}) + \sum_{i=1}^{n}(\tilde{\mathbf{y}}_i - \hat{\mathbf{y}}_i)\mathbf{w}^\top\hat{\mathbf{x}}_i + \frac{1}{C(\tilde{\mathbf{y}})}\sum_{i=1}^{n}(1-\tilde{\mathbf{y}}_i)\hat{\mathbf{y}}_i\mathbf{w}^\top\hat{\mathbf{x}}_i\right)$$

Recall that we are using $\hat{\mathbf{y}}$ to denote the true labels of the sample points and $\tilde{\mathbf{y}}$ to denote the candidate labellings while defining the surrogates. We also define, for any $\beta \in [0,1]$, the following quantities

$$\Delta(\mathbf{w},\beta) := \max_{\substack{\|\tilde{\mathbf{y}}\|_1=\kappa pn \\ K(\mathbf{y},\tilde{\mathbf{y}})=\beta pn}} \{\Delta(\mathbf{w},\tilde{\mathbf{y}})\}$$

$$\hat{\Delta}(\mathbf{w},\beta) := \max_{\substack{\|\tilde{\mathbf{y}}\|_1=\kappa\hat{p}b \\ K(\hat{\mathbf{y}},\tilde{\mathbf{y}})=\beta\hat{p}b}} \left\{\hat{\Delta}(\mathbf{w},\tilde{\mathbf{y}})\right\}$$

Note that β denotes a target true positive *rate* and consequently, can only take values between 0 and κ. Given the above, we claim the following lemmata

Lemma 26. *For every $\mathbf{w}$ and any $\beta,\beta' \in [0,\kappa]$, we have*

$$|\Delta(\mathbf{w},\beta) - \Delta(\mathbf{w},\beta')| \leq \mathcal{O}\left(|\beta-\beta'|\right).$$

Lemma 27. *For any fixed β, we have, with probability at least $1-\delta$ over the choice of the sample*

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\Delta(\mathbf{w},\beta) - \hat{\Delta}(\mathbf{w},\beta)\right| \leq \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right).$$

Using the above two lemmata as given, we can now prove the desired uniform convergence result for the $\ell^{\text{avg}}_{\text{prec}@\kappa}(\cdot)$ surrogate:

Theorem 28. *With probability at least $1-\delta$ over the choice of the samples, we have*

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\ell^{avg}_{\text{prec}@\kappa}(\mathbf{w};\mathbf{z}_1,\ldots,\mathbf{z}_n) - \ell^{avg}_{\text{prec}@\kappa}(\mathbf{w};\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)\right| \leq \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right).$$

Proof. We note that given the definitions of $\Delta(\mathbf{w},\beta)$ and $\hat{\Delta}(\mathbf{w},\beta)$, we can redefine the performance measure as follows

$$\ell^{\text{avg}}_{\text{prec}@\kappa}(\mathbf{w};\mathbf{z}_1,\ldots,\mathbf{z}_n) = \max_{\beta\in[0,\kappa]}\Delta(\mathbf{w},\beta)$$

We now note that for the population, the set of achievable values of true positive rates i.e. β is

$$B = \left\{0, \frac{1}{\kappa pn}, \frac{2}{\kappa pn}, \ldots, \frac{\kappa pn - 1}{\kappa pn}, 1\right\},$$

which correspond, respectively, to classifiers for which the *number* of true positives equals $\{0,1,2\ldots\kappa pn-1,\kappa pn\}$. Similarly, the set of achievable values of true positive rates i.e. β for the sample is

$$\hat{B} = \left\{0, \frac{1}{\kappa\hat{p}b}, \frac{2}{\kappa\hat{p}b}, \ldots, \frac{\kappa\hat{p}b - 1}{\kappa\hat{p}b}, 1\right\}.$$

Clearly, for any $\beta \in B$, there exists a $\pi_{\hat{B}}(\beta) \in \hat{B}$ such that

$$\left|\pi_{\hat{B}}(\beta) - \beta\right| \leq \frac{1}{\kappa\hat{p}b}.$$

Given this, let us define

$$\beta^*(\mathbf{w}) = \arg\max_{\beta\in[0,\kappa]}\Delta(\mathbf{w},\beta)$$

$$\hat{\beta}^*(\mathbf{w}) = \arg\max_{\hat{\beta}\in[0,\kappa]} \hat{\Delta}(\mathbf{w},\hat{\beta})$$

We shall assume, for the sake of simplicity, that $s|n$ so that $\hat{B} \subset B$. This gives us the following set of inequalities for any $\mathbf{w} \in \mathcal{W}$:

$$\begin{aligned}
\Delta(\mathbf{w},\beta^*(\mathbf{w})) &\leq \Delta(\mathbf{w},\pi_{\hat{B}}(\beta^*(\mathbf{w}))) + \left|\beta^*(\mathbf{w}) - \pi_{\hat{B}}(\beta^*(\mathbf{w}))\right| \\
&\leq \hat{\Delta}(\mathbf{w},\pi_{\hat{B}}(\beta^*(\mathbf{w}))) + \sup_{\mathbf{w}\in\mathcal{W}} \left|\Delta(\mathbf{w},\pi_{\hat{B}}(\beta^*(\mathbf{w}))) - \hat{\Delta}(\mathbf{w},\pi_{\hat{B}}(\beta^*(\mathbf{w})))\right| + \frac{1}{\kappa\hat{p}b} \\
&\leq \hat{\Delta}(\mathbf{w},\pi_{\hat{B}}(\beta^*(\mathbf{w}))) + \sup_{\mathbf{w}\in\mathcal{W},\hat{\beta}\in\hat{B}} \left|\Delta(\mathbf{w},\hat{\beta}) - \hat{\Delta}(\mathbf{w},\hat{\beta})\right| + \frac{1}{\kappa\hat{p}b} \\
&\leq \hat{\Delta}(\mathbf{w},\pi_{\hat{B}}(\beta^*(\mathbf{w}))) + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{b}{\delta}}\right) + \frac{1}{\kappa\hat{p}b} \\
&\leq \hat{\Delta}(\mathbf{w},\hat{\beta}^*(\mathbf{w})) + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{b}{\delta}}\right) + \frac{1}{\kappa\hat{p}b},
\end{aligned}$$

where the first step follows from Lemma 26, the third step follows since $\pi_{\hat{B}}(\beta^*(\mathbf{w})) \in \hat{B}$, the fourth step follows from an application of the union bound with Lemma 27 over the set of elements in $\hat{B}$ and noting $\left|\hat{B}\right| \leq \mathcal{O}(b)$, and the last step follows from the optimality of $\hat{\beta}^*(\mathbf{w})$. Similarly we can write, for any $\mathbf{w} \in \mathcal{W}$,

$$\begin{aligned}
\hat{\Delta}(\mathbf{w},\hat{\beta}^*(\mathbf{w})) &\leq \Delta(\mathbf{w},\hat{\beta}^*(\mathbf{w})) + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{b}{\delta}}\right) \\
&\leq \Delta(\mathbf{w},\beta^*(\mathbf{w})) + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{b}{\delta}}\right),
\end{aligned}$$

where the first step uses Lemma 27 with a union bound over elements in $\hat{B}$ and the fact that $\hat{\beta}^*(\mathbf{w}) \in \hat{B} \subset B$ (note that this assumption is not crucial to the argument – indeed, even if $\hat{\beta}^*(\mathbf{w}) \notin B$, we would only incur an extra $\mathcal{O}\left(\frac{1}{n}\right)$ error by an application of Lemma 26 since given the granularity of B, we would always be able to find a value in B that is no more than $\mathcal{O}\left(\frac{1}{n}\right)$ far from $\hat{\beta}^*(\mathbf{w})$), and the last step uses the optimality of $\beta^*(\mathbf{w})$. Thus, we can write

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{W}} \left|\ell^{\text{avg}}_{\text{prec}@\kappa}(\mathbf{w};\mathbf{z}_1,\ldots,\mathbf{z}_n) - \ell^{\text{avg}}_{\text{prec}@\kappa}(\mathbf{w};\hat{\mathbf{z}}_1,\ldots,\hat{\mathbf{z}}_b)\right| &= \sup_{\mathbf{w}\in\mathcal{W}} \left|\Delta(\mathbf{w},\beta^*(\mathbf{w})) - \hat{\Delta}(\mathbf{w},\hat{\beta}^*(\mathbf{w}))\right| \\
&\leq \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{b}{\delta}}\right) + \frac{1}{\kappa\hat{p}b} \\
&\leq \tilde{\mathcal{O}}\left(\sqrt{\frac{1}{b}\log\frac{1}{\delta}}\right),
\end{aligned}$$

since $\hat{p} \geq \Omega(1)$ with probability at least $1-\delta$. Thus, all we are left is to prove Lemmata 26 and 27 which we do below. To proceed with the proofs, we first write the form of $\Delta(\mathbf{w},\beta)$ for a fixed $\mathbf{w}$ and β and simplify the expression for ease of further analysis. We shall assume, for sake of simplicity, that βpn, κpn, $\beta\hat{p}b$, and $\kappa\hat{p}b$ are all integers.

$$\begin{aligned}
\Delta(\mathbf{w},\beta) &= \max_{\substack{\|\tilde{\mathbf{y}}\|_1=\kappa pn\\ K(\mathbf{y},\tilde{\mathbf{y}})=\beta pn}} \left\{\frac{1}{\kappa pn}\left(\Delta(\mathbf{y},\tilde{\mathbf{y}}) + \sum_{i=1}^n(\tilde{\mathbf{y}}_i - \mathbf{y}_i)\mathbf{w}^\top\mathbf{x}_i + \frac{1}{C(\tilde{\mathbf{y}})}\sum_{i=1}^n(1-\tilde{\mathbf{y}}_i)\mathbf{y}_i\mathbf{w}^\top\mathbf{x}_i\right)\right\} \\
&= 1 - \frac{\beta}{\kappa} - \underbrace{\frac{1}{\kappa pn}\left(\frac{\kappa-\beta}{1-\beta}\right)\sum_{i=1}^n\mathbf{y}_i\mathbf{w}^\top\mathbf{x}_i}_{A(\mathbf{w},\beta)} + \underbrace{\max_{\substack{\|\tilde{\mathbf{y}}\|_1=\kappa pn\\ K(\mathbf{y},\tilde{\mathbf{y}})=\beta pn}} \left\{\frac{1}{\kappa pn}\sum_{i=1}^n\tilde{\mathbf{y}}_i\left(1 - \frac{1-\kappa}{1-\beta}\cdot\mathbf{y}_i\right)\mathbf{w}^\top\mathbf{x}_i\right\}}_{B(\mathbf{w},\beta)}
\end{aligned}$$

We can similarly define $\hat{A}(\mathbf{w}, \beta)$ and $\hat{B}(\mathbf{w}, \beta)$ for the samples.

Proof of Lemma 26. We have, by the above simplification,

$$\left|\Delta(\mathbf{w}, \beta) - \Delta(\mathbf{w}, \beta')\right| = \frac{1}{\kappa}\left|\beta - \beta'\right| + \left|A(\mathbf{w}, \beta) - A(\mathbf{w}, \beta')\right| + \left|B(\mathbf{w}, \beta) - B(\mathbf{w}, \beta')\right|,$$

as well as, assuming without loss of generality, that $\left|\mathbf{w}^\top \mathbf{x}\right| \leq 1$ for all $\mathbf{w}$ and $\mathbf{x}$,

$$\begin{aligned}
\left|A(\mathbf{w}, \beta) - A(\mathbf{w}, \beta')\right| &\leq \left|\frac{\kappa - \beta}{1 - \beta} - \frac{\kappa - \beta'}{1 - \beta'}\right| \cdot \left|\frac{1}{\kappa pn}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{w}^\top \mathbf{x}_i\right| \\
&\leq \frac{(1-\kappa)\left|\beta - \beta'\right|}{\kappa(1-\beta)(1-\beta')} \leq \frac{1}{\kappa(1-\kappa)}\left|\beta - \beta'\right|,
\end{aligned}$$

where the last step follows since $\beta, \beta' \leq \kappa$. To analyze the third term i.e. $|B(\mathbf{w}, \beta) - B(\mathbf{w}, \beta')|$, we analyze the nature of the assignment $\tilde{\mathbf{y}}$ which defines $B(\mathbf{w}, \beta)$. Clearly $\tilde{\mathbf{y}}$ must assign βpn positives and $(\kappa - \beta)pn$ negatives a label of 1 and the rest, a label of 0. Since it is supposed to maximize the scores thus obtained, it clearly assigns the top ranked $(\kappa - \beta)pn$ negatives a label of 1. As far as positives are concerned, $\beta < \kappa$, we have $\left(1 - \frac{1-\kappa}{1-\beta}\right) \geq 0$ which means that the βpn top ranked positives will get assigned a label of 1.

To formalize this, let us set some notation. Let $s_1^+ \geq s_2^+ \geq \ldots \geq s_{pn}^+$ denote the scores of the positive points arranged in descending order. Similarly, let $s_1^- \geq s_2^- \geq \ldots \geq s_{(1-p)n}^-$ denote the scores of the negative points arranged in descending order. Given this notation, we can rewrite $B(\mathbf{w}, \beta)$ as follows:

$$B(\mathbf{w}, \beta) = \frac{1}{\kappa pn}\left(\left(\frac{\kappa - \beta}{1 - \beta}\right)\sum_{i=1}^{\beta pn} s_i^+ + \sum_{i=1}^{(\kappa-\beta)pn} s_i^-\right).$$

Thus, assuming without loss of generality that $\left|s_i^+\right|, \left|s_i^-\right| \leq 1$, we have,

$$\begin{aligned}
\left|B(\mathbf{w}, \beta) - B(\mathbf{w}, \beta')\right| &= \frac{1}{\kappa pn}\left|\left(\frac{\kappa - \beta}{1 - \beta}\right)\sum_{i=1}^{\beta pn} s_i^+ + \sum_{i=1}^{(\kappa-\beta)pn} s_i^- - \left(\frac{\kappa - \beta'}{1 - \beta'}\right)\sum_{i=1}^{\beta' pn} s_i^+ - \sum_{i=1}^{(\kappa-\beta')pn} s_i^-\right| \\
&\leq \frac{1}{\kappa pn}\left|\left(\frac{\kappa - \beta}{1 - \beta}\right)\sum_{i=1}^{\beta pn} s_i^+ - \left(\frac{\kappa - \beta'}{1 - \beta'}\right)\sum_{i=1}^{\beta' pn} s_i^+\right| + \frac{1}{\kappa pn}\left|\sum_{i=1}^{(\kappa-\beta)pn} s_i^- - \sum_{i=1}^{(\kappa-\beta')pn} s_i^-\right| \\
&\leq \left|\frac{\kappa - \beta}{1 - \beta} - \frac{\kappa - \beta'}{1 - \beta'}\right| \cdot \left|\frac{1}{\kappa pn}\sum_{i=1}^{\min\{\beta,\beta'\}pn} s_i^+\right| + \frac{1}{\kappa pn}\frac{\kappa - \max\{\beta, \beta'\}}{1 - \max\{\beta, \beta'\}}\left|\beta - \beta'\right| pn + \frac{\left|\beta - \beta'\right| pn}{\kappa pn} \\
&\leq \frac{1}{\kappa(1-\kappa)}\left|\beta - \beta'\right|\frac{\min\{\beta, \beta'\} pn}{\kappa pn} + \frac{1}{\kappa}\frac{\kappa - \max\{\beta, \beta'\}}{1 - \max\{\beta, \beta'\}}\left|\beta - \beta'\right| + \frac{\left|\beta - \beta'\right|}{\kappa} \\
&\leq \frac{2}{\kappa(1-\kappa)}\left|\beta - \beta'\right|,
\end{aligned}$$

where the last step uses the fact that $0 \leq \beta, \beta' \leq \kappa$. This tells us that

$$\left|\Delta(\mathbf{w}, \beta) - \Delta(\mathbf{w}, \beta')\right| \leq \frac{4 - \kappa}{\kappa(1-\kappa)}\left|\beta - \beta'\right|,$$

which finishes the proof. □

Proof of Lemma 27. We will prove the theorem by showing that the terms $A(\mathbf{w}, \beta)$ and $B(\mathbf{w}, \beta)$ exhibit uniform convergence.

It is easy to see that $A(\mathbf{w}, \beta)$ exhibits uniform convergence since it is a simple average of population scores. The only thing to be taken care of is that $A(\mathbf{w}, \beta)$ contains p in the normalization whereas $\hat{A}(\mathbf{w}, \beta)$ contains $\hat{p}$. However, since p and $\hat{p}$ are very close with high probability, an argument similar to the one used in the proof of Theorem 25 can be used to conclude that with probability at least $1 - \delta$, we have

$$\sup_{\mathbf{w} \in \mathcal{W}} \left| A(\mathbf{w}, \beta) - \hat{A}(\mathbf{w}, \beta) \right| \leq \mathcal{O}\left(\sqrt{\frac{1}{b} \log \frac{1}{\delta}} \right).$$

To prove uniform convergence for $B(\mathbf{w}, \beta)$ we will use our earlier method of showing that this function exhibits pointwise convergence and that this function is Lipschitz with respect to $\mathbf{w}$. The Lipschitz property of $B(\mathbf{w}, \beta)$ is evident from an application of Corollary 29. To analyze its pointwise convergence property

Thus the function $B(\mathbf{w}, \beta)$, as analyzed in the proof of Lemma 26, is composed by sorting the positives and negatives separately and taking the top few positions in each list and adding the scores present therein. This allows an application of Lemma 22, as used in the proof of Theorem 25, separately to the positive and negative lists, to conclude the pointwise convergence bound for $B(\mathbf{w}, \beta)$. □

This concludes the proof of the uniform convergence bound for $\ell^{\text{avg}}_{\text{prec}@\kappa}(\cdot)$. □

D.4 Proof of Lemma 21

Lemma 21. *Let $f_1, \ldots, f_m$ be m real valued functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ such that every f_i is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm. Then the function*

$$g(\mathbf{v}) = \max_{i \in [m]} f_i(\mathbf{v})$$

is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm too.

Proof. Fix $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^n$. The premise guarantees us that for any $i \in [m]$, we have

$$|f_i(\mathbf{v}) - f_i(\mathbf{v}')| \leq \|\mathbf{v} - \mathbf{v}'\|_\infty .$$

Now let $g(\mathbf{v}) = f_i(\mathbf{v})$ and $g(\mathbf{v}') = f_j(\mathbf{v}')$. Then we have

$$g(\mathbf{v}) - g(\mathbf{v}') = f_i(\mathbf{v}) - f_j(\mathbf{v}') \leq f_i(\mathbf{v}) - f_i(\mathbf{v}') \leq \|\mathbf{v} - \mathbf{v}'\|_\infty ,$$

since $f_j(\mathbf{v}') \geq f_i(\mathbf{v}')$. Similarly we have $g(\mathbf{v}') - g(\mathbf{v}) \leq \|\mathbf{v} - \mathbf{v}'\|_\infty$. This completes the proof. □

The following corollary would be most useful in our subsequent analyses.

Corollary 29. *Let $\Psi : \mathcal{W} \rightarrow \mathbb{R}$ be a function defined as follows*

$$\Psi(\mathbf{w}) = \max_{\substack{\hat{\mathbf{y}} \in \{0,1\}^n \\ \|\hat{\mathbf{y}}\|_1 = k}} \frac{1}{k} \sum \hat{\mathbf{y}}_i (\mathbf{w}^\top \mathbf{x}_i - c_i),$$

where c_i are constants independent of $\mathbf{w}$ and we assume without loss of generality that $\|\mathbf{x}_i\|_2 \leq 1$ for all i. Then $\Psi(\cdot)$ is 1- Lipschitz with respect to the L_2 norm i.e. for all $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$

$$|\Psi(\mathbf{w}) - \Psi(\mathbf{w}')| \leq \|\mathbf{w} - \mathbf{w}'\|_2 .$$

Proof. Note that for any $\hat{\mathbf{y}}$ such that $\|\hat{\mathbf{y}}\|_1 = k$, the function $f_{\hat{\mathbf{y}}}(\mathbf{v}) = \frac{1}{k} \sum \hat{\mathbf{y}}_i (\mathbf{v}_i - c_i)$ is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm. Thus if we define

$$\Phi(\mathbf{v}) = \max_{\|\hat{\mathbf{y}}\|_1 = k} f_{\hat{\mathbf{y}}}(\mathbf{v}),$$

then an application of Lemma 21 tells us that $\Phi(\cdot)$ is 1-Lipschitz with respect to the $\|\cdot\|_\infty$ norm as well. Also note that if we define

$$\mathbf{v}(\mathbf{w}) = \left(\mathbf{w}^\top \mathbf{x}_1 - c_1, \ldots, \mathbf{w}^\top \mathbf{x}_n - c_n\right),$$

then we have

$$\Psi(\mathbf{w}) = \Phi(\mathbf{v}(\mathbf{w}))$$

We now note that by an application of Cauchy-Schwartz inequality, and the fact that $\|\mathbf{x}_i\|_2 \leq 1$ for all i, we have

$$\|\mathbf{v}(\mathbf{w}) - \mathbf{v}(\mathbf{w}')\|_\infty \leq \|\mathbf{w} - \mathbf{w}'\|_2$$

Thus we have

$$|\Psi(\mathbf{w}) - \Psi(\mathbf{w}')| = |\Phi(\mathbf{v}(\mathbf{w})) - \Phi(\mathbf{v}(\mathbf{w}'))| \leq \|\mathbf{v}(\mathbf{w}) - \mathbf{v}(\mathbf{w}')\|_\infty \leq \|\mathbf{w} - \mathbf{w}'\|_2$$

which gives us the desired result. □

D.5 Proof of Lemma 22

Lemma 22. *Let $\mathcal{V}$ be a universe with a total order $\succeq$ established on it and let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a population of n items arranged in decreasing order. Let $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_b$ be a sample chosen i.i.d. (or without replacement) from the population and arranged in decreasing order as well. Then for any fixed $h : \mathcal{V} \to [-1, 1]$ and $\kappa \in (0, 1]$, we have, with probability at least $1 - \delta$ over the choice of the samples,*

$$\left| \frac{1}{\lceil \kappa n \rceil} \sum_{i=1}^{\lceil \kappa n \rceil} h(\mathbf{v}_i) - \frac{1}{\lceil \kappa b \rceil} \sum_{i=1}^{\lceil \kappa b \rceil} h(\hat{\mathbf{v}}_i) \right| \leq 4\sqrt{\frac{\log \frac{2}{\delta}}{\kappa b}}$$

Proof. We will assume, for sake of simplicity, that κn and κb are both integers so that there are no rounding off issues. Let $\mathbf{v}_n^* := \mathbf{v}_{\kappa n}$ and $\mathbf{v}_b^* := \hat{\mathbf{v}}_{\kappa b}$ denote the elements at the bottom of the κ-th fraction of the top in the sorted population and sample lists (recall that the population and the sample lists are sorted in descending order). Also let $\mathbb{T}(\mathbf{v}) := \mathbb{I}[\mathbf{v} \succeq \mathbf{v}_n^*]$ and $\hat{\mathbb{T}}(\mathbf{v}) := \mathbb{I}[\mathbf{v} \succeq \mathbf{v}_b^*]$ (note that $\mathbb{I}[E]$ is the indicator variable for the event E) so that we have

$$\begin{aligned}
\left| \frac{1}{\kappa n} \sum_{i=1}^{\kappa n} h(\mathbf{v}_i) - \frac{1}{\kappa b} \sum_{i=1}^{\kappa b} h(\hat{\mathbf{v}}_i) \right| &= \left| \frac{1}{\kappa n} \sum_{i=1}^{n} \mathbb{T}(\mathbf{v}_i) \cdot h(\mathbf{v}_i) - \frac{1}{\kappa b} \sum_{i=1}^{b} \hat{\mathbb{T}}(\hat{\mathbf{v}}_i) \cdot h(\hat{\mathbf{v}}_i) \right| \\
&\leq \left| \frac{1}{\kappa n} \sum_{i=1}^{n} \mathbb{T}(\mathbf{v}_i) \cdot h(\mathbf{v}_i) - \frac{1}{\kappa b} \sum_{i=1}^{b} \mathbb{T}(\hat{\mathbf{v}}_i) \cdot h(\hat{\mathbf{v}}_i) \right| + \left| \frac{1}{\kappa b} \sum_{i=1}^{b} \left(\mathbb{T}(\hat{\mathbf{v}}_i) - \hat{\mathbb{T}}(\hat{\mathbf{v}}_i) \right) \cdot h(\hat{\mathbf{v}}_i) \right| \\
&\leq 2\sqrt{\frac{\log \frac{2}{\delta}}{\kappa b}} + \underbrace{\left| \frac{1}{\kappa b} \sum_{i=1}^{b} \left(\mathbb{T}(\hat{\mathbf{v}}_i) - \hat{\mathbb{T}}(\hat{\mathbf{v}}_i) \right) \cdot h(\hat{\mathbf{v}}_i) \right|}_{(A)},
\end{aligned}$$

where the third step follows from Bernstein's inequality (which holds in situations with sampling without replacement as well Boucheron et al. [2004]) since $|\mathbb{T}(\mathbf{v}) \cdot h(\mathbf{v})| \leq 1$ for all $\mathbf{v}$ and we have assumed $b \geq \frac{1}{\kappa} \log \frac{2}{\delta}$. Now if $\mathbf{v}_n^* \succeq \mathbf{v}_b^*$, then we have $\hat{\mathbb{T}}(\mathbf{v}) \geq \mathbb{T}(\mathbf{v})$ for all $\mathbf{v}$. On the other hand if $\mathbf{v}_b^* \succeq \mathbf{v}_n^*$, then we have $\hat{\mathbb{T}}(\mathbf{v}) \leq \mathbb{T}(\mathbf{v})$ for all $\mathbf{v}$. This means that since $|h(\mathbf{v})| \leq 1$ for all $\mathbf{v}$, we have

$$(A) \leq \left| \frac{1}{\kappa b} \sum_{i=1}^{b} \left(\mathbb{T}(\hat{\mathbf{v}}_i) - \hat{\mathbb{T}}(\hat{\mathbf{v}}_i) \right) \right| = \left| \frac{1}{\kappa b} \sum_{i=1}^{b} \mathbb{T}(\hat{\mathbf{v}}_i) - 1 \right| \leq 2\sqrt{\frac{\log \frac{2}{\delta}}{\kappa b}},$$

where the second step follows since $\frac{1}{\kappa b} \sum_{i=1}^{b} \hat{\mathbb{T}}(\hat{\mathbf{v}}_i) = 1$ by definition and the last step follows from another application of Bernstein's inequality. This completes the proof. □

D.6 A Uniform Convergence Bound for the $\ell^{\text{max}}_{\text{prec}@\kappa}(\cdot)$ Surrogate

Having proved a generalization bound for the $\ell^{\text{avg}}_{\text{prec}@\kappa}(\cdot)$ surrogate, we note that similar techniques, that involve partitioning the candidate label space into labels that have a fixed true positive rate β, and arguing uniform convergence for each partition, can be used to prove a generalization bound for the $\ell^{\text{max}}_{\text{prec}@\kappa}(\cdot)$ surrogate as well. We postpone the details of the argument to a later version of the paper.

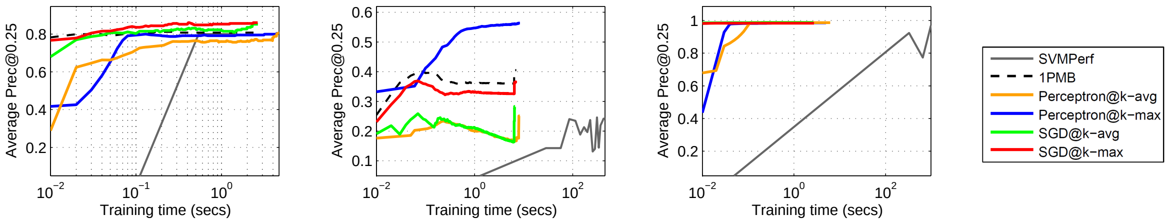
E Proof of Theorem 15

Theorem 15. *Let $\bar{\mathbf{w}}$ be the model returned by Algorithm 3 when executed on a stream with T batches of length b. Then with probability at least $1-\delta$, for any $\mathbf{w}^* \in \mathcal{W}$, we have*

$$\ell^{avg}_{\text{prec}@\kappa}(\bar{\mathbf{w}}; \mathcal{Z}) \leq \ell^{avg}_{\text{prec}@\kappa}(\mathbf{w}^*; \mathcal{Z}) + \mathcal{O}\left(\sqrt{\frac{1}{b}\log\frac{T}{\delta}}\right) + \mathcal{O}\left(\sqrt{\frac{1}{T}}\right)$$

Proof. The proof of this theorem closely follows that of Theorems 7 and 8 in Kar et al. [2014]. More specifically, Theorem 6 from Kar et al. [2014] ensures that any convex loss function demonstrating uniform convergence would ensure a result of the kind we are trying to prove. Since Theorem 12 confirms that $\ell^{\text{avg}}_{\text{prec}@\kappa}(\cdot)$ exhibits uniform convergence, the proof follows. □

F Additional Empirical Results



(a) KDD08 (b) Covtype (c) Cod-RNA

Figure 4: A comparison of the proposed perceptron and SGD based methods with baseline methods (SVMPerf and **1PMB**) on prec@0.25 maximization tasks.