Appendices

A. Proofs of Section 4.1

Proof of Theorem 1. For any positive x^+ and negative x^-, $\mathbb{1}(w^T x^+ \le w^T x^-)$ is monotonically decreasing in $w^T x^+$ assigned to x^+. Thus the quantity $\frac{1}{\beta}\sum_{x^+ \in Z_+} \mathbb{1}(w^T x^+ \le w^T x^-)$ for any $x^- \in Z_-$ is minimized by top ranked β positives. Define $r(w; x^+_{(1)_w}, \dots x^+_{(\beta)_w}, x^-) \coloneqq \frac{1}{\beta}\sum_{i=1}^{\beta} \mathbb{1}(w^T x^+_{(i)_w} \le w^T x^-)$. This function is monotonically increasing in $w^T x^-$ assigned to x^-. Hence,

$$\widehat{R}_{\text{pAp@k}}(w; x^+_{(1)_w}, \dots x^+_{(\beta)_w}, Z_-) = \frac{1}{k}\sum_{x^- \in Z_-} r(w, x^+_{(1)_w}, \dots x^+_{(\beta)_w}, x^-)$$

is maximized by top k negatives. Therefore,

$$\begin{aligned}\widehat{R}_{\text{pAp@k}}(w; S) &= \max_{\substack{Z_- \subseteq S_-,\\ |Z_-|=k}} \min_{\substack{Z_+ \subseteq S_+,\\ |Z_+|=\beta}} \frac{1}{\beta k}\sum_{x^+ \in Z_+}\sum_{x^- \in Z_-} \mathbb{1}(w^T x^+ \le w^T x^-)\\ &= \max_{\substack{Z_- \subseteq S_-,\\ |Z_-|=k}} \min_{\substack{Z_+ \subseteq S_+,\\ |Z_+|=\beta}} \widehat{R}_{AUC}(w; Z_+, Z_-),\end{aligned}$$

where $\widehat{R}_{AUC}(w; Z_+, Z_-)$ is the AUC between the subset of positives Z_+ and the subset of negatives Z_-. By using the same argument first on negatives and then on positives, we get

$$\widehat{R}_{\text{pAp@k}}(w; S) = \min_{\substack{Z_+ \subseteq S_+,\\ |Z_+|=\beta}} \max_{\substack{Z_- \subseteq S_-,\\ |Z_-|=k}} \widehat{R}_{AUC}(w; Z_+, Z_-),$$

as well. Hence, the order of the min-max over subsets Z_- and Z_+ does not affect pAp@k and can be interchanged. □

Lemma 1. *Let $\bar{Z}_+ = \{\bar{z}_1^+, \dots, \bar{z}_\beta^+\}$ and $\bar{Z}_- = \{\bar{z}_1^-, \dots, \bar{z}_k^-\}$ be the set of instances in the top β and top k positions in the ranking of positive instances and negative instances (in descending order of scores) by $w^T x$, respectively. Then the (outer) maximum of the minimum value in Eq. (6) is attained at $\bar{Z}_+$ and $\bar{Z}_-$.*

Proof. By expanding (6), we have that

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w; S) = \max_{\substack{Z_- \subseteq S_-\\ |Z_-|=k}} \min_{\substack{Z_+ \subseteq S_+\\ |Z_+|=\beta}} \max_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} + \sum_{j=1}^{k} q_j w^T z_j^- - \sum_{i=1}^{\beta} p_i w^T z_i^+\right], \tag{18}$$

where $p_i = \sum_{j=1}^{k}\pi_{i,j} \ge 0$ and $q_j = \sum_{i=1}^{\beta}\pi_{i,j} \ge 0$. For any subset of positive instances $Z_+ = \{z_1^+, \dots, z_\beta^+\} \subseteq S_+$, we assume w.l.o.g. that $w^T z_1^+ \ge \cdots \ge w^T z_\beta^+$ (this ensures that the identity of each z_i^+ is unique). Similarly, for any subset of negative instances $Z_- = \{z_1^-, \dots, z_k^-\} \subseteq S_-$, we assume w.l.o.g. that $w^T z_1^- \ge \cdots \ge w^T z_k^-$. Notice that in (18), only the last term depends on the subset Z_+. Moreover, in general, the min and the max (over π) cannot be exchanged; however, notice that since p_i's are always non-negative for any $\pi \in \Pi_{\beta \times k}$ and set $Z_- \subseteq S_-$, we may push the minimum inside as shown below:

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w; S) = \max_{\substack{Z_- \subseteq S_-\\ |Z_-|=k}} \max_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} + \sum_{j=1}^{k} q_j w^T z_j^- - \max_{\substack{Z_+ \subseteq S_+\\ |Z_+|=\beta}} \sum_{i=1}^{\beta} p_i w^T z_i^+\right].$$

For any fixed π (or equivalently p_i's), the last term above is maximized when the subset Z_+ contains the positives with the highest scores, and in particular, the top β ranked positives by w, denoted by $\bar{Z}_+ = \{\bar{z}_1^+, \dots, \bar{z}_\beta^+\}$. Similarly, notice that in (18), only the second term depends on the subset Z_-. Therefore, Eq.(18) can be written as follows:

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w; S) = \max_{\pi \in \Pi_{\beta \times k}} \left[\frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} - \frac{1}{\beta k}\sum_{i=1}^{\beta} p_i w^T \bar{z}_i^+ + \max_{\substack{Z_- \subseteq S_-\\ |Z_-| = k}} \frac{1}{\beta k}\sum_{j=1}^{k} q_j w^T z_j^-\right].$$

Since $q_j \geq 0$ for all j for any fixed π (or equivalently q_j's), the last term above is maximized when the subset Z_- contains the negative with the highest scores, and in particular, the top k ranked negatives by w. □

Proof of Proposition 1. Since the term in (4) upper bounds the AUC performance measure between the set of positives S_+ and the set of negatives S_-, the ramp surrogate (6), by construction, upper bounds the pAp@k metric.

Moreover, since $p_i, q_j \geq 0$ for all i, j in (18), we observe from Lemma 1 that the set of positives and the set of negatives selected are $\bar{Z}_+ = \{\bar{z}_1^+, \dots z_\beta^+\} = \{z^+_{(1)_w}, \dots z^+_{(\beta)_w}\}$ i.e. the top β positives and $\bar{Z}_- = \{\bar{z}_1^-, \dots z_k^-\} = \{z^-_{(1)_w}, \dots z^-_{(k)_w}\}$ i.e. the top k negatives, respectively, ranked in decreasing order by $w^T x$. The ramp loss can be written as:

$$\begin{aligned}
\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w; S) &= \max_{\pi \in \Pi_{\beta \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} - \frac{1}{\beta k} \sum_{i=1}^{\beta} p_{(i)_w} w^T z^+_{(i)_w} + \frac{1}{\beta k} \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w} \right] \\
&= \max_{\pi \in \Pi_{\beta \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} - \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} w^T z^+_{(i)_w} + \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} w^T z^-_{(j)_w} \right] \\
&= \max_{\pi \in \Pi_{\beta \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} [1 - w^T (z^+_{(i)_w} - w^T z^-_{(j)_w})] \right], \qquad (19)
\end{aligned}$$

where the ramp loss in the first step is another way to write eq. (18) when the subset $\bar{Z}_+$ and $\bar{Z}_-$ are chosen. From (19), it is easy to find the optimum $\bar{\pi}$, i.e., $\bar{\pi}_{(i)_w,(j)_w} = 1$ if $(1 - w^T(\bar{z}^+_{(i)_w} - \bar{z}^-_{(j)_w})) \geq 0$, and 0 otherwise. Therefore, the ramp surrogate is 0 iff the weak β-margin condition (Definition 2) holds in the data, i.e., there is set of β positives which are separated by negatives by a margin of 1. □

Proof of Proposition 2. One simple way to construct a surrogate for the pAp@k metric is by replacing the indicator function in (1) by a convex, monotone, Lipschitz, classification surrogate, e.g., the hinge surrogate:

$$\widehat{R}^{\text{hinge}}_{\text{pAp@k}}(w; S) := \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} (1 - (w^T x^+_{(i)_w} - w^T x^-_{(j)_w}))_+, \qquad (20)$$

where $x^+_{(i)_w}$ and $x^-_{(j)_w}$ denotes the positive and negative instances in S_+ and S_- ranked in i-th and j-th position (among positives and negatives, in decreasing order of scores) by w, respectively. From (19), we observe that the hinge loss surrogate is equal to the ramp surrogate for the pAp@k metric, i.e.

$$\widehat{R}^{\text{hinge}}_{\text{pAp@k}}(w; S) = \widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w; S).$$

When $n_+ \leq k$, we consider all the positives in the data, and as discussed, the pAp@k metric reduces to partial-AUC with false positive range being $[0, \frac{k}{n_-}]$. From Theorem 3 of (Narasimhan & Agarwal, 2017), the above hinge loss based surrogate (equiv. the ramp surrogate) becomes convex.

However, when $n_+ > k$, then the above surrogate is non-convex. We provide a counter example to support our claim. Consider the following 2-dimensional example. Let $k = 2$, $w_1 = [-1, 0], w_2 = [0, -1]$, and $\lambda = 0.5$. We take $\widetilde{w} = \lambda w_1 + (1-\lambda) w_2$. Consider the feature matrix $x = [[-1, 0], [-1, -1], [1, 0], [1, 0], [0, 1]]$ and the true labels $y = [0, 0, 1, 1, 1]$. We can observe that in this case, $\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(\widetilde{w}; S) = 2.25$, $\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w_1; S) = 2.5$, and $\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w_2; S) = 1.5$. This means that $\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(\widetilde{w}; S) > \lambda \widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w_1; S) + (1-\lambda) \widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w_2; S)$. Thus, the ramp surrogate (hinge surrogate) is non-convex. □

B. Proofs of Section 4.2, Section 4.3, and Section 4.4

Proof of Proposition 3. The average surrogate is constructed by replacing the last maximum in (9) by average over all the subsets. Thus, the average surrogate upper bounds the ramp surrogate and hence the metric pAp@k. For conditional consistency, let us recall the definition of average surrogate:

$$\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w; S) = \max_{\substack{Z_- \subseteq S_- \\ |Z_-| = k}} \max_{\pi \in \Pi_{\beta \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} w^T z_j^- - \frac{1}{\beta k} \left(\frac{1}{n_+} \sum_{l=1}^{n_+} w^T x_l^+ \right) \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} \right].$$

The maximum over finite sets can be interchanged. Thus, we can write the average surrogate as follows:

$$\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) = \max_{\pi\in\Pi_{\beta\times k}} \left[\frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} - \frac{1}{\beta k}\left(\frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+\right)\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} + \max_{\substack{Z_-\subseteq S_-\\ |Z_-|=k}} \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} w^T z_j^-\right].$$

Similar to the proof of Lemma 1, the inner maximum is achieved by the top-k negatives as scored by the scoring function $w^T x$. Thus,

$$\begin{aligned}
\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) &= \max_{\pi\in\Pi_{\beta\times k}} \left[\frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} + \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,(j)_w} w^T z^-_{(j)_w} - \frac{1}{\beta k}\left(\frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+\right)\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j}\right] \\
&= \max_{\pi\in\Pi_{\beta\times k}} \left[\frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,(j)_w} + \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,(j)_w} w^T z^-_{(j)_w} - \frac{1}{\beta k}\left(\frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+\right)\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,(j)_w}\right] \\
&= \max_{\pi\in\Pi_{\beta\times k}} \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,(j)_w}\left[1 + w^T z^-_{(j)_w} - \left(\frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+\right)\right] \\
&= \max_{\pi\in\Pi_{\beta\times k}} \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,(j)_w}\left[1 + w^T \bar{z}^-_{(j)_w} - \left(\frac{(n_+-\beta)!\beta!}{n_+!}\sum_{\widetilde{Z}_+\in\mathcal{Z}_{uo}}\frac{1}{\beta}\sum_{i=1}^{\beta} w^T x_{i\in\widetilde{Z}_+}\right)\right],
\end{aligned}$$

where $\mathcal{Z}_{uo}$ is the set of all (unordered) sets of positives of size β, and the last step follows from the fact that average score of all the positives is equal to the average of the mean over all subsets of size β. From the above equation, it is easy to find the optimum $\bar{\pi}\in\Pi_{\beta\times k}$, i.e., $\bar{\pi}_{i,(j)_w} = 1$ if $(1 + w^T z^-_{(j)_w} - \frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+) >= 0$, and 0 otherwise. Clearly, if there exists a scoring function w which satisfies the β-margin condition, then $\widehat{R}_{\text{pAp@k}}(w;S) = \widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) = 0$. □

Before proving Proposition 4, we prove the following Lemma which states that that the argument maximum over $\pi\in\Pi_{n_+\times k}$ in TS surrogate can be found in a restricted space of ordering matrices where any two positives separated by a negative are sorted in decreasing order of scores by $w^T x$. To this end, recall that the TS surrogate is defined as:

$$\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S) := \max_{\substack{Z_-\subseteq S_-\\ |Z_-|=k}} \max_{\pi\in\Pi_{n_+\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{(i)_\pi,j} - \sum_{i=1}^{n_+} p_i w^T x_i^+ + \sum_{j=1}^{k} q_j w^T z_j^-\right],$$

where $p_i = \sum_{j=1}^{k}\pi_{i,j}$ and $q_j = \sum_{i=1}^{n_+}\pi_{i,j}$. Similar to the proof of Proposition 3, we observe that the argmax over $Z_-\subseteq S_-, |Z_-| = k$ is attained at the top-k negatives $\overline{Z}_-$ according to $w^T x$, and the combinatorial optimization problem over $\Pi_{n_+\times k}$ becomes equivalent to:

$$\max_{\pi\in\Pi_{n_+\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{(i)_\pi,j} - \sum_{i=1}^{n_+} p_i w^T x_i^+ + \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w}\right]. \tag{OP1}$$

Notice that the top-β positives in the first term above are different for different ordering matrices $\pi\in\Pi_{n_+\times k}$. Hence, one may simplify the above optimization problem and search the argmax, for any given $w\in\mathbb{R}^d$, over a restricted space of ordering matrices defined as:

$$\Pi^w_{n_+\times k} = \{\pi\in\Pi_{n_+\times k} \mid \forall\ j, i_1 < i_2 : \pi_{(i_1)_w,j} \le \pi_{(i_2)_w,j}\}, \tag{21}$$

where $(i)_w$ denotes the index of the i-th ranked positive (among the positive instances) , when the instances are sorted in descending order by $w^T x$. The set in (21) is the set of all ordering matrices in which any two positive instances that are separated by a negative instance are sorted according to the score $w^T x$ in decreasing order. Notice that this approach of searching the optimum in the restricted search space is similar to the approach for optimizing $\widehat{R}^{\text{tight}}_{\text{pAUC}}(\gamma, \delta)$ surrogate, where $0 < \gamma < \delta < 1$, for the pAUC performance measure (Narasimhan & Agarwal, 2017) in the false positive range (γ, δ). This is expected since both $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S)$ for pAp@k and $\widehat{R}^{\text{tight}}_{\text{pAUC}}$ for pAUC take features from all the positives after restricting to (top-k) negatives. The difference is that the $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S)$ further restricts ordering of positives. Thus the solutions are entirely different as we seek positives to be further restricted in a certain order. Restricting our search to $\Pi^w_{n_+ \times k}$, we have:

Lemma 2. *The solution $\bar{\pi}$ to the optimization problem* (OP1) *lies in $\Pi^w_{n_+ \times k}$.*

Proof. The proof follows by contradiction. Let us suppose that $\bar{\pi} \notin \Pi^w_{n_+ \times k}$. Then, $\exists j, i_1 < i_2$ such that $\bar{\pi}_{(i_1)_w, j} > \bar{\pi}_{(i_2)_w, j}$, which essentially means that $\pi_{(i_2)_w, j} = 0$ and $\pi_{(i_1)_w, j} = 1$. That is, $\bar{\pi}$ ranks $x^+_{(i_2)_w}$ above x^-_j, which is further ranked above than $x^+_{(i_1)_w}$, and hence the number of negatives above the $(i_1)_w$-th positive is greater than the number of negatives above the $(i_2)_w$-th positive, i.e. $\bar{p}_{(i_1)_w} \geq \bar{p}_{(i_2)_w}$. Now let us construct another ordering π' in which the instances $x^+_{(i_1)_w}$ and $x^+_{(i_2)_w}$ are swapped, i.e. for all j' with $\bar{\pi}_{(i_1)_w, j'} = 1$ and $\bar{\pi}_{(i_2)_w, j'} = 0$, we set $\pi'_{(i_1)_w, j'} = 0$ and $\pi'_{(i_2)_w, j'} = 1$. This would entail that $p'_{(i_1)_w} \leq p'_{(i_2)_w}$. Then it can be seen that the loss term (first term in (OP1)) is the same for $\bar{\pi}$ as for π'. Define $\bar{\Delta} := -(\bar{p}_{(i_1)_w} w^T x^+_{(i_1)_w} + \bar{p}_{(i_2)_w} w^T x^+_{(i_2)_w})$ and $\Delta' := -(p'_{(i_1)_w} w^T x^+_{(i_2)_w} + p'_{(i_2)_w} w^T x^+_{(i_1)_w})$. Clearly, we have $\bar{\Delta}' \geq \bar{\Delta}$ since $w^T x^+_{(i_1)_w} \geq w^T x^+_{(i_2)_w}$. Therefore, the second term in (OP1) increases leading to a higher objective value in (OP1). This contradicts the fact that $\bar{\pi}$ is a maximizer. □

Given Lemma 2, we may now prove Proposition (4).

Proof of Proposition 4. As discussed in Section 4.3, the TS surrogate upper bounds the pAp@k metric by construction. For consistency, let us recall that the definition of TS surrogate:

$$\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S) := \max_{\substack{Z_- \subseteq S_- \\ |Z_-| = k}} \max_{\pi \in \Pi_{n_+ \times k}} \frac{1}{\beta k} \left[\sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_\pi, j} - \sum_{i=1}^{n_+} p_i w^T x_i^+ + \sum_{j=1}^{k} q_j w^T z_j^- \right].$$

Again the outer max is satisfied by the top-k negatives. Moreover, using Lemma 2, we may write the TS surrogate as:

$$\begin{aligned} \widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S) &= \max_{\pi \in \Pi^w_{n_+ \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_\pi, j} - \frac{1}{\beta k} \sum_{i=1}^{n_+} p_i w^T x_i^+ + \frac{1}{\beta k} \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w} \right] \\ &= \max_{\pi \in \Pi^w_{n_+ \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_\pi, (j)_w} - \frac{1}{\beta k} \sum_{i=1}^{n_+} p_i w^T x_i^+ + \frac{1}{\beta k} \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w} \right]. \end{aligned}$$

Now observe that for any $\pi \in \Pi^w_{n_+ \times k}$, $\pi_{(i)_\pi, j} = \pi_{(i)_w, j}$, as there always exists an ordering consistent with π in which all the positives are sorted in decreasing order by the score function $w^T x$. This means that the ordering of positives is also fixed, now we must fill the entries of the matrix π. So, our objective is to solve the following:

$$\begin{aligned} &\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S) \\ &= \max_{\pi \in \Pi^w_{n_+ \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w, (j)_w} - \frac{1}{\beta k} \sum_{i=1}^{n_+} p_i w^T x_i^+ + \frac{1}{\beta k} \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w} \right] \\ &= \max_{\pi \in \Pi^w_{n_+ \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w, (j)_w} - \frac{1}{\beta k} \sum_{i=1}^{n_+} \sum_{j=1}^{k} \pi_{(i)_w, (j)_w} w^T x^+_{(i)_w} + \frac{1}{\beta k} \sum_{i=1}^{n_+} \sum_{j=1}^{k} \pi_{(i)_w, (j)_w} w^T z^-_{(j)_w} \right] \\ &= \max_{\pi \in \Pi^w_{n_+ \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w, (j)_w} [1 - w^T x^+_{(i)_w} + w^T z^-_{(j)_w}] + \frac{1}{\beta k} \sum_{i=\beta+1}^{n_+} \sum_{j=1}^{k} \pi_{(i)_w, (j)_w} [-w^T x^+_{(i)_w} + w^T z^-_{(j)_w}] \right] \qquad (22) \end{aligned}$$

The second term goes to zero if all the positives are just separated by negatives, and the first term goes to zero if the top-β positives further outrank all the negatives by a margin of 1. Thus, in conclusion, the TS surrogate $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S) = 0$ under the moderate β-margin condition. □

Proof of Proposition 5. By construction, the ramp surrogate (6) upper bounds the pAp@k metric (1). Moreover, by writing ramp surrogate as in equation (9) and using the first inequality in the following

$$\max_{\substack{Z_+ \subseteq S_+ \\ |Z_+| = \beta}} \sum_{i=1}^{\beta} p_i w^T z_i^+ \geq \frac{(n_+ - \beta)!}{n_+!} \sum_{\widetilde{Z}_+ \in \mathcal{Z}} \sum_{i=1}^{\beta} p_i w^T z^+_{i \in \widetilde{Z}} \geq \min_{\substack{Z_+ \subseteq S_+ \\ |Z_+| = \beta}} \sum_{i=1}^{\beta} p_i w^T z_i^+,$$

where $\mathcal{Z}$ is the set of all ordered sets of size β, we construct the average surrogate (10). Since maximum is greater than the mean, the average surrogate upper bounds ramp surrogate. Lastly, since mean is greater than the minimum, the max surrogate (defined in Appendix C, equation (24)) upper bounds the average surrogate. This establishes the following:

$$\widehat{R}_{\text{pAp@k}}(w;S) \leq \widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) \leq \widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) \leq \widehat{R}^{\text{max}}_{\text{pAp@k}}(w;S).$$

Now, from (22), it is easy to write the TS surrogate as:

$$\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S) = \widehat{R}^{\text{ramp}}_{\text{pAp@k}} + \max_{\pi \in \Pi^w_{(n_+ - \beta) \times k}} \left[\frac{1}{\beta k} \sum_{i=\beta+1}^{n_+} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} [-w^T x^+_{(i)_w} + w^T z^-_{(j)_w}] \right].$$

The second term on the right hand side above is non-negative, hence we establish that

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) \leq \widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S).$$

□

C. The Max Surrogate for pAp@k

Let us take the form of ramp surrogate as defined in (9), i.e.:

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_- \subseteq S_- \\ |Z_-| = k}} \max_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k} \left[\sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \sum_{j=1}^{k} q_j w^T z_j^- - \max_{\substack{Z_+ \subseteq S_+ \\ |Z_+| = \beta}} \sum_{i=1}^{\beta} p_i w^T z_i^+ \right]. \tag{23}$$

Since $p_i \geq 0 \; \forall \; i$, the last term is maximized by the top-β positives in descending order of scores by $w^T x$. Thus by replacing the maximum over Z_+ in (23) by a minimum over Z_+ and pushing that outside, we relax the ramp surrogate in order to obtain the max surrogate defined below:

$$\begin{aligned} \widehat{R}^{\text{max}}_{\text{pAp@k}}(w;S) &= \max_{\substack{Z_- \subseteq S_- \\ |Z_-| = k}} \max_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k} \left[\sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \sum_{j=1}^{k} q_j w^T z_j^- - \min_{\substack{Z_+ \subseteq S_+ \\ |Z_+| = \beta}} \sum_{i=1}^{\beta} p_i w^T z_i^+ \right] \\ &= \max_{\substack{Z_- \subseteq S_- \\ |Z_-| = k}} \max_{\substack{Z_+ \subseteq S_+ \\ |Z_+| = \beta}} \max_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k} \left[\sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} w^T z_j^- - \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} w^T z_i^+ \right]. \end{aligned} \tag{24}$$

This surrogate is a point-wise maximum over convex functions in w, thus it is convex. It also upper bounds pAp@k, since it upper bounds the ramp surrogate. This surrogate is consistent w.r.t. pAp@k under the strong (β, δ)-margin condition defined as follows:

Definition 5. ***Strong*** (β, δ)***-margin*** *(Kar et al., 2015)**:** A dataset S satisfies the strong (β, δ)-margin condition if for some scoring function f,*

$$\min_{i \in S_+} f_i - \max_{j \in S_-} f_j \geq \delta. \tag{25}$$

We say that f realizes this margin. We refer the strong $(\beta, 1)$-margin condition as simply the strong β-margin condition.

Proposition 6. *For any scoring function $w^T x$, we have $\widehat{R}^{\text{max}}_{\text{pAp@k}}(w;S) \geq \widehat{R}_{\text{pAp@k}}(w;S)$. Moreover, if the scoring function $w^T x$ realizes the strong β-margin condition over a dataset S, then $\widehat{R}^{\text{max}}_{\text{pAp@k}}(w;S) = \widehat{R}_{\text{pAp@k}}(w;S) = 0$.*

Proof. From (24), we have:

$$\widehat{R}^{\max}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_- \subseteq S_- \\ |Z_-|=k}} \max_{\pi \in \Pi_{\beta\times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \frac{1}{\beta k} \sum_{j=1}^{k} q_j w^T z_j^- - \min_{\substack{Z_+ \subseteq S_+ \\ |Z_+|=\beta}} \frac{1}{\beta k} \sum_{i=1}^{\beta} p_i w^T z_i^+ \right].$$

Since $p_i \geq 0$ for all i, the min in the last term is minimized by the bottom β positives according to the score function $w^T x$. Let us denote the set of bottom β positives by $\underline{Z}_+ = \{z^+_{(n_+-\beta+1)_w}, \ldots, z^+_{(n_+)_w}\}$. Also, only the second term depends on the set of negatives and since q_j's are non-negative for all j, the objective is maximized by the top-k negatives $\bar{Z}_- = \{z^-_{(1)_w}, \ldots, z^-_{(k)_w}\}$ according to the score function $w^T x$. So, $\pi \in \Pi_{\beta\times k}$ only measures the relative ordering of the bottom β positives and the top k negatives as discussed below:

$$\begin{aligned}
\widehat{R}^{\max}_{\text{pAp@k}}(w;S) &= \max_{\pi \in \Pi_{\beta\times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \frac{1}{\beta k} \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w} - \frac{1}{\beta k} \sum_{i=1}^{\beta} p_{(n_+-\beta+i)_w} w^T z^+_{(n_+-\beta+i)_w} \right] \\
&= \max_{\pi \in \Pi_{\beta\times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(n_+-\beta+i)_w,(j)_w} + \frac{1}{\beta k} \sum_{j=1}^{k} q_{(j)_w} w^T z^-_{(j)_w} - \frac{1}{\beta k} \sum_{i=1}^{\beta} p_{(n_+-\beta+i)_w} w^T z^+_{(n_+-\beta+i)_w} \right] \\
&= \max_{\pi \in \Pi_{\beta\times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(n_+-\beta+i)_w,(j)_w} [1 - w^T (z^+_{(n_+-\beta+i)_w} - z^-_{(j)_w})] \right] \qquad (26)
\end{aligned}$$

It is easy to see that $\bar{\pi}_{(n_+-\beta+i)_w,(j)_w} = \mathbb{1}[1 - w^T(z^-_{(n_+-\beta+i)_w} - z^+_{(j)_w}) \geq 0]$, and under strong (β, δ)-margin condition (Definition 5), the upper bounding surrogate $\widehat{R}^{\max}_{\text{pAp@k}}(w;S) = \widehat{R}_{\text{pAp@k}}(w;S) = 0$. □

D. Subgradients of the Proposed Surrogates for pAp@k

In this section, we discuss the subgradients of the proposed surrogates. This involves first finding the argument maximum over the subsets Z_-, Z_+ (if any) and the ordering matrix π, and then computing the gradient step w.r.t the model w. We will also mention some interesting observations regarding the surrogates while computing their subgradients.

D.1. Subgradient of $\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S)$

In $\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S)$ (10), the arguments over which the maximum is searched are restricted to Z_- and π. Following the same arguments as in the proof of Proposition 3, we observe that the argmax over $Z_- \subseteq S_-, |Z_-| = k$ is attained at the top-k negatives $\overline{Z}_-$ according to the score function $w^T x$ (line 1 of Algorithm 2). Moreover, the remaining combinatorial optimization problem becomes equivalent to:

$$\arg\max_{\pi \in \Pi_{\beta\times k}} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} \left[1 - w^T \left(\frac{1}{n_+} \sum_{l=1}^{n_+} x_l^+ - \bar{z}_j^- \right) \right]. \qquad \text{(OP2)}$$

Notice that the objective (OP2) can be maximized by maximizing each term separately; the optimal matrix is then given by (line 3 of Algorithm 2):

$$\bar{\pi}_{i,j} = \mathbb{1} \left[1 - w^T \left(\frac{1}{n_+} \sum_{l=1}^{n_+} x_l^+ - \bar{z}_j^- \right) \geq 0 \right].$$

This optimal matrix $\bar{\pi}$ is also a valid ordering matrix in $\Pi_{\beta,k}$, which interestingly, puts all the positives below the j-th negative if the average score of the positives is less than one plus score of the j-th negative. In particular, we obtain a rank-1 $\bar{\pi}$ matrix. After obtaining $(\overline{Z}_-, \bar{\pi})$ for $\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S)$, we can compute the sub-gradient which is shown (in line 4 of Algorithm 2) as follows:

$$\partial_w \widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) = \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \bar{\pi}_{i,j} \left[\bar{z}_j^- - \frac{1}{n_+} \sum_{l=1}^{n_+} x_l^+ \right].$$

D.2. Subgradient of $\widehat{R}^{\text{max}}_{\text{pAp@k}}$

Since maximum over finite sets can be interchanged in the definition of $\widehat{R}^{\text{max}}_{\text{pAp@k}}(w; S)$in (24) and since $\pi_{i,j} \geq 0$ for all i, j, we find that the argmax over $Z_- \subseteq S_-, |Z_-| = k$ is attained at the top-k negatives $\overline{Z}_- := \{z_{(1)_w}, \dots, z_{(k)_w}\} = \{\bar{z}_1, \dots, \bar{z}_k\}$ according to $w^T x$ (line 1 of Algorithm 2). Similarly, from the proof of Proposition 6, we find that the argmax over $Z_+ \subseteq S_+, |Z_+| = \beta$ is attained at the bottom-β positives $\underline{Z}_+ := \{z_{(n_+-\beta+1)_w}, \dots, z_{(n_+)_w}\} = \{\underline{z}_1, \dots, \underline{z}_\beta\}$ according to $w^T x$ (line 6 of Algorithm 2). All that remains is a combinatorial optimization problem to compute the relative ordering matrix between the bottom-β positives and top-k negatives as scored by $w^T x$:

$$\operatorname*{arg\,max}_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} \left[1 - w^T(\underline{z}_i^+ - \bar{z}_j^-)\right]. \tag{OP3}$$

Notice that the subgradient of $\widehat{R}^{\text{max}}_{\text{pAp@k}}(w; S)$ requires the relative ordering of the bottom-β positives and top-k negatives, which is unlike the ramp surrogate. The objective in (OP3) can now be maximized by optimizing each term separately. The optimal matrix is given by (line 7 of Algorithm 2)

$$\overline{\pi}_{i,j} = \mathbb{1}\left[1 - w^T(\underline{z}_i^+ - \bar{z}_j^-) \geq 0\right].$$

This optimal matrix $\overline{\pi}$ is a valid ordering matrix in $\Pi_{\beta,k}$ as well. Hence $(\underline{Z}_+, \overline{Z}_-, \overline{\pi})$ gives us the desired argument maximums for $\widehat{R}^{\text{max}}_{\text{pAp@k}}(w; S)$. We can compute the sub-gradient which is shown (in line 8 of Algorithm 2) as follows:

$$\partial_w \widehat{R}^{\text{max}}_{\text{pAp@k}}(w; S) = \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \overline{\pi}_{i,j} \left[\bar{z}_j^- - \underline{z}_i^+\right].$$

D.3. Subgradient of $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S)$

Similar to the procedure for $\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w; S)$, we observe that the argmax over $Z_- \subseteq S_-, |Z_-| = k$ in (14) is attained at the top-k negatives $\overline{Z}_- := \{z_{(1)_w}, \dots, z_{(k)_w}\}$ according to $w^T x$ (line 1 of Algorithm 2), and the combinatorial optimization problem over $\Pi_{n_+ \times k}$ becomes equivalent to (OP1):

$$\operatorname*{arg\,max}_{\pi \in \Pi_{n_+ \times k}} \frac{1}{\beta k} \left[\sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_\pi, j} - \sum_{i=1}^{n_+} \sum_{j=1}^{k} \pi_{i,j} w^T x_i^+ + \sum_{i=1}^{n_+} \sum_{j=1}^{k} \pi_{i,(j)_w} w^T \bar{z}_{(j)_w}^-\right].$$

As shown in Lemma 2, the above argmax can be computed over a restricted set of ordering matrices $\Pi^w_{n_+ \times k}$ (21), given for any $w \in \mathbb{R}^d$. This reduces (OP1) to a simpler optimization problem (22):

$$\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S) = \max_{\pi \in \Pi^w_{n_+ \times k}} \left[\frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} [1 - w^T x^+_{(i)_w} + w^T z^-_{(j)_w}] + \frac{1}{\beta k} \sum_{i=\beta+1}^{n_+} \sum_{j=1}^{k} \pi_{(i)_w,(j)_w} [-w^T x^+_{(i)_w} + w^T z^-_{(j)_w}]\right].$$

The above objective can be decomposed into a sum of terms involving individual elements $\pi_{i,j} \in \{0, 1\}$ and thus can be maximized by optimizing each term separately. The optimal matrix is given by (line 11 of Algorithm 2):

$$\overline{\pi}_{(i)_w,j} = \left\{\begin{array}{ll} 1, & \text{if } i \leq \beta \text{ and } w^T(x^+_{(i)_w} - \bar{z}_j^-) \leq 1 \\ 1, & \text{if } i > \beta \text{ and } w^T(x^+_{(i)_w} - \bar{z}_j^-) \leq 0 \\ 0, & \text{o.w.} \end{array}\right\} = \mathbb{1}\left[w^T(x^+_{(i)_w} - \bar{z}_j^-) \leq \mathbb{1}(i \leq \beta)\right]. \tag{27}$$

This optimal matrix $\overline{\pi}$ is a valid ordering matrix in $\Pi_{\beta,k}$. Hence $(\overline{Z}_-, \overline{\pi})$ gives us the desired argument maximums for $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S)$. We can now compute the sub-gradient shown (in line 12 of Algorithm 2) as follows:

$$\partial_w \widehat{R}^{\text{TS}}_{\text{pAp@k}}(w; S) = \frac{1}{\beta k} \sum_{i=1}^{n_+} \sum_{j=1}^{k} \overline{\pi}_{i,j} \left[\bar{z}_j^- - x_i^+\right].$$

E. Proof of Section 6

Proof of Theorem 2. Recall that the population pAp@k risk is defined (16) as:

$$R_{\text{pAp@k}}[f;\mathcal{D}] = \frac{1}{\gamma_+\gamma_-}\mathbb{E}_{x^+\sim\mathcal{D}_+,x^-\sim\mathcal{D}_-}[\mathbb{1}(f(x^+)\le f(x^-))T_{\gamma^-}(f,x^-)T_{\gamma^+}(f,x^+))], \tag{28}$$

where

$$\gamma_+ = \left\{\begin{array}{ll} 1, & \text{if } \mathbb{P}[x\sim\mathcal{D}_+]\le\gamma_- \\ \gamma_-, & \text{o.w.}\end{array}\right\}, \tag{29}$$

$T_{\gamma^-}(f,x^-)$ is 1 if $\mathbb{P}_{\widetilde{x}^-\sim\mathcal{D}_-}[f(\widetilde{x}^-)>f(x^-)]\le\gamma_-$ and 0 otherwise, and $T_{\gamma^+}(f,x^+)$ is 1 if $\mathbb{P}_{\widetilde{x}^+\sim\mathcal{D}_+}[f(\widetilde{x}^+)>f(x^+)]\le\gamma_+$ and 0 otherwise. Similarly, let us define the empirical version of the pAp@k risk as follows:

$$\widehat{R}_{\text{pAp@k}}[f;S] = \frac{1}{i_{\gamma_+}j_{\gamma_-}}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\mathbb{1}[f(x_i^+)\le f(x_j^-)]\widehat{T}_{\gamma^+}(f,x_i^+)\widehat{T}_{\gamma^-}(f,x_j^-), \tag{30}$$

where $i_{\gamma_+}=\min(n_+,j_{\gamma_-})$, $\widehat{T}_{\gamma^-}(f,x^-)$ is 1 if x_j^- lies in the top-j_{γ_-} negatives and 0 otherwise, and $\widehat{T}_{\gamma^+}(f,x^+)$ is 1 if x_j^+ lies in the top-i_{γ_+} positives and 0 otherwise.

Assuming that f has no ties and that γ_-n_-,γ_+n_+ are integers, let us define the population and sample thresholds on both positives and negatives as:

$$t_{\mathcal{D}_+,f,\gamma_+} = \arg\inf_{t\in\mathbb{R}}\{t \mid \mathbb{P}_{x^+\sim\mathcal{D}_+}[f(x^+)>t]=\gamma_+\}, \qquad \widehat{t}_{S_+,f,\gamma_+} = \arg\min_{t\in\mathbb{R}}\{t \mid \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>t]\ge\gamma_+\} \tag{31}$$

$$t_{\mathcal{D}_-,f,\gamma_-} = \arg\inf_{t\in\mathbb{R}}\{t \mid \mathbb{P}_{x^-\sim\mathcal{D}_-}[f(x^-)>t]=\gamma_-\}, \qquad \widehat{t}_{S_-,f,\gamma_-} = \arg\min_{t\in\mathbb{R}}\{t \mid \frac{1}{n_-}\sum_{i=1}^{n_-}\mathbb{1}[f(x_i^-)>t]\ge\gamma_-\}. \tag{32}$$

Given that there are no ties by f, $\widehat{t}_{S_+,f,\gamma_+}$ is the threshold on f above which $n_+\gamma_+$ of the positives in S_+ are ranked by f. This implies $\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_{S_+,f,\gamma_+}]=n_+\gamma_+$. Analogous is the case for the negatives in S_-. Now let us rewrite the population and empirical risk as follows:

$$R_{\gamma_+,\gamma_-}[f;\mathcal{D}] = \frac{1}{\gamma_+\gamma_-}\mathbb{E}_{x^+\sim\mathcal{D}_+,x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-)\ge f(x^+), f(x^+)>t_{\mathcal{D}_+,f,\gamma_+}, f(x^-)>t_{\mathcal{D}_-,f,\gamma_-}] \tag{33}$$

$$\widehat{R}_{\gamma_+,\gamma_-}[f;S] = \frac{1}{n_+\gamma_+n_-\gamma_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)\ge f(x_i^+), f(x_i^+)>\widehat{t}_{S_+,f,\gamma_+}, f(x_j^-)>\widehat{t}_{S_-,f,\gamma_-}]. \tag{34}$$

Let us first use $t_+,t_-,\widehat{t}_+$, and $\widehat{t}_-$ as shorthand notations for $t_{\mathcal{D}_+,f,\gamma_+}, t_{\mathcal{D}_-,f,\gamma_-}, t_{S_+,f,\gamma_+}$, and t_{S_-,f,γ_+}, respectively. Furthermore, let us define six more terms as follows:

$$\widetilde{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] = \frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+)>t_+]\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)\ge f(x^+)]\mathbb{1}[f(x_j^-)>t_-]\right] \tag{35}$$

$$\widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] = \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-)>t_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-)\ge f(x_i^+)]\mathbb{1}[f(x_i^+)>t_+]\right] \tag{36}$$

$$\bar{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] = \frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+)>t_+]\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)\ge f(x^+)]\mathbb{1}[f(x_j^-)>\widehat{t}_-]\right] \tag{37}$$

$$\bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] = \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > t_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > \widehat{t}_+]\right] \tag{38}$$

$$\breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] = \frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+) > \widehat{t}_+]\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) \geq f(x^+)]\mathbb{1}[f(x_j^-) > \widehat{t}_-]\right] \tag{39}$$

$$\breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] = \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > \widehat{t}_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > \widehat{t}_+]\right] \tag{40}$$

We then have for any $f \in \mathcal{F}$,

$$\begin{aligned}
R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - R_{\gamma_+,\gamma_-}[f;S] &= \frac{1}{2}(2R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - 2R_{\gamma_+,\gamma_-}[f;S]) \\
&= \frac{1}{2}\Big[(R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - \widetilde{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-]) + (R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - \widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+]) \\
&+ (\widetilde{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \bar{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-]) + (\widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+]) \\
&+ (\bar{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-]) + (\bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+]) \\
&+ (\breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \widehat{R}_{\gamma_+,\gamma_-}[f;S]) + (\breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \widehat{R}_{\gamma_+,\gamma_-}[f;S])\Big]
\end{aligned} \tag{41}$$

Using (41), for any $\epsilon > 0$, we have that:

$$\begin{aligned}
&\mathbb{P}_{S\sim\mathcal{D}_+^{n_+}\times\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\{R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - R_{\gamma_+,\gamma_-}[f;S] \geq \epsilon\}\right) \leq \underbrace{\mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - \widetilde{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{A}} \\
&+ \underbrace{\mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\left\{R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - \widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{B}} + \underbrace{\mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\widetilde{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \bar{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{C}} \\
&+ \underbrace{\mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\left\{\widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{D}} + \underbrace{\mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\bar{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{E}} \\
&+ \underbrace{\mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\left\{\bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{F}} + \underbrace{\mathbb{P}_{S\sim\mathcal{D}_+^{n_+}\times\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \widehat{R}_{\gamma_+,\gamma_-}[f;S] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{G}} \\
&+ \underbrace{\mathbb{P}_{S\sim\mathcal{D}_+^{n_+}\times\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \widehat{R}_{\gamma_+,\gamma_-}[f;S] \geq \frac{\epsilon}{4}\right\}\right)}_{\text{H}}
\end{aligned}$$

We will now bound every term separately. Let us start with term A.

$$\begin{aligned}
&R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - \widetilde{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] \\
&= \frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t_+]\left[\mathbb{E}_{x^-\sim\mathcal{D}_-}\frac{1}{\gamma_-}[\mathbb{1}[f(x^-) \geq f(x^+)]\mathbb{1}[f(x^-) > t_-]] - \frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) \geq f(x^+)]\mathbb{1}[f(x_j^-) > t_-]\right] \\
&= \frac{1}{\gamma_+\gamma_-}\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t_+]\left[\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-) > \max\{f(x^+), t_-\}] - \frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \max\{f(x^+), t_-\}]\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\gamma_+\gamma_-}\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t_+]] \sup_{x^+\in\mathcal{X}} \left|\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-) > \max\{f(x^+), t_-\}] - \frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \max\{f(x^+), t_-\}]\right| \\
&\leq \frac{1}{\gamma_-}\sup_{x^+\in\mathcal{X}}\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > \max\{f(x^+), t_-\}] - \frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \max\{f(x^+), t_-\}]\right]\right| \quad (\text{as } \mathbb{1}[f(x^+) > t_+] \in \{0,1\}) \\
&\leq \frac{1}{\gamma_-}\sup_{t\in\mathbb{R}}|\mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > t] - \frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > t]| \qquad (42)
\end{aligned}
$$

Using (42), we have the following:

$$
\begin{aligned}
A &\leq \mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > t] - \frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > t]\right| \geq \frac{\gamma_-\epsilon}{4}\right\}\right) \\
&= \mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\bigcup_{t\in\mathbb{R}}\left\{\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > t] - \frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > t]\right| \geq \frac{\gamma_-\epsilon}{4}\right\}\right) \\
&\leq C_1 n_-^d e^{-2n_-\gamma_-^2\epsilon^2/16} \qquad (43)
\end{aligned}
$$

The last step follows from the usual VC dimension bound. Now we bound the B term.

$$
\begin{aligned}
&R_{\gamma_+,\gamma_-}[f;\mathcal{D}] - \widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] \\
&= \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > t_-]\left[\mathbb{E}_{x^+\sim\mathcal{D}_+}\frac{1}{\gamma_+}[\mathbb{1}[f(x^-) \geq f(x^+)]\mathbb{1}[f(x^+) > t_+]] - \frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > t_+]\right] \\
&\leq \frac{1}{\gamma_-\gamma_+}\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-) > t_-]]\sup_{x^-\sim\mathcal{X}}\Bigg|\Big[\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^-) \geq f(x^+)]\mathbb{1}[f(x^+) > t_+]] \\
&-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > t_+]\Big]\Bigg| \\
&\leq \frac{1}{\gamma_+}\sup_{x^-\sim\mathcal{X}}\left|\left[\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^-) \geq f(x^+)]\mathbb{1}[f(x^+) > t_+]] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > t_+]\right]\right| \\
&\leq \frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[t \geq f(x^+)]\mathbb{1}[f(x^+) > t_+]] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[t \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > t_+]\right| \\
&\leq \frac{1}{\gamma_+}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t_+] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t_+]\right| + \frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \\
&\leq \frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| + \frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \\
&= \frac{2}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \qquad (44)
\end{aligned}
$$

Using (44), we have the following:

$$
\begin{aligned}
B &\leq \mathbb{P}_{S_+ \sim \mathcal{D}_+^{n_+}} \left(\bigcup_{f \in \mathcal{F}} \left\{ \sup_{t \in \mathbb{R}} \left| \mathbb{E}_{x^+ \sim \mathcal{D}_+} [\mathbb{1}[f(x^+) > t] - \frac{1}{n_+} \sum_{i=1}^{n_+} \mathbb{1}[f(x_i^+) > t] \right| \geq \frac{\gamma_+ \epsilon}{4 \times 2} \right\} \right) \\
&= \mathbb{P}_{S_+ \sim \mathcal{D}_+^{n_+}} \left(\bigcup_{f \in \mathcal{F}} \bigcup_{t \in \mathbb{R}} \left\{ \left| \mathbb{E}_{x^+ \sim \mathcal{D}_+} [\mathbb{1}[f(x^+) > t] - \frac{1}{n_+} \sum_{i=1}^{n_+} \mathbb{1}[f(x_i^+) > t] \right| \geq \frac{\gamma_+ \epsilon}{8} \right\} \right) \\
&\leq C_2 n_+^d e^{-2n_+ \gamma_+^2 \epsilon^2 / 64} \qquad (45)
\end{aligned}
$$

The last step follows from the usual VC dimension bound. Now we bound the C term.

$$
\begin{aligned}
&\widetilde{R}^+_{\gamma_+, \gamma_-}[f; \mathcal{D}, S_-] - \bar{R}^+_{\gamma_+, \gamma_-}[f; \mathcal{D}, S_-] \\
&= \frac{1}{\gamma_+} \mathbb{E}_{x^+ \sim \mathcal{D}_+} \left[\mathbb{1}[f(x^+) > t_+] \frac{1}{n_- \gamma_-} \sum_{j=1}^{n_-} \mathbb{1}[f(x_j^-) \geq f(x^+)] \mathbb{1}[f(x_j^-) > t_-] \right] \\
&- \frac{1}{\gamma_+} \mathbb{E}_{x^+ \sim \mathcal{D}_+} \left[\mathbb{1}[f(x^+) > t_+] \frac{1}{n_- \gamma_-} \sum_{j=1}^{n_-} \mathbb{1}[f(x_j^-) \geq f(x^+)] \mathbb{1}[f(x_j^-) > \widehat{t}_-] \right] \\
&\leq \frac{1}{\gamma_+ \gamma_-} \mathbb{E}_{x^+ \sim \mathcal{D}_+} \mathbb{1}[f(x^+) > t_+] \sup_{x^+ \sim \mathcal{D}_+} \left[\frac{1}{n_-} \sum_{j=1}^{n_-} \underbrace{\mathbb{1}[f(x_j^-) \geq f(x^+)]}_{\in \{0,1\}} \{ \mathbb{1}[f(x_j^-) > t_-] - \mathbb{1}[f(x_j^-) > \widehat{t}_-] \} \right] \\
&\leq \frac{1}{\gamma_-} \left| \frac{1}{n_-} \sum_{j=1}^{n_-} \{ \mathbb{1}[f(x_j^-) > t_-] - \mathbb{1}[f(x_j^-) > \widehat{t}_-] \} \right| \\
&= \frac{1}{\gamma_-} \left| \frac{1}{n_-} \sum_{j=1}^{n_-} \mathbb{1}[f(x_j^-) > t_-] - \gamma_- \right| \\
&= \frac{1}{\gamma_-} \left| \frac{1}{n_-} \sum_{j=1}^{n_-} \mathbb{1}[f(x_j^-) > t_-] - \mathbb{E}_{x^- \sim \mathcal{D}_-} \mathbb{1}[f(x^-) > t_-] \right| \\
&= \frac{1}{\gamma_-} \sup_{t \in \mathbb{R}} \left| \frac{1}{n_-} \sum_{j=1}^{n_-} \mathbb{1}[f(x_j^-) > t] - \mathbb{E}_{x^- \sim \mathcal{D}_-} \mathbb{1}[f(x^-) > t] \right| \qquad (46)
\end{aligned}
$$

Using (46), we have the following:

$$
\begin{aligned}
C &\leq \mathbb{P}_{S_- \sim \mathcal{D}_-^{n_-}} \left(\bigcup_{f \in \mathcal{F}} \left\{ \sup_{t \in \mathbb{R}} \left| \mathbb{E}_{x^- \sim \mathcal{D}_-} [\mathbb{1}[f(x^-) > t] - \frac{1}{n_-} \sum_{i=1}^{n_-} \mathbb{1}[f(x_j^-) > t] \right| \geq \frac{\gamma_- \epsilon}{4} \right\} \right) \\
&= \mathbb{P}_{S_- \sim \mathcal{D}_-^{n_-}} \left(\bigcup_{f \in \mathcal{F}} \bigcup_{t \in \mathbb{R}} \left\{ \left| \mathbb{E}_{x^- \sim \mathcal{D}_-} \mathbb{1}[f(x^-) > t] - \frac{1}{n_-} \sum_{j=1}^{n_-} \mathbb{1}[f(x_j^-) > t] \right| \geq \frac{\gamma_- \epsilon}{4} \right\} \right) \\
&\leq C_3 n_-^d e^{-2n_- \gamma_-^2 \epsilon^2 / 16} \qquad (47)
\end{aligned}
$$

This is same as (43). The last step follows from the usual VC dimension bound. We now consider the term D.

$$
\begin{aligned}
&\widetilde{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] \\
&= \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > t_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > t_+]\right] \\
&- \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > t_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)]\mathbb{1}[f(x_i^+) > \widehat{t}_+]\right] \\
&= \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > t_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)][\mathbb{1}[f(x_i^+) > t_+] - \mathbb{1}[f(x_i^+) > \widehat{t}_+]]\right] \\
&\leq \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-) > t_-]] \sup_{x^-\sim\mathcal{D}_-}\left|\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \geq f(x_i^+)][\mathbb{1}[f(x_i^+) > t_+] - \mathbb{1}[f(x_i^+) > \widehat{t}_+]]\right| \\
&\leq \frac{1}{\gamma_+}|\frac{1}{n_+}\sum_{i=1}^{n_+}[\mathbb{1}[f(x_i^+) > t_+] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > \widehat{t}_+]]| \\
&\leq \frac{1}{\gamma_+}|\frac{1}{n_+}\sum_{i=1}^{n_+}[\mathbb{1}[f(x_i^+) > t_+] - \gamma_+]| \\
&= \frac{1}{\gamma_+}|\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t_+] - \mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t_+]| \\
&\leq \frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t] - \mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t]\right|,
\end{aligned}
\tag{48}
$$

where the third step follows from the fact that $\mathbb{1}[f(x_i^+) > t_+] - \mathbb{1}[f(x_i^+) > \widehat{t}_+] \geq 0$ if $t_+ \leq \widehat{t}_+$, and $\mathbb{1}[f(x_i^+) > t_+] - \mathbb{1}[f(x_i^+) > \widehat{t}_+] \leq 0$ otherwise. Using (48), we have the following:

$$
\begin{aligned}
D &\leq \mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \geq \frac{\gamma_+\epsilon}{4}\right\}\right) \\
&= \mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\bigcup_{t\in\mathbb{R}}\left\{\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \geq \frac{\gamma_+\epsilon}{4}\right\}\right) \\
&\leq C_4 n_+^d e^{-2n_+\gamma_+^2\epsilon^2/16}
\end{aligned}
\tag{49}
$$

The last step follows from the usual VC dimension bound. We now consider the term E.

$$
\begin{aligned}
&\bar{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] - \breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-] \\
&= \frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+) > t_+]\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) \geq f(x^+)]\mathbb{1}[f(x_j^-) > \widehat{t}_-]\right] \\
&- \frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+) > \widehat{t}_+]\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) \geq f(x^+)]\mathbb{1}[f(x_j^-) > \widehat{t}_-]\right] \\
&= \frac{1}{\gamma_+}\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \widehat{t}_-]\{\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x_j^-) \geq f(x^+)](\mathbb{1}[f(x^+) > t_+] - \mathbb{1}[f(x^+) > \widehat{t}_+])\}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\gamma_+}\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \widehat{t}_-]\mathbb{E}_{x^+\sim\mathcal{D}_+}\Big[\mathbb{1}[f(x^+) > t_+] - \mathbb{1}[f(x^+) > f(x_j^-)] - \mathbb{1}[f(x^+) > \widehat{t}_+] + \mathbb{1}[f(x^+) > f(x_j^-)]\Big] \\
&= \frac{1}{\gamma_+}\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \widehat{t}_-]\mathbb{E}_{x^+\sim\mathcal{D}_+}\Big[\mathbb{1}[f(x^+) > t_+] - \mathbb{1}[f(x^+) > \widehat{t}_+]\Big] \\
&\le \frac{1}{\gamma_+}\Big[\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t_+] - \mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > \widehat{t}_+]\Big] \\
&= \frac{1}{\gamma_+}\Big[\gamma_+ - \mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > \widehat{t}_+]\Big] \\
&= \frac{1}{\gamma_+}\left[\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > \widehat{t}_+] - \mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > \widehat{t}_+]\right] \\
&\le \frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t] - \mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t]\right| \qquad (50)
\end{aligned}
$$

Using (50), we have the following:

$$
\begin{aligned}
E &\le \mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \ge \frac{\gamma_+\epsilon}{4}\right\}\right) \\
&= \mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\bigcup_{t\in\mathbb{R}}\left\{\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+) > t] - \frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > t]\right| \ge \frac{\gamma_+\epsilon}{4}\right\}\right) \le C_5 n_+^d e^{-2n_+(\gamma_+)^2\epsilon^2/16}
\end{aligned}
\qquad (51)
$$

The last step follows from the usual VC dimension bound. We now consider the term F.

$$
\begin{aligned}
&\bar{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] - \breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+] \\
&= \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > t_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \ge f(x_i^+)]\mathbb{1}[f(x_i^+) > \widehat{t}_+]\right] \\
&- \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-) > \widehat{t}_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-) \ge f(x_i^+)]\mathbb{1}[f(x_i^+) > \widehat{t}_+]\right] \\
&= \frac{1}{\gamma_-}\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > \widehat{t}_+]\mathbb{E}_{x^-\sim\mathcal{D}_-}\Big[\mathbb{1}[f(x^-) \ge f(x_i^+)]\{\mathbb{1}[f(x^-) > t_-] - \mathbb{1}[f(x^-) > \widehat{t}_-]\Big] \\
&\le \frac{1}{\gamma_-}\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+) > \widehat{t}_+]\mathbb{E}_{x^-\sim\mathcal{D}_-}\Big[\{\mathbb{1}[f(x^-) > t_-] - \mathbb{1}[f(x^-) > \widehat{t}_-]\Big] \\
&\le \frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\Big[\{\mathbb{1}[f(x^-) > t_-] - \mathbb{1}[f(x^-) > \widehat{t}_-]\Big] \\
&= \frac{1}{\gamma_-}\Big[\mathbb{E}_{x^-\sim\mathcal{D}_-}\{\mathbb{1}[f(x^-) > t_-] - \mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > \widehat{t}_-]\Big] = \frac{1}{\gamma_-}\Big[\gamma_- - \mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > \widehat{t}_-]\Big] \\
&= \frac{1}{\gamma_-}\left[\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > \widehat{t}_-] - \mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > \widehat{t}_-]\right] \le \frac{1}{\gamma_-}\sup_{t\in\mathbb{R}}\left|\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-) > t] - \mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-) > t]\right|
\end{aligned}
\qquad (52)
$$

Using (52), we have the following:

$$
\begin{aligned}
F &\leq \mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-)>t]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>t]\right|\geq\frac{\gamma_-\epsilon}{4}\right\}\right)\\
&= \mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\bigcup_{t\in\mathbb{R}}\left\{\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-)>t]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>t]\right|\geq\frac{\gamma_-\epsilon}{4}\right\}\right)\\
&\leq C_6 n_-^d e^{-2n_-(\gamma_-)^2\epsilon^2/16}
\end{aligned}
\tag{53}
$$

The last step follows from the usual VC dimension bound. We now consider the term G.

$$
\begin{aligned}
&\breve{R}^+_{\gamma_+,\gamma_-}[f;\mathcal{D},S_-]-\widehat{R}_{\gamma_+,\gamma_-}[f;S]\\
&=\frac{1}{\gamma_+}\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+)>\widehat{t}_+]\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)\geq f(x^+)]\mathbb{1}[f(x_j^-)>\widehat{t}_-]\right]\\
&-\frac{1}{n_+\gamma_+n_-\gamma_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\mathbb{1}[f(x_j^-)\geq f(x_i^+)]\mathbb{1}[f(x_j^-)>\widehat{t}_-]\\
&=\frac{1}{\gamma_+}\frac{1}{n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\widehat{t}_-]\mathbb{E}_{x^+\sim\mathcal{D}_+}\left[\mathbb{1}[f(x^+)>\widehat{t}_+]\mathbb{1}[f(x_j^-)\geq f(x^+)]\right]\\
&-\frac{1}{\gamma_+n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\widehat{t}_-]\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\mathbb{1}[f(x_j^-)\geq f(x_i^+)]\\
&=\frac{1}{\gamma_+n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\widehat{t}_-]\left[\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+)>\widehat{t}_+]\mathbb{1}[f(x_j^-)\geq f(x^+)]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\mathbb{1}[f(x_j^-)\geq f(x_i^+)]\right]\\
&\leq\frac{1}{\gamma_+n_-\gamma_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\widehat{t}_-]\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+)>\widehat{t}_+]\mathbb{1}[t\geq f(x^+)]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\mathbb{1}[t\geq f(x_i^+)]\right|\\
&\leq\frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}\mathbb{1}[f(x^+)>\widehat{t}_+]\mathbb{1}[t\geq f(x^+)]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\mathbb{1}[t\geq f(x_i^+)]\right|\\
&=\frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[t\geq f(x^+)>\widehat{t}_+]]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[t\geq f(x_i^+)>\widehat{t}_+]\right|\\
&=\frac{1}{\gamma_-}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>\widehat{t}_+]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]-\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>t]+\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>t]\right|\\
&\leq\frac{1}{\gamma_+}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>\widehat{t}_+]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\right|+\frac{1}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>t]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>t]\right|\\
&=\frac{2}{\gamma_+}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>t]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>t]\right|
\end{aligned}
\tag{54}
$$

Using (54), we have the following:

$$
\begin{aligned}
G &\leq \mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>t]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>t]\right|\geq\frac{\gamma_+\epsilon}{4\times 2}\right\}\right)\\
&=\mathbb{P}_{S_+\sim\mathcal{D}_+^{n_+}}\left(\bigcup_{f\in\mathcal{F}}\bigcup_{t\in\mathbb{R}}\left\{\left|\mathbb{E}_{x^+\sim\mathcal{D}_+}[\mathbb{1}[f(x^+)>t]-\frac{1}{n_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>t]\right|\geq\frac{\gamma_+\epsilon}{8}\right\}\right)\leq C_7 n_+^d e^{-2n_+(\gamma_+)^2\epsilon^2/64}
\end{aligned}
\tag{55}
$$

The last step follows from the usual VC dimension bound. Now we bound the H term.

$$
\begin{aligned}
&\breve{R}^-_{\gamma_+,\gamma_-}[f;\mathcal{D},S_+]-\widehat{R}_{\gamma_+,\gamma_-}[f;S]\\
&=\frac{1}{\gamma_-}\mathbb{E}_{x^-\sim\mathcal{D}_-}\left[\mathbb{1}[f(x^-)>\widehat{t}_-]\frac{1}{n_+\gamma_+}\sum_{i=1}^{n_+}\mathbb{1}[f(x^-)\geq f(x_i^+)]\mathbb{1}[f(x_i^+)>\widehat{t}_+]\right]\\
&-\frac{1}{n_+\gamma_+n_-\gamma_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\widehat{t}_-]\mathbb{1}[f(x_j^-)\geq f(x_i^+)]\mathbb{1}[f(x_i^+)>\widehat{t}_+]\\
&=\frac{1}{n_+\gamma_+\gamma_-}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\left[\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-)>\widehat{t}_-]\mathbb{1}[f(x^-)\geq f(x_i^+)]]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\widehat{t}_-]\mathbb{1}[f(x_j^-)\geq f(x_i^+)]\right]\\
&=\frac{1}{n_+\gamma_+\gamma_-}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\left[\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-)>\max\{\widehat{t}_-,f(x_i^+)\}]]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>\max\{\widehat{t}_-,f(x_i^+)\}]\right]\\
&=\frac{1}{n_+\gamma_+\gamma_-}\sum_{i=1}^{n_+}\mathbb{1}[f(x_i^+)>\widehat{t}_+]\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-)>t]]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>t]\right|\\
&\leq\frac{1}{\gamma_-}\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-)>t]]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>t]\right|
\end{aligned}
\tag{56}
$$

Using (56), we have the following:

$$
\begin{aligned}
H &\leq \mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\left\{\sup_{t\in\mathbb{R}}\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}[\mathbb{1}[f(x^-)>t]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>t]\right|\geq\frac{\gamma_-\epsilon}{4}\right\}\right)\\
&=\mathbb{P}_{S_-\sim\mathcal{D}_-^{n_-}}\left(\bigcup_{f\in\mathcal{F}}\bigcup_{t\in\mathbb{R}}\left\{\left|\mathbb{E}_{x^-\sim\mathcal{D}_-}\mathbb{1}[f(x^-)>t]-\frac{1}{n_-}\sum_{j=1}^{n_-}\mathbb{1}[f(x_j^-)>t]\right|\geq\frac{\gamma_-\epsilon}{4}\right\}\right)\leq C_8 n_-^d e^{-2n_-(\gamma_-)^2\epsilon^2/16}
\end{aligned}
\tag{57}
$$

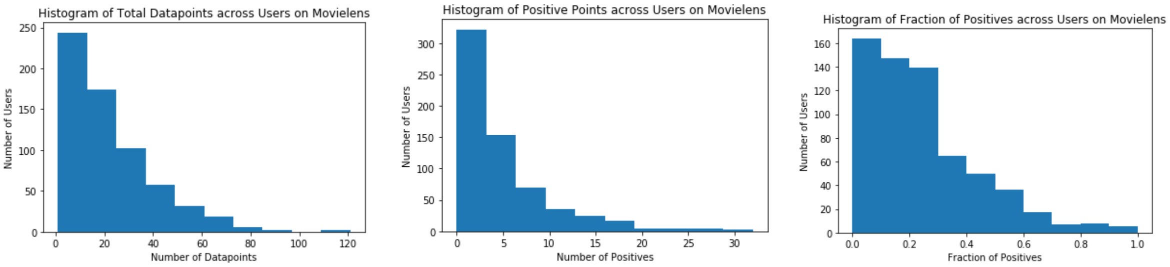
We may now use the bounds for the eight terms, i.e. (43), (45), (47), (49), (51), (53), (55), and (57) to get the desired result of Theorem 2. □

F. Experimental and Dataset Details

As discussed, we fix $\eta_t = \eta/\sqrt{t+1}$ in our methods and use a regularized version of the surrogates by adding $\lambda\|w\|^2$ for real-world experiments. For all the methods, including baselines, the learning rate and regularization parameters are

Table 4: Datasets Statistics

	Movielens ($d = 90$)			Citation ($d = 157$)			Behance ($d = 150$)		
Statistic	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Datapoints	41934	13870	14099	90213	22264	29317	402083	133913	134062
Positives	9273	3081	3132	13726	3348	4125	66861	22074	22473
Users	638	637	638	1573	402	502	2498	2498	2498



(a) Total datapoints across users

(b) Positives across users

(c) Fraction of positives across users.

Figure 2: Histograms of total datapoints, positives, and ratio of positives across users for Movielens test data.

cross validated on the set $\{10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, \ldots, 0.5\}$ and $10^{\{-3,\ldots,1\}}$, respectively. The reported results in Table 3 are averaged over 5 random runs. Since we evaluate performance by Micro-pAp@k (17), all the methods including baselines optimize the micro-version of the respective risks, i.e, average risk over users. Moreover, for fair comparisons, baseline methods are also cross-validated on Micro-pAp@k (17) instead of the metrics for which they were introduced. That is, the best hyper-parameters for each dataset and each method were chosen based on the highest Micro-pAp@k (17) value on the validation data.

We take three publicly available datasets and process them to reflect the challenges in modern recommender systems, i.e. data imbalance and heterogeneity in per-user fraction of positives. Moreover, we only focus on recommending limited (top-k) items. The schema for our datasets is <*user-feat, item-feat, prod-feat, label*>, where *prod-feat* is the Hadamard product of the user and item features. The data statistics are summarized in Table 4. In all the datasets, 60% data is for training, 20% data is for validating, and 20% data is for testing purposes.

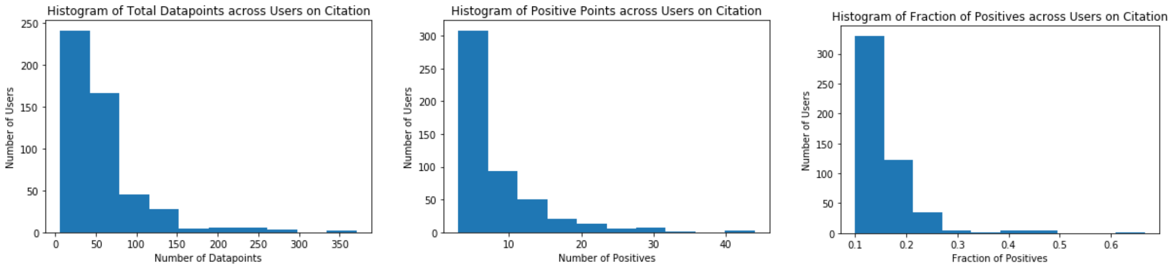
F.1. Movielens Dataset

We use the Movielens 100K dataset (Harper & Konstan, 2015), where the task is to recommend movies (items) to users.[5] We create a rating matrix by considering the first 20 movies rated by the users. Then we apply non-negative matrix factorization (Lee & Seung, 2001) to construct 30-dimensional user and item features. The non-negative matrix factorization is run for a maximum of 1000 iterations and stopped earlier if the change in loss reaches below 10^{-6}. The rest of the ratings are used in training and inference. We remove the users who do not have at least one rating in the remaining dataset. The number of features $d = 90$ after including the hadamard product of user and item features. Label 1 is provided if the rating of the movie is 5, and 0 otherwise. The train-validation-test data statistics is provided in Table 4. Histograms of total instances, positive instances, and fraction of positives across users on the test data are provided in Figure 2. Depending on the number of positives per user, we vary k from 8 to 24 with a step of 4 for this dataset.

F.2. Citation Dataset

The task in the citation dataset (Budhiraja et al., 2020) is to recommend research papers for a paper in progress. Both the paper being written, which acts as the user, and the candidate citations, which act as items, are embedded into 50 dimensional features using Glove embedding (Pennington et al., 2014). There are additional 7 features denoting the past interactions between authors and conferences. We further add the hadamard product of user and item features in the feature set, so the total number of features is $d = 157$. Both Glove embedding and binary labels denoting relevance are provided in the dataset. We remove the users who have less than three positives overall and less than 10% positives. A challenging aspect of this dataset is that there is no overlap among users in train, test, and validation data. The train-validation-test data statistics is provided in Table 4. Histograms of total instances, positive instances, and fraction of positives across users on the test data

[5] Download: https://grouplens.org/datasets/movielens/100k/

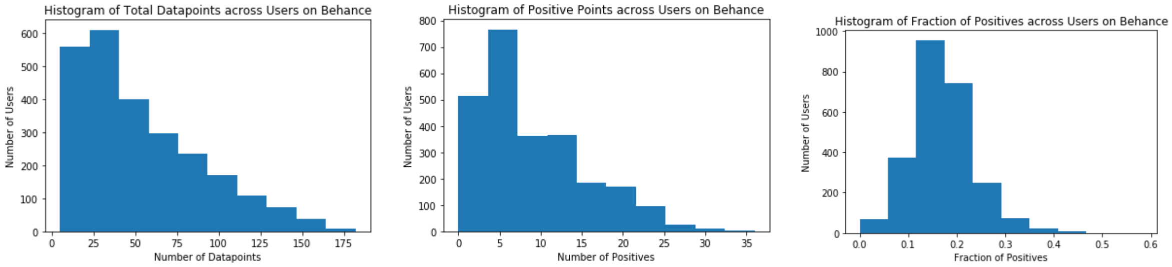


(a) Total datapoints across users

(b) Positives across users

(c) Fraction of positives across users.

Figure 3: Histograms of total datapoints, positives, and ratio of positives across users for Citation test data.



(a) Total datapoints across users

(b) Positives across users

(c) Fraction of positives across users.

Figure 4: Histograms of total datapoints, positives, and ratio of positives across users for Behance test data.

are provided in Figure 3. Depending on the number of positives per user, we vary k from 6 to 18 with a step of 3 for this dataset.

F.3. Behance Dataset

We consider the Behance dataset (He et al., 2016), where the task is to recommend images (items) to users.[6] We first apply UMAP (McInnes et al., 2018) (nearest neighbors = 10, minimum distance = 0.5) to reduce the 4096 dimensions of images to 50 dimensions. We filter users who have liked 60 to 170 images (just to control the number of users in the data). We then randomly select 50 liked images for each user and denote the average of those features as user features. The remaining images are used for training-test-validation. Label 1 is given if the user has liked an image. The label 0 is generated by random sampling (by generating either four, five, or six times the positives for each user). We again take the hadamard product of the user and item features, making the number of dimensions $d = 150$. The train-validation-test data statistics is provided in Table 4. Histograms of total instances, positive instances, and fraction of positives across users on the test data are provided in Figure 4. Depending on the number of positives per user, we vary k from 5 to 25 with a step of 5 for this dataset.

[6] Download: https://cseweb.ucsd.edu/ jmcauley/datasets.html#behance