A. Structural SVM Surrogate for prec@k

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

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

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

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

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

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

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

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

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

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

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

For $w' = -w^*$,

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

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

B. Proofs of Claims from Section 3

B.1. Proof of Claim 1

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

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

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

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

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

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

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

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

which proves the claim. □

B.2. Proof of Claim 3

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

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

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

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

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

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

Case 2 ($K(\mathbf{y}, \hat{\mathbf{y}}) = k' < k$): In this case let S^* be the set of k top ranked positive points according to the scoring function s. Also let S^*_1 be the set of $k' (= K(\mathbf{y}, \hat{\mathbf{y}}))$ top ranked positives and let $S^*_2 = S^* \backslash S^*_1$. Then we have

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

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

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

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

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

B.3. A Useful Supplementary Lemma

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

Thus we have

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

B.5. Proof of Claim 6

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

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

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

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

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

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

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

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

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

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

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

as well as

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

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

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

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

C. Proofs from Section 4

C.1. Proof of Theorem 7

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

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

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

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

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

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

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

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

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

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

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

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 $\mathbf{v}_t = D_t \cdot \sum_{i\in[b]}(1-\hat{\mathbf{y}}_i)\mathbf{y}_i \cdot \mathbf{x}^i_t - \sum_{i\in[b]}(1-\mathbf{y}_i)\hat{\mathbf{y}}_i \cdot \mathbf{x}^i_t$ 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 $s_i = \mathbf{w}^\top_{t-1}\mathbf{x}^i_t$. 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

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

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

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

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

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

as required.

[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 $x \geq l$.

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

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

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

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

This concludes the proof of the mistake bound. □

C.2. Proof of Theorem 9

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

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

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

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

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

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

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

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

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

as required.

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

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

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

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

This concludes the proof of the theorem. □

D. Proof of Theorem 12

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We can now write

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

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

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

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

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

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

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

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

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

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

We have

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

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

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

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

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

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

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

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

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

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

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

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

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

which concludes the uniform convergence proof.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which correspond, respectively, to classifiers for which the *number* of true positives equals $\{0, 1, 2 \ldots \kappa p n - 1, \kappa p n\}$. 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which finishes the proof. □

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

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

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

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

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

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

D.4. Proof of Lemma 21

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

then we have

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

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

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

Thus we have

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

which gives us the desired result. □

D.5. Proof of Lemma 22

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

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

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. □

D.6. 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.

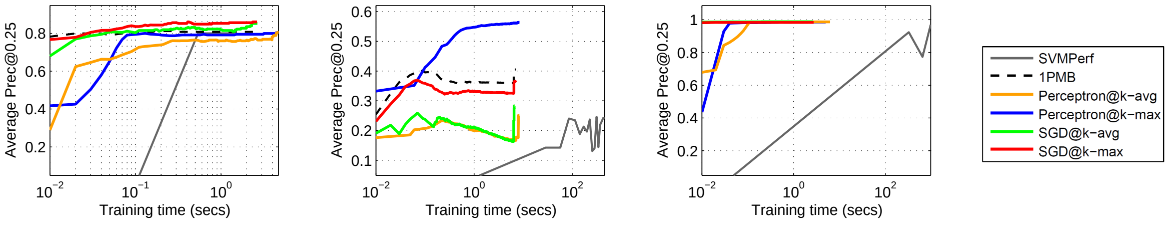
E. Proof of Theorem 15

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

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

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

F. Additional Empirical Results



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

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