Surrogate Functions for Maximizing Precision at the Top

Purushottam Kar T-PURKAR@MICROSOFT.COM
Microsoft Research, INDIA

Harikrishna Narasimhan[*] HARIKRISHNA@CSA.IISC.ERNET.IN
Indian Institute of Science, Bangalore, INDIA

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research, INDIA

Abstract

The problem of maximizing precision at top, also dubbed Precision@k, finds relevance in myriad learning applications such as ranking, multi-label classification, and learning with severe label imbalances. Despite its popularity, Precision@k is not known to have a surrogate function that upper bounds it. Similarly, notions of consistency under certain noise/margin conditions are also not explored.

In this work, we devise two novel convex surrogate functions for Precision@k, that upper bound it and are motivated by certain natural notions of margin for Precision@k performance measure. We also provide two novel perceptron algorithms for Precision@k that have interesting mistake bounds w.r.t. the proposed surrogates. Moreover, we devise scalable stochastic gradient descent style methods for our proposed surrogates and prove convergence bounds for the same. Our convergence bounds rely on a strong uniform convergence bound for Precsion@k and crucially exploit the structural simplicity of Precision@k. We conclude with experimental evidence of superiority of our surrogates when compared to the structural SVM surrogate (Joachims, 2005), a state-of-the-art approach to optimize Precision@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 rare-class classification problems like 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 present in a data point (Tsoumakas & Katakis, 2007).

Naturally, the ranking of points/labels at the top is of utmost importance to an application. Consequently, several performance measures have been designed to promote accuracy at top. Popular examples include Precision@k, Average Precision, NDCG (Tsoumakas & Katakis, 2007) etc.

While Precision@k (prec@k) is used as a key performance measure in several domains, there are a very few approaches to directly optimize Precision@k. In fact, to the best of our knowledge, there is only one known surrogate function for prec@k in literature, namely, the struct-SVM surrogate by (Joachims, 2005). However, as we reveal in this work, the struct-SVM surrogate is not a proper surrogate as it does not upper bound prec@k (see Appendix A for more details).

In this paper, our goal is to design efficient and consistent algorithms for optimizing prec@k. Given the intractability of even binary classification in the agnostic model (Guruswami & Raghavendra, 2009), we would instead focus on natural notions of *benign-ness* that are frequently satisfied by real life data. Indeed, the notion of margin in binary classification is well-established in several real-world scenarios and has led to tremendous progress in the area.

prec@k Margin: Motivated by the success of margin based frameworks in classification domains, we first develop a natural notion of margin for prec@k. In particular, we say that a dataset has a (k, γ) margin w.r.t. prec@k if there exist at least k positively labeled points that are all separated (with a margin of γ) from all the negatively labeled points. This notion of margin is well motivated for prec@k since we are only interested in making accurate predictions at the

Proceedings of the 31st International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

[*] Part of the work was done while H.N. was an intern at MSR India.

top-k positions in our ranked list. Moreover, it can be easily seen that this is a strictly weaker notion of margin than the margin notion for binary classification which requires *every* positive point to be separated from *every* negative point by a certain margin. Consequently, existing methods for binary classification such as perceptron/SVM do not apply directly to the prec@k problem, as the positive points (except for k privileged points) and the negative points might be adversarially mixed in an arrangement as proposed by (Guruswami & Raghavendra, 2009).

Using insights from the above defined margin notion for prec@k, we design a novel surrogate function (called $ramp$-surrogate) that upper bounds prec@k loss and can also be shown to be *consistent* w.r.t. prec@k as long as the dataset satisfies the above mentioned margin condition. However, the $ramp$-surrogate has a term that computes minimum of certain linear functions. Hence, it turns out to be a non-convex function and is not amenable to convex optimization techniques. To ameliorate this issue, we provide two successive relaxations of the $ramp$-surrogate. We first relax the above mentioned *min* function to the *average* (avg) function, which leads us to a surrogate that we call avg-surrogate. We then relax the average function to the *max* function, leading to the max-surrogate. Naturally, the obtained surrogates can also be shown to upper bound the true prec@k loss. Moreover, we can show that both of our proposed convex surrogates are also consistent w.r.t. prec@k, albeit under certain stronger margin notions.

Next, we propose two perceptron algorithms for the above mentioned convex surrogates. We show that under certain natural but stronger margin conditions, our proposed algorithms exhibit a mistake bound similar to the one obtained by the standard perceptron algorithm. Our notion of margin for the avg surrogate requires that the positive points on an average be separated (by some margin) from all of the negatives. Note that this margin notion is also significantly weaker than the standard margin notion for classification. Our notion of margin for the max surrogate is exactly the same as the one for binary classification.

We also devise stochastic gradient descent (SGD) based methods for optimizing the proposed surrogates. Note that in general, prec@k cannot be written as a sum of loss functions for each individual training points. Hence standard convergence analyses for SGD do not apply for our methods (Shalev-Shwartz et al., 2011). Instead, we prove the the convergence bound by combining a novel uniform convergence bound for our surrogates along with a generic technique by (Kar et al., 2014). Our uniform convergence bounds need to crucially exploit the structure of the surrogates, as the surrogates are dependent on all the training points and hence can potentially change significantly by perturbing one data point. Our key structural lemmas for these surrogates rule out such a possibility.

Finally, we validate our proposed surrogate functions and the corresponding methods on benchmark datasets. Empirical validation on several benchmark datasets reveals that our methods are in general significantly more accurate than the struct-SVM loss (Joachims, 2005) based methods, hence matching our theoretical bounds. We also observe that our SGD based methods scale much better to large-scale datasets as compared to the existing methods (Joachims, 2005; Kar et al., 2014) while providing competitive or better accuracy.

Paper Organization: Section 2 presents the problem formulation and sets up the notation. We present our three novel surrogates for prec@k in Section 3. Next, in Section 4 we present two perceptron algorithms for prec@k and their mistake bounds. Section 5 discussed uniform convergence and generalization error bounds for our algorithms. We conclude with empirical results in Section 6.

1.1. 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 the 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 in the top portion of the ranked list, and in ranking methods that directly optimize these measures (Clémençon & Vayatis, 2007; Rudin, 2009; Agarwal, 2011; Boyd et al., 2012; Narasimhan & Agarwal, 2013a;b; Li et al., 2014). In this work, we focus on one such evaluation measure, Precision@k, used widely in practice. The only prior algorithms that we are aware of that directly optimizes this evaluation measure is a structSVM based method due to (Joachims, 2005), and an efficient stochastic implementation of this method due to (Kar et al., 2014). However, as pointed out earlier the convex surrogate used in these methods is not well-suited for Precision@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 the subset/list-wise ranking settings, which arises in several information retrieval applications, where again there has been much work in optimizing performance measures that emphasize accuracy at the top (e.g. NDCG) (Valizadegan et al., 2009; Cao et al., 2007; Yue et al., 2007; Le & Smola, 2007; Chakrabarti et al., 2008; Yun et al., 2014). There has also been some recent work on perceptron style ranking methods for this list-wise ranking settings (Chaudhuri & Tewari, 2014), but these methods are tailored to optimize the NDCG and MAP measures, which are different from the Precision@k measure that we consider here. Other less related work include online ranking algorithms for optimizing popular ranking measures in a certain adversarial setting with limited feedback (Chaudhuri & Tewari, 2015).

2. Problem Formulation and Notations

We now set some notation. We are interested in supervised learning settings where we are presented with a set of labeled points $(\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_n, y_n)$, where each $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{0, 1\}$. We abbreviate this data set using the notation $(\mathbf{X}, \mathbf{y})$ where $\mathbf{X} \in \mathcal{X}^n$ and $\mathbf{y} \in \{0,1\}^n$. $\mathbf{z} = (\mathbf{x}, y)$ denotes a labeled data point, and $\mathbf{X}_+$ and $\mathbf{X}_-$ refer to the set of positive and negatively labeled points, respectively. Our results readily extend to multi-label and ranking models, but for simplicity of exposition, we focus only on binary classification problems in this paper.

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

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

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 notation $\text{prec@k}(s) := \text{prec@k}(s;\ \mathbf{z}_1, \ldots, \mathbf{z}_n)$ and suppress mention of the data points if the set of points is clear from context. The same will hold true for any surrogates that we introduce later. We also use the shorthand $s_i = s(\mathbf{x}_i)$. For any label vectors $\mathbf{y}', \mathbf{y}'' \in \{0,1\}^n$, define

$$\Delta(\mathbf{y}', \mathbf{y}'') = \sum_{i=1}^{n} (1 - \mathbf{y}'_i)\mathbf{y}''_i, \;\; K(\mathbf{y}', \mathbf{y}'') = \sum_{i=1}^{n} \mathbf{y}'_i \mathbf{y}''_i. \quad (2)$$

Note that $\|\mathbf{y}'\|_1 = K(\mathbf{y}', \mathbf{y}')$ denotes the number of positive points in a label vector $\mathbf{y}'$. Hence, $n_+(\mathbf{y}) = K(\mathbf{y}, \mathbf{y})$ where y is the *true* label vector. We also use shorthand n_+ when the context is clear. $\mathbf{y}^{(s,k)}$ denotes the predicted label vector for a given scoring function $s : \mathcal{X} \to \mathbb{R}$. 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, it is natural to seek easy to optimize surrogate functions that act as a good proxy for prec@k. There will be two properties that we shall desire of such a surrogate: a) the surrogate should always *upper bound* prec@k loss, so that minimizing the surrogate indeed leads to small prec@k, b) the surrogate should be *conditionally consistent* w.r.t. prec@k. This is to say, under some regularity assumptions, it should be possible to show that optimizing the surrogate implies an optimal solution for prec@k as well.

Motivated by the above requirements, we provide a family of surrogates which upper bound the prec@k loss function. Furthermore, our surrogates are designed so that for certain natural notions of margin (w.r.t the prec@k loss), i.e. for appropriate noise conditions, we can prove that our surrogates are actually consistent with the prec@k loss function.

Although prec@k is a popularly used performance measure used by several works to evaluate models in a variety of settings such as multi-label learning (Prabhu & Varma, 2014), results aimed at directly optimizing this performance measure are few and far between. In fact, the only known direct surrogate for prec@k is a structural SVM based surrogate by (Joachims, 2005), that we refer to as $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$.

Unfortunately, this surrogate falls short of meeting our aforementioned requirements since it does not even upper bound the prec@k loss, let alone be consistent with respect to it. We direct the reader to Appendix A for more details and a counter example that proves this claim.

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

To design our surrogate, we first revisit the struct-SVM surrogate for prec@k to better understand the reason for its failure. As it turns out, the very reason this surrogate fails would end up motivating the design of our surrogates. 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 the structured output prediction problems that can have exponentially large output spaces. Given a set of n labeled data points, the $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$ is defined as

$$\ell^{\text{struct}}_{\text{prec@k}}(s) = \max_{\substack{\hat{\mathbf{y}} \in \{0,1\}^n \\ \|\hat{\mathbf{y}}\|_1 = k}} (\Delta(\mathbf{y}, \hat{\mathbf{y}}) + \sum_{i=1}^{n} (\hat{\mathbf{y}}_i - \mathbf{y}_i)\, s_i).$$

The above surrogate tries to penalize a scoring function if it is possible to label k points as positives with very large scores (i.e., the second term is large) but which are labeled as "negatives" by the true label vector $\mathbf{y}$ (i.e., the first term is also large). However, one issue with this setup is that the candidate labeling $\hat{\mathbf{y}}$ is restricted to predict only k positives whereas the true label vector $\mathbf{y}$ has $n_+ \geq k$ positives. Hence, a non-optimal labeling can exploit the remaining $n_+ - k$ labels to hide the high scoring negative points thus confusing the loss function. Consequently a poor scoring function might have end up having very small $\ell^{\text{struct}}_{\text{prec@k}}(s)$ loss. See Appendix A for an explicit example.

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

The goal behind ranking in a bipartite setting is to select a subset of relevant items and rank them at the top k positions. Now 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 makes it possible to assign 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:

$$\max_{\|\hat{\mathbf{y}}\|_1=k}\left\{\Delta(\mathbf{y},\hat{\mathbf{y}})+\sum_{i=1}^{n}\hat{\mathbf{y}}_i s_i\right\}-\underbrace{\max_{\substack{\|\tilde{\mathbf{y}}\|_1=k\\ K(\mathbf{y},\tilde{\mathbf{y}})=k}}\sum_{i=1}^{n}\tilde{\mathbf{y}}_i s_i}_{(P)}. \quad (3)$$

Note that the above loss function is similar to the "ramp" losses for binary classification, variants of which have been proposed in (Do et al., 2008). We now show that the above loss function is indeed a surrogate of prec@k, in the sense that it upper bounds prec@k.

Claim 1. *For any $k \leq n_+$ and scoring function s, we have: $\ell^{ramp}_{prec@k}(s) \geq prec@k(s)$. Moreover, if $\ell^{ramp}_{prec@k}(s) \leq \xi$ 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: $\sum_{i\in S} s_i \geq \sum_{i=1}^{n}\hat{\mathbf{y}}_i s_i + \Delta(\mathbf{y},\hat{\mathbf{y}}) - \xi$.*

See Appendix B for a detailed proof. We now show that this surrogate satisfies our second condition, that of conditional consistency as well. We can show that if a scoring function s assigns the top k scores to only positive points which are greater than the score of any negative point by at least one, then $\ell^{\text{ramp}}_{\text{prec@k}}(s) = 0$ (see Claim 3). In fact, this "separation" condition for the scoring function motivates the following 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 some $S_+ \subseteq \mathbf{X}_+$ 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.*

Informally, a dataset has *weak* (k,γ)-margin if there exists at least one set of k positive points that are substantially far away from all the negatives. Note that this margin notion is strictly weaker than the usual margin condition for binary classification, as this notion allows many positives to be completely mingled with the negatives, so long as a small fraction of positives is separated from the negatives. We believe that the this notion of margin is one of the most natural notions of margin for prec@k. The following lemma shows that $\ell^{\text{ramp}}_{\text{prec@k}}$ is consistent w.r.t. prec@k for any dataset that exhibits *weak* k-margin.

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

$$\ell^{ramp}_{prec@k}(s) = prec@k(s)$$

Claims 1 and 3 suggest that $\ell^{\text{ramp}}_{\text{prec@k}}$ is indeed a tight surrogate for prec@k. Unfortunately, $\ell^{\text{ramp}}_{\text{prec@k}}$ is also a non-convex loss function, mainly due to the second term (P) in its definition (3). To alleviate this issue, we further relax the surrogate so as to obtain more tractable convex surrogates. To this end, we first re-write the term (P):

$$(P) = \sum_{i=1}^{n}\mathbf{y}_i s_i - \underbrace{\min_{\substack{\tilde{\mathbf{y}}\preceq\mathbf{y}\\ \|\tilde{\mathbf{y}}\|_1=n_+-k}}\sum_{i=1}^{n}\tilde{\mathbf{y}}_i s_i}_{(Q)}, \quad (4)$$

where $\tilde{\mathbf{y}} \preceq \mathbf{y}$ implies that $\mathbf{y}_i = 0 \Rightarrow \tilde{\mathbf{y}}_i = 0$. Thus, to convexify the surrogate $\ell^{\text{ramp}}_{\text{prec@k}}(s)$, 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 the different surrogate functions.

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

An immediate upper bound on (Q) is to relax the $\min$ function in (Q) with the $\max$ function. Since the $\max$ function is convex, this should convexify the surrogate. Noticing the fact that the "candidate labeling" $\hat{\mathbf{y}}$ in (4) has to predict at least $n_+ - k$ false negatives, we obtain the following upper bound: $(Q) \leq \max_{\substack{\tilde{\mathbf{y}}\preceq(1-\hat{\mathbf{y}})\cdot\mathbf{y}\\ \|\tilde{\mathbf{y}}\|_1=n_+-k}} \sum_{i=1}^{n}\tilde{\mathbf{y}}_i s_i,$

which gives us the $\ell^{\text{max}}_{\text{prec@k}}(s)$ surrogate defined below:

$$\max_{\|\hat{\mathbf{y}}\|_1=k}(\Delta(\mathbf{y},\hat{\mathbf{y}})+\sum_{i=1}^{n}(\hat{\mathbf{y}}_i-\mathbf{y}_i)s_i+\max_{\substack{\tilde{\mathbf{y}}\preceq(1-\hat{\mathbf{y}})\cdot\mathbf{y}\\ \|\tilde{\mathbf{y}}\|_1=n_+-k}}\sum_{i=1}^{n}\tilde{\mathbf{y}}_i s_i). \quad (5)$$

The above surrogate, being a point-wise max over convex functions, is convex, as well as an upper bound on prec@k(s) since it upper bounds $\ell^{\text{ramp}}_{\text{prec@k}}(\cdot)$ which itself upper bounds prec@k(s). The max surrogate (5) also exhibits *consistency* w.r.t. prec@k as long as the data satisfies the *strong* (k,γ)-margin defined below:

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

We notice that the strong margin condition is exactly equivalent to the notion of margin used in binary classification and hence strictly stronger than our *weak* (k,γ)-margin. This leads us to believe that there might exist tighter convex relaxations to the term (Q). Indeed the following relaxation gives us a tighter surrogate.

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

A tighter upper bound on (Q) (than the max) is to replace Q by the average of the false negatives, which can be rewritten as:

$$(Q) \le \frac{1}{C(\hat{\mathbf{y}})}\sum_{i=1}^{n}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i,$$

where $C(\hat{\mathbf{y}}) = \frac{n_+ - K(\mathbf{y},\hat{\mathbf{y}})}{n_+ - k} \ge 1$ whenever $k \le n_+$. By combining the above upper bound on (Q) with (3), we get a new convex relaxation $\ell^{\text{avg}}_{\text{prec@k}}(s) = \max_{\|\hat{\mathbf{y}}\|_1 = k} \Delta(\mathbf{y},\hat{\mathbf{y}}; s)$, where $\Delta(\mathbf{y},\hat{\mathbf{y}}; s)$ is given by:

$$\Delta(\mathbf{y},\hat{\mathbf{y}}) + \sum_{i=1}^{n} s_i(\hat{\mathbf{y}}_i - \mathbf{y}_i) + \frac{1}{C(\hat{\mathbf{y}})}\sum_{i=1}^{n}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i. \tag{6}$$

Now, it is easy to see that $\ell^{\text{avg}}_{\text{prec@k}}(s) \ge$ prec@k because $\ell^{\text{avg}}_{\text{prec@k}}(s)$ upper bounds $\ell^{\text{ramp}}_{\text{prec@k}}$ (3) and $\ell^{\text{ramp}}_{\text{prec@k}} \ge$ prec@k using Claim 1. For completeness, we provide formal claim and proof in Appendix B.4.

It is notable that for $k = n_+$ (when the performance measure reduces to the well-known precision-recall break even point or PRBEP), the surrogate $\ell^{\text{avg}}_{\text{prec@k}}(\cdot)$ reduces to Joachims' formulation $\ell^{\text{struct}}_{\text{prec@k}}(\cdot)$.

Also, similar to the $ramp$-surrogate, we can show that the avg-surrogate is consistent with the performance measure prec@k$(\cdot)$ under the following margin assumptions:

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 note that the above margin definition is a strictly weaker condition that the usual notion of margin for binary classification since it only requires the existence of a score function that assigns a higher average score to the bottom most $n_+ - k + 1$ positive points than the score to the highest ranked negative point than by a unit which still allows a non negligible fraction of the positive points to be assigned a lower score than those assigned to negatives. We also note that whenever a classifier $\mathbf{w}$ realizes the (k,γ)-margin, the scaled classifier $\frac{\mathbf{w}}{\gamma}$ realizes the k-margin condition.

On the other hand, the above margin condition is strictly stronger than the *weak* (k,γ)-margin condition (Definition 2). The weak-margin condition only requires one set of k-positives to be separated from the negatives, while the above margin condition require the average of *all* positives to be separated from the negatives.

We now show that under the above defined (k,γ)-margin, our avg-surrogate (6) is consistent with the prec@k performance measure. That is, our surrogate presents a tight convex upper bound to the prec@k performance measure.

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

See Appendix B for a detailed proof. as well as (k,γ)-margin (see Definition 2, 5)

Hence, our all three surrogates presented above fall in a nice hierarchy so that for any score function s, we have

$$\boxed{\text{prec@k}(s) \le \underbrace{\ell^{\text{ramp}}_{\text{prec@k}}(s)}_{\text{non-convex}} \le \underbrace{\ell^{\text{avg}}_{\text{prec@k}}(s) \le \ell^{\max}_{\text{prec@k}}(s)}_{\text{convex}}}$$

Figure 1. A hierarchy describing the three surrogates for prec@k

In the next section, we formulate two perceptron algorithms that can be shown to optimize our two convex surrogate functions: avg-surrogate and max-surrogate. Moreover, we provide mistake bounds for the two algorithms based on the (k,γ)-margin as well as the *strong* (k,γ)-margin, defined above.

4. Perceptron Algorithms for prec@k

We now present perceptron-style algorithms for maximizing the prec@k performance measure by using our proposed convex surrogates.

Our first perceptron algorithm PERCEPTRON@K (see Algorithm 1) works with an incoming stream of binary labeled points and processes them in *mini-batches* of a predetermined size b. Recently, mini-batch methods have been popular and also have been used for maximizing the struct-SVM surrogate ($\ell^{\text{struct}}_{\text{prec@k}}$) as well (Kar et al., 2014). Note that, for ranking and multi-label classification settings, mini-batches are not required and the algorithm can be applied to a single data points.

At a high level, our algorithm receives a batch of b points and predicts the label vector $\mathbf{y}_t \in \{0,1\}^b$ using the existing model $\mathbf{w}^{t-1}$. If prec@k loss is 0 then $\mathbf{w}^{t-1}$ is not updated. For non-zero prec@k, $\mathbf{w}^{t-1}$ is updated using all the false-positives as well as the false-negatives in the current mini-batch (see Line 11, 12 of Algorithm 1). Note that in the limiting case of $n_+ = k = 1$ (with $b = 1$), Perc@k-avg reduces to the standard perceptron algorithm(Rosenblatt, 1958; Minsky & Papert, 1988).

Next, we show that the above algorithm actually enjoys a mistake bound similar to those known for the traditional perceptron algorithm (Novikoff, 1962) with the hinge loss function replaced with our surrogate $\ell^{\text{avg}}_{\text{prec@k}}(s)$.

Theorem 7. *Suppose $\left\|\mathbf{x}_t^i\right\| \le R$ 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*

Algorithm 1 PERCEPTRON@K-AVG

```
Input: Batch length $b$
1:  $\mathbf{w}^0 \leftarrow \mathbf{0}, t \leftarrow 0$
2:  while stream not exhausted do
3:      $t \leftarrow t + 1$
4:      Receive $b$ data points $\mathbf{X}_t = [\mathbf{x}_t^1, \ldots, \mathbf{x}_t^b], \mathbf{y}_t \in \{0,1\}^b$
5:      Calculate $s_t = \mathbf{w}^{t-1}\mathbf{X}_t$ and let $\hat{\mathbf{y}}_t = \mathbf{y}^{(s_t,k)}$
6:      $\Delta_t \leftarrow \Delta(\mathbf{y}_t, \hat{\mathbf{y}}_t)$
7:      if $\Delta_t = 0$ then
8:          $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1}$
9:      else
10:         $D_t \leftarrow \frac{\Delta_t}{\|\mathbf{y}_t\|_1 - K(\mathbf{y}_t, \hat{\mathbf{y}}_t)}$
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 pos.}
12:         $\mathbf{w}^t \leftarrow \mathbf{w}^t + D_t \cdot \sum_{i \in [b]} (1 - \hat{\mathbf{y}}_i)\mathbf{y}_i \cdot \mathbf{x}_t^i$      {false neg.}
13:     end if
14: end while
15: 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 pos.}
12':        $\mathbf{w}^t \leftarrow \mathbf{w}^t + \sum_{i \in S_t} \mathbf{x}_t^i$      {top ranked false neg.}
```

$\hat{\mathcal{L}}_T^{avg}(\mathbf{w}) = \sum_{t=1}^T \ell_{prec@k}^{avg}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)$. *Then we have*

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

For "separable settings", we can trade-off the two terms in the mistake bound (RHS above) so that it reduces to a form that is similar to the mistake bound for standard perceptron (Novikoff, 1962).

Corollary 8. *Suppose there exists a unit norm classifier $\mathbf{w}^*$ 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: $\Delta_T^C \leq \frac{4kR^2}{\gamma^2}$.*

Hence as the dataset becomes "easier" in the (k, γ)-margin sense, Perc@k-avg converges to an optimal hyperplane at a faster rate. Here we would like to stress that (k, γ)-margin is strictly weaker than the standard classification margin. Hence for several datasets, Perc@k-avg might achieve 0 prec@k loss while the standard binary classification methods might not be able to find any reasonable classifier in poly-time (Guruswami & Raghavendra, 2009).

Note that Perc@k-avg updates all the false negatives for a given mini-batch. A natural question here might be that can we design an algorithm that requires to update $\mathbf{w}^t$ according to much smaller number of points. Such updates are slightly faster and ensures that $\mathbf{w}^t$ is more sparse in large scale settings. Our Perc@k-max algorithm (Algorithm 2) answers this question in the affirmative.

Perc@k-max differs from Perc@k-avg in that it performs updates using only a few of the *top ranked* false negatives.

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 $\mathbf{X}_t = [\mathbf{x}_t^1, \ldots, \mathbf{x}_t^b], \mathbf{y}_t \in \{0,1\}^b$
    Set $\mathbf{g}_t \in \partial_{\mathbf{w}} \ell_{\text{prec@k}}^{\text{avg}}(\mathbf{w}_{t-1}; \mathbf{X}_t, \mathbf{y}_t)$      {See Algorithm 4}
    $\mathbf{w}_t \leftarrow \Pi_{\mathcal{W}} [\mathbf{w}_{t-1} - \eta_t \cdot \mathbf{g}_t]$      {project onto set $\mathcal{W}$}
end while
return $\bar{\mathbf{w}} = \frac{1}{t} \sum_{\tau=1}^t \mathbf{w}_t$
```

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

```
Input: A model $\mathbf{w}_{\text{in}}$, $n$ data points $\mathbf{X}, \mathbf{y}$, parameter $k$
Output: A subgradient $\mathbf{g} \in \partial_{\mathbf{w}} \ell_{\text{prec@k}}^{\text{avg}}(\mathbf{w}_{\text{in}}; \mathbf{X}, \mathbf{y})$
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 $k' = 0 \to k$ 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 $\mathbf{g}_{k^*}$
```

More specifically, for any scoring function s and $m > 0$, define:

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

as the set of top m-ranked false negatives. The algorithm makes updates for false positives in the set FN(s, Δ_t) i.e. the top Δ_t ranked false negatives which can significantly smaller than the number of false negatives. Moreover, as we show below, Perc@k-max also enjoys a mistake bound but with respect to the max-surrogate $\ell_{\text{prec@k}}^{\text{max}}(\cdot)$ which offers a weaker upper bound to the prec@k objective.

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 values when Algorithm 2 is run. Also, for any predictor $\mathbf{w}$, let $\hat{\mathcal{L}}_T^{max}(\mathbf{w}) = \sum_{t=1}^T \ell_{prec@k}^{max}(\mathbf{w}; \mathbf{X}_t, \mathbf{y}_t)$. 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 Algorithm 1, we can give a simplified mistake bound in situations when the separability condition specified by Definition 4 is satisfied.

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

As our *strong* (k, γ)-margin is exactly the same as the standard margin notion for classification, the above bound is equivalent to the standard perceptron mistake bound. However, we observe that in practice, many times Perc@k-max outperforms Perc@k-avg algorithm, even though the later optimizes a tighter surrogate. This hints at a possible loose analysis that fails to exploit some other structure that might be present in the dataset.

4.1. Stochastic Gradient Descent for $\ell^{\text{avg}}_{\text{prec@k}}$

Next, we present a stochastic gradient descent (SGD) algorithm for prec@k loss. SGD methods are known to be very successful for optimizing empirical risk minimization (ERM) problems as they require only a few passes over the data to achieve the optimal statistical accuracy. However, prec@k loss is a non-convex function that is also non-additive over the entire training set. Hence, standard SGD techniques for classification do not apply directly.

By combining our proposed avg-surrogate ($\ell^{\text{avg}}_{\text{prec@k}}$) with mini-batches, we can provide a scalable SGD algorithm for optimizing prec@k (see Algorithm 3). At a high level, our SGD@k-avg algorithm uses mini-batches to update the current model vector using standard gradient descent. As our avg-surrogate is also non-additive over training points, we need to obtain an estimate of the gradient of $\ell^{\text{avg}}_{\text{prec@k}}$ using mini-batches. Algorithm 4 details the gradient calculation for the $\ell^{\text{avg}}_{\text{prec@k}}$ and Algorithm 3 uses the obtained gradient estimates over the current mini-batch to update the model vector ($\mathbf{w}^t$).

Note that the standard analysis of SGD methods (with point-wise loss functions) crucially exploits the fact that the loss functions are additive and hence at each step, we get unbiased estimate of the gradient. Unfortunately, for general non-decomposable loss function, such an unbiased estimate is not possible. So, one needs to show that each mini-batch gives an accurate (but potentially biased) estimate of the gradient. To this end, we need to prove a *uniform convergence bound* for our surrogate functions.

Note that (Kar et al., 2014) also used a similar technique to devise an SGD algorithm for the *struct-SVM* surrogate ($\ell^{\text{struct}}_{\text{prec@k}}$) for prec@k. Naturally, such uniform convergence bounds need to exploit the structure of the loss function, as such bounds are not possible for arbitrary non-decomposable losses. At a high level, we need to show that although $\ell^{\text{avg}}_{\text{prec@k}}$ is non-decomposable over the entire mini-batch, it still is "Lipschitz" in the sense that it does not get perturbed heavily by changing one training point. In the next section, we provide uniform convergence bounds for both of our convex surrogate and use them to provide convergence guarantee for SGD@k-avg.

5. Generalization Bounds

In this section, we provide uniform convergence (UC) bounds for our proposed convex surrogates (avg and max surrogates). We use our novel UC bounds along with the mistake bounds (Theorem 7, 9) to prove two different results: i) precise online-to-batch conversion bounds for the Perc@k-avg and Perc@k-max algorithms, ii) convergence guarantee for the SGD@k-avg algorithm (Algorithm 3).

To present our generalization and convergence bounds, we use normalized versions of prec@k and our proposed surrogates. More specifically, we set $k = \kappa \cdot n_+$ as a fraction of the number of positives. For any scoring function s, its prec@k loss is now denoted as:

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

For uniformity, we will also normalize the surrogate loss functions $\ell^{\text{max}}_{\text{prec@}\kappa}(\cdot)$, $\ell^{\text{avg}}_{\text{prec@}\kappa}(\cdot)$ by dividing throughout by $k = \kappa \cdot n_+$.

Definition 11 (Uniform Convergence)**.** *A performance measure $\Psi : \mathcal{W} \times (\mathcal{X}, \{0,1\})^n \mapsto \mathbb{R}_+$ 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. $1-\delta$,*

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

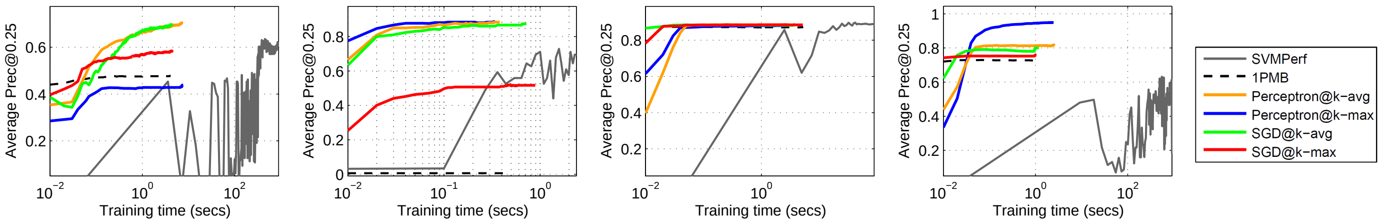
We now provide UC bounds for prec@k as well as all of our surrogates.

Theorem 12. *The performance measure prec@$\kappa(\cdot)$, as well as the surrogates $\ell^{ramp}_{prec@\kappa}(\cdot)$, $\ell^{avg}_{prec@\kappa}(\cdot)$ and $\ell^{max}_{prec@\kappa}(\cdot)$, 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 true positives predicted by the candidate labeling $\hat{\mathbf{y}}$. The above results allow us to establish strong online-to-batch conversion bounds for Perc@k-avg and Perc@k-max, and convergence rate for SGD@k-avg method. For the bounds given below, we shall assume that the points received in the stream for each of our three algorithms are chosen i.i.d. from some fixed population $\mathcal{Z}$.

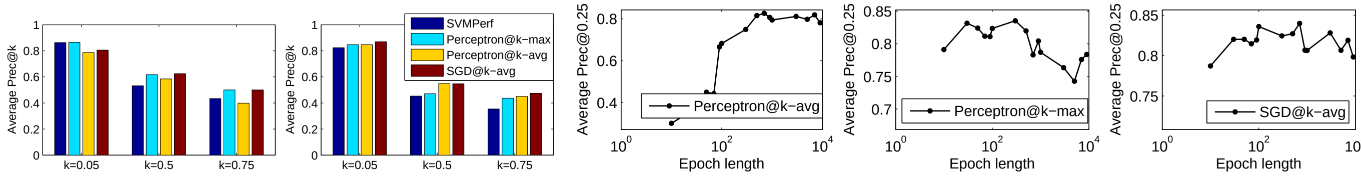
Theorem 13. *Suppose an algorithm, when faced with a random stream of data points, and batch length b, generates an ensemble of classifiers $\mathbf{w}_1,\ldots,\mathbf{w}_T$ which incur a prec@k mistake bound M_T. Then we have, with probability at least $1-\delta$, we have*

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



(a) PPI (b) Letter (c) a9a (d) IJCNN1

Figure 2. Comparison of our perceptron and SGD based methods (Perc@k-avg, Perc@k-max, SGD@k-avg) with baseline methods (SVMPerf, 1PMB) on Prec@0.25 maximization tasks. Clearly, Perc@k-avg, SGD@k-avg (both of which are based on $\ell^{\text{avg}}_{\text{prec@k}}$ loss) are the most consistent methods while accuracies of Perc@k-max can have large variations.



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

Figure 3. (a), (b): Comparison of performance of Prec@k optimization methods for different values of k. (c), (d), (e): Comparison of performance of proposed methods for different epoch lengths on Prec@0.25 maximization tasks

The proof of this theorem follows from the uniform convergence bound for the prec@k$(\cdot)$ performance measure. In particular, combining this with the mistake bound from Theorem 7 and the uniform convergence bound from Theorem 12, ensures the following generalization guarantee for the ensemble generated by Algorithm 1.

Corollary 14. *Suppose we encounter a stream of data points randomly chosen from the population $\mathcal{Z}$ and let $\mathbf{w}^1, \ldots, \mathbf{w}^T$ be the ensemble of classifiers returned by the* PERCEPTRON@K-AVG *algorithm. Then we have, with probability at least $1-\delta$, for any $\mathbf{w}^*$*

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

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

A similar statement holds for the PERCEPTRON@K-MAX algorithm with respect to the $\ell^{\text{max}}_{\text{prec@}\kappa}(\cdot)$ surrogate.

Theorem 15. *Suppose we execute Algorithm 3 with batch length b, then with probability at least $1-\delta$ over the random ordering of the points, for any $\mathbf{w}^* \in \mathcal{W}$, the predictor $\bar{\mathbf{w}}$ returned by the algorithm satisfies*

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

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

6. Experiments

In this section, we apply our methods to several benchmark datasets for rare-class binary classification. The goal of our experimental evaluation is two-fold: a): demonstrate that our surrogate functions which are well-motivated theoretically indeed outperforms struct-SVM surrogate (SVMPerf) in practice as well, b): our SGD based method quickly learns an accurate (in terms of prec@k) classifier. In addition to SVMPerf, we also compare our methods to the 1PMB method by (Kar et al., 2014) which also attempts to optimize prec@k by using the struct-SVM loss. Recall that while struct-SVM loss is intended to optimize prec@k, it does not upper bound prec@k loss (see Section 3) and is not known to be consistent for any interesting margin notion or noise condition.

Implementation Details: For both the baseline methods (SVMPerf and 1PMB), we use the C code provided by the respective authors. Our method is also implemented in C. We randomly split each of the dataset with 70% used for training (out of which 10% was used for validation) and the remaining 30% for testing. All of our results are averaged over 10 random train-test splits.

In the first set of experiments, we evaluated different our methods (Perc@k-avg, Perc@k-max, SGD@k-avg, SGD@k-max) on several benchmark UCI datasets. We use Precision@(0.25) for evaluating the different methods. Figure 2 plots prec@k achieved by different methods vs the training time required by each method (see Appendix H for results on more datasets). Clearly, out of the six methods that we evaluated, SVMPerf (which is based on cutting plane method) is computationally most expensive. Perceptron and SGD methods frequently updates the classifier using a few points, hence, they tend to find reasonably accurate solutions much earlier than the cutting plane methods.

We observe that our avg-surrogate (6) based methods

(Perc@k-avg, SGD@k-avg) are the most consistent methods and achieve nearly the best accuracy for each of the dataset. This also matches our theoretical results which show that avg-surrogate is a tighter convex relaxation of prec@k as compared to the max-surrogate. Moreover, on almost all of the dataset Perc@k-avg and SGD@k-avg are more accurate than SVMPerf and 1PMB. The max-surrogate seems to be the most inconsistent of all the loss functions. For the a9a dataset, it is about 10% more accurate than all the other methods, while for PPI dataset its prec@k is around 10% *less* than the Perc@k-avg method.

Next, we compare Precision@k obtained by different methods with varying k. Here again, we observe that Perc@k-avg, SGD@k-avg consistently outperforms SVMPerf. Finally, we study the effect of the selected epoch-lengths for our methods. We observe that the accuracy for Perc@k-avg increase with larger epochs and then stabilizes at epochs of length around 10K. SGD@k-avg method seems to be more or less invariant to the epoch length, while the accuracy of Perc@k-max method suffers for epochs of length > 1000.

References

Agarwal, S. The Infinite Push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 839—850, 2011.

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

Boyd, Stephen, Cortes, Corinna, Mohri, Mehryar, and Radovanovic, Ana. Accuracy at the top. In *Advances in neural information processing systems*, pp. 953–961, 2012.

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

Cao, Zhe, Qin, Tao, Liu, Tie-Yan, Tsai, Ming-Feng, and Li, Hang. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pp. 129–136. ACM, 2007.

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

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

Chaudhuri, Sougata and Tewari, Ambuj. Online ranking with top-1 feedback. 2015.

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

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

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

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

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

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, 2002.

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

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

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

Li, Nan, Jin, Rong, and Zhou, Zhi-Hua. Top rank optimization in linear time. In *Advances in Neural Information Processing Systems*, pp. 1502–1510, 2014.

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

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

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

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

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

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

Rudin, C. 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.

Shalev-Shwartz, Shai, Singer, Yoram, Srebro, Nathan, and Cotter, Andrew. Pegasos: primal estimated sub-gradient solver for SVM. *Math. Program.*, 127(1):3–30, 2011.

Tsoumakas, Grigorios and Katakis, Ioannis. Multi-label classification: An overview. *IJDWM*, 3(3):1–13, 2007.

Valizadegan, Hamed, Jin, Rong, Zhang, Ruofei, and Mao, Jianchang. Learning to rank by optimizing ndcg measure. In *Advances in neural information processing systems*, pp. 1883–1891, 2009.

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

Yun, Hyokun, Raman, Parameswaran, and Vishwanathan, S. Ranking via robust binary classification. In *Advances in Neural Information Processing Systems*, pp. 2582–2590, 2014.

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

A. Structural SVM Surrogate for prec@k

The structural SVM surrogate for prec@k for a set of n points $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathbb{R}^d \times \{0,1\})^n$ and model $w \in \mathbb{R}^d$ can be written as $\ell^{\text{struct}}_{\text{prec@k}}(w)$:

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

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.

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*

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

Moreover, if for some scoring function s, we have $\ell^{ramp}_{prec@k}(s) \leq \xi$, 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,*

$$\ell^{ramp}_{prec@k}(s) = prec@k(s)$$

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

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

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

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

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

This finishes the proof. □

B.3. Proof of Lemma 16

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 Claim 17

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

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

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 $\ell^{avg}_{prec@k}(\mathbf{w})$ 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 $\ell^{\text{avg}}_{\text{prec@k}}(s)$

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

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

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

$$
\begin{aligned}
&= \Delta(\mathbf{y}, \tilde{\mathbf{y}}) + \sum_{i=1}^{n} s_i(\tilde{\mathbf{y}}_i(1-\mathbf{y}_i) - \mathbf{y}_i(1-\tilde{\mathbf{y}}_i)) + \frac{n_+ - k}{n_+ - K(\mathbf{y}, \tilde{\mathbf{y}})} \sum_{i=1}^{n}(1-\tilde{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= \Delta(\mathbf{y}, \tilde{\mathbf{y}}) + \underbrace{\sum_{i=1}^{n} \tilde{\mathbf{y}}_i(1-\mathbf{y}_i)s_i}_{(A)} - \underbrace{\frac{k - K(\mathbf{y}, \tilde{\mathbf{y}})}{n_+ - K(\mathbf{y}, \tilde{\mathbf{y}})} \sum_{i=1}^{n}(1-\tilde{\mathbf{y}}_i)\mathbf{y}_i s_i}_{(B)} .
\end{aligned}
$$

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,*

$$
\ell^{avg}_{prec@k}(s) = prec@k(s)
$$

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

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

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

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

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

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

$$\Delta_T^C \quad \leq \quad P_T + \hat{\mathcal{L}}_T(\mathbf{w})$$

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

which gives us the desired result upon solving the quadratic inequality[1]. 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

$$\begin{aligned}
\langle \mathbf{w}_{t-1}, \mathbf{v}_t \rangle &= D_t \cdot \sum_{i\in[b]}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i - \sum_{i\in[b]}(1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i \\
&= \Delta_t \left(\underbrace{\frac{1}{\|\mathbf{y}_t\|_1 - K(\mathbf{y}_t, \hat{\mathbf{y}}_t)} \sum_{i\in[b]}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i}_{(A)} - \underbrace{\frac{1}{\Delta_t} \sum_{i\in[b]}(1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i}_{(B)} \right) \\
&\leq 0,
\end{aligned}$$

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 $\ell_{\text{prec@k}}^{\text{avg}}(\mathbf{w}) \geq \text{prec@k}(\mathbf{w}) \geq 0$ for all $\mathbf{w} \in \mathcal{W}$ (by Claim 17), we still get

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

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, $D_t = 1 - \frac{1}{C(\hat{\mathbf{y}})}$.

$$\begin{aligned}
P_t &= P_{t-1} - \sum_{i\in[b]}(1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i + D_t \cdot \sum_{i\in[b]}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \sum_{i\in[b]}(1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i + \left(1 - \frac{1}{C(\hat{\mathbf{y}})}\right) \sum_{i\in[b]}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i
\end{aligned}$$

[1] 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 $x \geq l$.

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

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

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

This concludes the proof of the mistake bound. □

C.2. Proof of Theorem 9

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

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

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*

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

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 $s_i = \mathbf{w}^\top \mathbf{x}^i_t$. 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 $\mathbf{w}^{t-1}$.

$$
\begin{aligned}
P_t &= P_{t-1} - \sum_{i\in[b]} (1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i + \sum_{i\in S_t} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \sum_{i\in[b]} (1-\mathbf{y}_i)\hat{\mathbf{y}}_i s_i - \sum_{i\in[b]} \mathbf{y}_i\hat{\mathbf{y}}_i s_i + \sum_{i\in[b]} \mathbf{y}_i\hat{\mathbf{y}}_i s_i + \sum_{i\in S_t} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i \\
&= P_{t-1} - \sum_{i\in[b]} \hat{\mathbf{y}}_i s_i + \sum_{i\in[b]} \mathbf{y}_i\hat{\mathbf{y}}_i s_i + \sum_{i\in S_t} (1-\hat{\mathbf{y}}_i)\mathbf{y}_i s_i
\end{aligned}
$$

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

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 concludes the proof of the theorem. □

D. Proof of Lemma 22

Lemma 22. *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.

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 21. *Let $\Psi : \mathcal{W} \to \mathbb{R}$ 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 22 tells us that $\Phi(\cdot)$ is 1-Lipschitz with respect to the $\left\|\cdot\right\|_\infty$ 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 $\left\|\mathbf{x}_i\right\|_2 \leq 1$ for all i, we have

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

Thus we have

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

which gives us the desired result. □

E. Proof of Lemma 23

Lemma 23. *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 $\mathbb{T}(\mathbf{v}) := \mathbb{I}\left[\mathbf{v} \succeq \mathbf{v}^*_n\right]$ and $\hat{\mathbb{T}}(\mathbf{v}) := \mathbb{I}\left[\mathbf{v} \succeq \mathbf{v}^*_b\right]$ (note that $\mathbb{I}\left[E\right]$ 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. □

F. Proof of Theorem 12

Our proof of Theorem 12 crucially utilize 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 22. *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.

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 23. *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 $h : \mathcal{V} \rightarrow [-1, 1]$ 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 prec@$\kappa(\cdot)$, as well as the surrogates $\ell^{\text{ramp}}_{\text{prec@}\kappa}(\cdot)$, $\ell^{\text{avg}}_{\text{prec@}\kappa}(\cdot)$ and $\ell^{\text{max}}_{\text{prec@}\kappa}(\cdot)$, 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.

F.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 23. 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 23 gives us the following result

Lemma 24. *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 23. 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

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

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 p n}\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 p n}\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 p n}\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 23 to show that

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

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

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

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

F.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 21 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

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

An application of Lemma 23 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 $\mathbf{x}$.

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

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

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

which concludes the uniform convergence proof.

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

The proof follows similarly here with a direct application of Corollary 21 showing us that $\Psi_2(\cdot)$ is Lipschitz and an application of Lemma 23 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 26. *We have, with probability at least $1-\delta$ over the choice of b samples,*

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

F.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

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

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

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

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

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 27. *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 28. *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 29. *With probability at least* $1-\delta$ *over the choice of the samples, we have*

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

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 $\mathbf{w}\in\mathcal{W}$:

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

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

where the first step follows from Lemma 27, 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 28 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 28 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 27 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 27 and 28 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 27. We have, by the above simplification,

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

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

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

$$\le \frac{(1-\kappa)\left|\beta-\beta'\right|}{\kappa(1-\beta)(1-\beta')} \le \frac{1}{\kappa(1-\kappa)}\left|\beta-\beta'\right|,$$

where the last step follows since $\beta, \beta' \le \kappa$. 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 $\left(1 - \frac{1-\kappa}{1-\beta}\right) \ge 0$ which means that the βpn top ranked positives will get assigned a label of 1.

To formalize this, let us set some notation. Let $s_1^+ \ge s_2^+ \ge \ldots \ge s_{pn}^+$ denote the scores of the positive points arranged in descending order. Similarly, let $s_1^- \ge s_2^- \ge \ldots \ge s_{(1-p)n}^-$ 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 $\left|s_i^+\right|, \left|s_i^-\right| \le 1$, we have,

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

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

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

which finishes the proof. □

Proof of Lemma 28. 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 26 can be used to conclude that with probability at least $1-\delta$, we have

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

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 21. To analyze its pointwise convergence property

Thus the function $B(\mathbf{w}, \beta)$, as analyzed in the proof of Lemma 27, 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 23, as used in the proof of Theorem 26, 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 $\ell^{\text{avg}}_{\text{prec}@\kappa}(\cdot)$. □

F.4. A Uniform Convergence Bound for the $\ell^{\textbf{max}}_{\textbf{prec}@\kappa}(\cdot)$ 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.

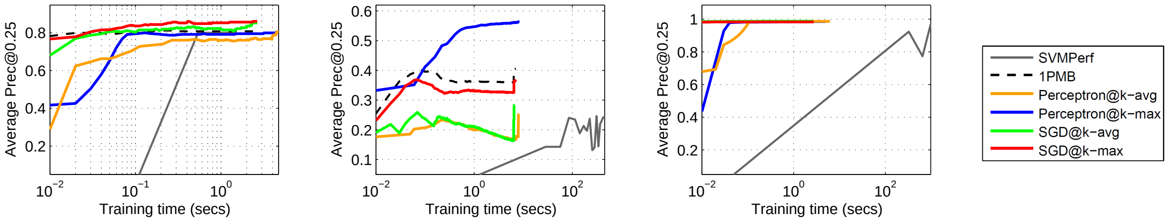
G. Proof of Theorem 15

Theorem 15. *Suppose we execute Algorithm 3 with batch length b, then with probability at least $1 - \delta$ over the random ordering of the points, for any $\mathbf{w}^* \in \mathcal{W}$, the predictor $\bar{\mathbf{w}}$ returned by the algorithm satisfies*

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

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. □

H. Additional Empirical Results



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

Figure 4. Comparison of proposed methods with baselines on Prec@0.25 maximization tasks.