Optimization and Analysis of the pAp@k Metric for Recommender Systems

Gaurush Hiranandani [1] **Warut Vijitbenjaronk** [1] **Oluwasanmi Koyejo** [1] **Prateek Jain** [2]

Abstract

Modern recommendation and notification systems must be robust to data imbalance, limitations on the number of recommendations/notifications, and heterogeneous engagement profiles across users. The pAp@k metric, which combines the partial-AUC and the precision@k metrics, was recently proposed to evaluate such recommendation systems and has been used in real-world deployments. Conceptually, pAp@k measures the probability of correctly ranking a top-ranked positive instance over top-ranked negative instances. Due to the combinatorial aspect surfaced by top-ranked points, little is known about the characteristics and optimization methods of pAp@k. In this paper, we analyze the learning-theoretic properties of pAp@k, particularly its benefits in evaluating modern recommender systems, and propose novel surrogates that are consistent under certain data regularity conditions. We then provide gradient descent based algorithms to optimize the surrogates directly. Our analysis and experimental evaluation suggest that pAp@k indeed exhibits a certain dual behavior with respect to partial-AUC and precision@k. Moreover, the proposed methods outperform all the baselines in various applications. Taken together, our results motivate the use of pAp@k for large-scale recommender systems with heterogeneous user-engagement.

1. Introduction

Modern recommendation services are quickly shifting from the search based *pull* paradigm, which requires users to be aware of the information in the first place, to *push* services, where the relevant information is recommended automatically. To devise high-quality push services, system designers must grapple with a variety of issues, which include data imbalance, space constraints, i.e., the recommendation of only a few items (top-k) out of a vast repository of items, and heterogeneity in the engagement profiles of users with the system, i.e., varied fraction of relevant items across users. Examples of platforms that must manage these constraints are ubiquitous – the Microsoft Teams platform[1] constructs the activity feed by displaying only a few essential messages to the users, the Behance[2] news feed includes only a few photos based on user's preferences, and Google Scholar[3] notifies users about a selected subset of the most relevant citations to the user's research. For such application domains, the recommendation problem can be formulated as a bipartite ranking problem. Further, in top-k settings, an ideal ranking metric rewards ordering the most relevant items at the top of the list, whereas accurate ranking in the remaining part is not of great importance (Rakotomamonjy, 2012).

To this end, several metrics have been proposed such as area under the receiver operating characteristic curve (AUC) (Rakotomamonjy, 2004), partial-AUC (McClish, 1989), and precision@k (Agichtein et al., 2006). Unfortunately, these standard metrics often fail to address one or more of the outlined challenges. For instance, while it is widely accepted that AUC is suitable for ranking problems characterized by imbalanced class priors (Cortes & Mohri, 2004), it can provide a misleading picture to the practitioners when the focus is on accuracy at the top of the list (Agarwal, 2011). Consider the example in Table 1, where three ranking functions f_1, f_2, and f_3 provide ranking to a dataset of five relevant (label = 1) and six irrelevant (label = 0) instances. While f_1 has the highest AUC, its accuracy at the top is the worst among the three functions.

The partial-AUC (pAUC) is designed to address this issue (Jiang et al., 1996; Agarwal, 2011); however, pAUC itself is often inadequate as it unnecessarily pits *all* the relevant items beyond the privileged set of k relevant items against the top scored irrelevant items – thus over-penalizing predictions that are sufficiently accurate at the top-k. Again, take the example in Table 1: f_2's accuracy at the top is not the most adequate; however, it achieves the highest pAUC when recommendations are limited to $k = 2$ items as pAUC rewards gain in ranking all the relevant items (even beyond top-2) over irrelevant items. This issue is exacerbated when the model is learned jointly across multiple users with a

[1]University of Illinois at Urbana-Champaign, Illinois, USA [2]Microsoft Research, Bengaluru, Karnataka, India. Correspondence to: Gaurush Hiranandani <gaurush2@illinois.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1] www.products.office.com/microsoft-teams

[2] www.behance.net [3] www.scholar.google.com

Table 1: Ranking of labeled instances by score functions.

	$k=2$				$k=6$	
Ranking	f_1	f_2	f_3	Ranking	f_4	f_5
(1)	0	1	1	(1)	1	1
(2)	1	0	1	(2)	1	1
(3)	1	1	0	(3)	0	1
(4)	0	1	0	(4)	1	0
(5)	1	1	0	(5)	1	1
(6)	1	0	0	(6)	1	1
(7)	1	0	0	(7)	0	0
(8)	0	0	0	(8)	0	0
(9)	0	0	1	(9)	0	0
(10)	0	0	1	(10)	0	0
(11)	0	1	1	(11)	0	0
AUC	22/30	21/30	12/30	prec@6	5/6	5/6
pAUC$(0, 2/6]$	2/10	5/10	4/10	pAp@6	27/30	28/30
pAp@2	2/4	3/4	1	-	-	-

large variance in the number of relevant items, as in that case the model would be overwhelmingly driven by preferences of a few users with loads of relevant items.

On the other extreme, the Precision@k (prec@k) metric ignores ranking of items anywhere but at the top of the list, and is widely used in both classification (Prabhu & Varma, 2014) and ranking (Le & Smola, 2007) domains. However, when there are fewer than k relevant items, perhaps for users with low engagement with the system, prec@k does not reward the few relevant items being ranked above the irrelevant items within the top-k and thus may give a false sense of achieving a perfect system. For example, consider the functions f_4 and f_5 in Table 1 which provide ranking on the same dataset. While prec@6 for both is at the highest achievable value, clearly f_5 provides better ranking since it puts more relevant items above the (inevitable) irrelevant items in top-6. This issue is exacerbated when there are a significant number of users with less than k relevant items.

Building on the disjoint properties of pAUC and prec@k metrics, Budhiraja et al. (2020) recently proposed a novel joint classification and ranking metric for recommender system - the 'partial-AUC + precision@k' (namely pAp@k). Intuitively, for a given score function, pAp@k measures AUC between the top-k irrelevant items and top-β relevant items, where β is the minimum of k and the number of relevant items. The metric behaves like prec@k when the number of relevant items are larger than k and like pAUC otherwise, and is currently used in Microsoft Teams' production system (Budhiraja et al., 2020). In this work, we shed light on how pAp@k eliminates the deficiencies of both prec@k and pAUC in top-k setting for systems having users with varied engagement. Furthermore, since pAp@k intertwines aspects of prec@k and pAUC in a complicated manner, existing optimization and analysis for the component metrics do not extend to pAp@k. This manuscript highlights the advantages of pAp@k in evaluating modern recommender systems and provides an analysis of the pAp@k metric. We further provide a novel optimization procedure for training models to optimize pAp@k. In summary, the key contributions of this paper are:

- We analyze the pAp@k metric, discuss its utility for modern recommendation systems, and further motivate its use to evaluate such systems.
- We propose four novel surrogates for pAp@k (Section 4), which are inspired by the structural surrogate for multivariate losses (Joachims, 2005). The surrogates are constructed as upper bounds of pAp@k and are shown to be consistent under certain data regularity conditions.
- We then provide procedures to compute sub-gradients for each of the surrogates and use them to enable sub-gradient descent methods for optimizing the surrogates (Section 5).
- We derive a uniform convergence generalization bound for the pAp@k performance measure, thus establishing that good training performance in terms of pAp@k also implies good generalization performance (Section 6).

Beyond conceptual and theoretical contributions, through a variety of simulated studies, we illustrate how pAp@k can be advantageous compared to pAUC and prec@k in outlined settings. We also conduct extensive experiments to show that the proposed methods optimize pAp@k better than a range of baselines in disparate recommendation applications involving image, text, or latent features (Section 7). Lastly, the proofs derived in this work are provided in the appendix.

2. Related Work

Many metrics have been proposed for bipartite ranking problem (Baeza-Yates et al., 1999; Menon & Williamson, 2016), each capturing different ranking notions. While the literature has focused on issues with data imbalance and ranking accuracy at the top (Joachims, 2005; Narasimhan & Agarwal, 2017; Kar et al., 2014), accommodating different engagement profiles of the system's users and thus different amount of data imbalance per user has largely been ignored. By incorporating properties of both prec@k and pAUC, the pAp@k metric (Budhiraja et al., 2020) is able to eliminate their deficiencies for a variety of ranking scenarios. For example, when the number of relevant items are more than k, pAp@k ignores rewards gained by performing better in other parts of the ranked list, unlike pAUC. On the other hand, when the number of relevant items are less than k, pAp@k focuses on ranking relevant items above irrelevant items in top-k, thus addressing this deficiency of prec@k.

Subset wise or list wise metrics such as NDCG@k (Wang et al., 2013) and MAP@k (Yilmaz & Aslam, 2006) are also popular ranking-at-the-top metrics, but they artificially assign different gain for different ranking positions which makes them challenging to use for bipartite ranking style problems that we encounter. For example, if the gain varies significantly with each position, then at the lower positions in top-k, presence of a positive or negative instance would not matter and hence it is easy to come up with scenarios where they can be swapped. In contrast, if the gains remain almost constant throughout top-k, then a ranking where top-

$k/2$ are negatives and bottom $k/2$ are positives might still give almost 90% accuracy. A broader family of bipartite ranking metrics are W-ranking measures (Clémençcon & Vayatis, 2009), which include metrics such as AUC and NDCG@k as special cases. Since W-ranking measures are defined on the positives and involve comparisons to all the negatives, no W-ranking measure is able to consider some specific (e.g. top-k) negatives. In particular, W-ranking measures have similar drawbacks as pAUC when the model is learned jointly across users, since under heterogeneous user engagement, the model can be driven by a few users with many negatives. Finally, two-way pAUC (Yang et al., 2019) is another related metric, but it requires the true positive rate to be lower-bounded which for many users, who have a large number of positives, implies that the measure will focus on a large part of the list (i.e., k would be high).

From the optimization point of view, there have been several techniques like structural support vector machines and direct gradient descent on surrogates that have been proposed and analyzed in literature (Herbrich, 2000; Joachims, 2002; Freund et al., 2003; Burges et al., 2005; Joachims, 2005; Kar et al., 2015). However, such techniques do not apply to the pAp@k metric as it combines challenging aspects of both pAUC and prec@k. To our knowledge, ours is the first work to develop conditionally consistent surrogates and principled gradient descent methods that can directly optimize the non-convex, computationally hard pAp@k metric.

3. Preliminaries and Background

Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{0, 1\}$ represent the instance and label space (0 = negative class, 1 = positive class), respectively. Let the training sample be $S = S_+ \cup S_-$ containing n_+ positive instances $S_+ = (x_1^+, \ldots, x_{n_+}^+) \in \mathcal{X}^{n_+}$ drawn independent and identically distributed (iid) according to a probability distribution $\mathcal{D}_+$ and n_- negative instances $S_- = (x_1^-, \ldots, x_{n_-}^-) \in \mathcal{X}^{n_-}$ drawn iid according to a probability distribution $\mathcal{D}_-$. Given S, the goal is to learn a scoring function $f : \mathcal{X} \to \mathbb{R}$ that provides optimal pAp@k value (defined below), where $k \in \mathbb{Z}_+$ is the number of item recommendations to the users. Typically, k is set to be a constant and as most users may have low engagement with the system (i.e., very few positives in train/test data), we study pAp@k with the assumption that $k \leq n_-$. On the other hand, prec@k requires a stronger $k \leq n_+$ assumption to ensure comparable metric across users (Kar et al., 2015)).

Definition 1. pAp@k *(risk): Let $f : \mathcal{X} \to \mathbb{R}$ be a score function and $\beta = \min(n_+, k)$. The* pAp@k *measures the probability of correctly ranking a pair of positive and negative instances, where the positive is one among the top-β positives and the negative is one among the top-k negatives, i.e.:*

$$\widehat{R}_{\text{pAp@k}}(f; S) = \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \mathbb{1}[f(x_{(i)_f}^+) \leq f(x_{(j)_f}^-)], \quad (1)$$

where $x_{(i)_f}^+$ and $x_{(j)_f}^-$ denote the positive instance ranked in i-th position (among positives in decreasing order of scores) and negative instance ranked in j-th position (among negatives in decreasing order of scores) by f, respectively.

The above definition can be extended for multiple users in the system leading to its Macro and Micro version (see (17)). Next, we compare the definitions of pAp@k to that of AUC and pAUC (Narasimhan & Agarwal, 2017). Recall that AUC considers all pairs of positive and negative instances, while pAUC$(0, k/n_-]$ considers pairs where the negative's score (according to f) is amongst the top-k scores of all negatives. In contrast, pAp@k restricts the positives also to be one among the top-β positives. This added restriction allows pAp@k to focus on the absolute top of the list while pAUC can be affected by positives with scores much lower down the list. On the other hand, when $n_+ \ll k$ and hence $\beta = n_+$, then unlike prec@k, pAp@k focuses on the pairwise ranking of the top-n_+ positives with the negatives; thus, reducing to pAUC$(0, k/n_-]$. In summary, pAp@k is maximized when relevant items are in the top-k items, and when there are less than k relevant items, pAp@k is maximized by placing the relevant items before the irrelevant items. Thus, as shown in Table 1, pAp@k can handle both large as well as small n_+ when compared to k (e.g. rewards better rankers f_3, f_5), while pAUC struggles when n_+ is large (e.g. compare rankers f_2, f_3) and prec@k can be misleading when n_+ is small (e.g. compare rankers f_4, f_5).

For the rest of the paper, we consider linear score functions of the form $f(x) = w^T x$ for $w \in \mathbb{R}^d$. The methods easily extend to non-linear functions/non-Euclidean spaces via Reproducing Hilbert Space kernels (Yu & Joachims, 2008).

3.1. Background: Structural surrogate for AUC

Like the majority of classification metrics, directly optimizing pAp@k (1) is computationally hard, so we propose a surrogate for the same. In preparation, we briefly outline the construction of the structural surrogate for AUC in the following. Let AUC *risk* be defined as:

$$\widehat{R}_{\text{AUC}}(f; S) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \mathbb{1}[w^T x_i^+ \leq w^T x_j^-]. \quad (2)$$

Let us denote the errors in relative ordering of the positives and negatives via the matrix $\pi \in \{0, 1\}^{n_+ \times n_-}$ as follows:

$$\pi_{ij} = \begin{cases} 1 & \text{if } x_i^+ \text{ is ranked below } x_j^- \\ 0 & \text{if } x_i^+ \text{ is ranked above } x_j^-. \end{cases}$$

Moreover, let Π_{n_+, n_-} denote the subset of matrices in $\{0, 1\}^{n_+ \times n_-}$ that correspond to valid orderings. Note that an optimal relative ordering π^* has entries $\pi_{ij}^* = 0 \; \forall i, j$. For any $\pi \in \Pi_{n_+, n_-}$, we may now define the AUC loss of π with respect to (w.r.t.) π^* as:

$$\Delta_{\text{AUC}}^{n_+ \times n_-}(\pi^*, \pi) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \pi_{i,j}. \tag{3}$$

Also let us define a joint feature map between the input training sample and an output ordering matrix π as: $\phi(S, \pi) := \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} (1 - \pi_{ij})(x_i^+ - x_j^-)$. This choice of joint features ensures that for any fixed $w \in \mathbb{R}^d$, $\arg\max_\pi w^T \phi(S, \pi)$ is a $\pi \in \Pi_{n_+, n_-}$ that is consistent with the score function $w^T x$ and hence for which the loss term evaluates to $\widehat{R}_{\text{AUC}}(w; S)$ (2). Consequently, minimizing AUC risk w.r.t. w reduces to a saddle point of the form: $\min_w \widehat{R}_{\text{AUC}}^{\text{struct}}(w; S)$, where the surrogate $\widehat{R}_{\text{AUC}}^{\text{struct}}(w; S) :=$

$$\max_{\pi \in \Pi_{n_+, n_-}} \Delta_{AUC}^{n_+ \times n_-}(\pi^*, \pi) - \frac{w^T(\phi(S, \pi^*) - \phi(S, \pi))}{n_+ n_-}. \tag{4}$$

The above structural surrogate is an upper bound of the empirical AUC (Joachims, 2005). Moreover, this surrogate is convex in w as it is a maximum of convex functions in w.

We may construct a convex surrogate for pAp@k by a direct application of the above structural surrogate framework, e.g., by replacing the first term in (4) by $\Delta_{\text{pAp@k}}(\pi^*, \pi) = \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{(i)\pi,(j)_\pi}$, where $(i)_\pi$ and $(j)_\pi$ denotes the index of the i-th and j-th ranked positive and negative, respectively, by any fixed ordering consistent with π. However, this surrogate is a loose upper bound on pAp@k since the joint feature map ϕ is defined on all the instances and not just the ones relevant for pAp@k. Clearly, these terms alter the emphasis of the surrogate for a given k; hence, we shall now focus on constructing tighter surrogates for pAp@k.

4. Novel Surrogates for pAp@k

A proxy for pAp@k is called a *surrogate* when (a) the proxy upper bounds pAp@k (1) so that minimizing the proxy promotes small pAp@k (aka *upper bounding property*), and (b) optimizing the proxy yields an optimal solution for pAp@k, under some regularity assumptions (aka *conditional consistency*). Having set the requirements for a surrogate, we next discuss a family of surrogates for pAp@k.

4.1. The Ramp Surrogate for pAp@k

Let $Z_+ \subseteq S_+$ and $Z_- \subseteq S_-$ be the sets of positives and negatives, respectively. Let $\widehat{R}_{\text{AUC}}(w; Z_+, Z_-)$ denote the full AUC risk of scoring function $w^T x$ calculated on a sample containing the subsets of positives Z_+ and negatives Z_-. Then pAp@k of $w^T x$ is equivalent to the maximum over all k-sized subsets of $Z_- \subseteq S_-$ of the minimum over all β-sized subsets $Z_+ \subseteq S_+$ of $\widehat{R}_{AUC}(w; Z_+, Z_-)$. This is formalized in the following theorem:

Theorem 1. *Given data S and a score function $w \in \mathbb{R}^d$:*

$$\widehat{R}_{\text{pAp@k}}(w; S) = \max_{\substack{Z_- \subseteq S_-, \\ |Z_-| = k}} \min_{\substack{Z_+ \subseteq S_+, \\ |Z_+| = \beta}} \widehat{R}_{\text{AUC}}(w; Z_+, Z_-) \tag{5}$$

$$= \max_{\substack{Z_- \subseteq S_-, \\ |Z_-| = k}} \min_{\substack{Z_+ \subseteq S_+, \\ |Z_+| = \beta}} \frac{1}{\beta k} \sum_{x^+ \in Z_+} \sum_{x^- \in Z_-} \mathbb{1}(w^T x^+ \leq w^T x^-).$$

Notice that the order of the min-max over subsets Z_- and Z_+ does not affect the metric pAp@k and can be interchanged. Now, similar to Section 3.1, we upper bound the inside term $\widehat{R}_{\text{AUC}}(w; Z_+, Z_-)$ with a convex function such that the features are independent of w. In particular, let us denote truncated ordering matrices $\pi \in \{0, 1\}^{\beta \times k}$ for any subset of positive instances $Z_+ = \{z_1^+, \cdots z_\beta^+\} \subseteq S_+$ and negative instances $Z_- = \{z_1^-, \cdots z_k^-\} \subseteq S_-$ as:

$$\pi_{ij} = \begin{cases} 1 & \text{if } z_i^+ \text{ is ranked below } z_j^- \\ 0 & \text{if } z_i^+ \text{ is ranked above } z_j^-. \end{cases}$$

The set of valid orderings is denoted by $\Pi_{\beta \times k}$, and the correct ordering is given by $\pi^* = 0^{\beta \times k}$. Let the joint feature map $\varphi : (\mathcal{X}^\beta \times \mathcal{X}^k) \times \Pi_{\beta \times k} \to R^d$ be defined as $\varphi(Z_+, Z_-, \pi) := \sum_{i=1}^{\beta} \sum_{j=1}^{k} (1 - \pi_{i,j})(z_i^+ - z_j^-)$. Further, let the AUC in terms of π among any two subsets Z_+ and Z_- of positives and negatives be denoted by:

$$\Delta_{\text{AUC}}(\pi^*, \pi) := \Delta_{\text{AUC}}^{\beta \times k}(\pi^*, \pi) := \frac{1}{\beta k} \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j}.$$

By applying the upper bound of AUC (4) from Section 3.1 to the inner term $\widehat{R}_{\text{AUC}}(w; Z_+, Z_-)$ in (5), we can define a ramp surrogate of the metric pAp@k as follows:

$$\widehat{R}_{\text{pAp@k}}^{\text{ramp}}(w; S) = \max_{\substack{Z_- \subseteq S_-, \\ |Z_-| = k}} \min_{\substack{Z_+ \subseteq S_+, \\ |Z_+| = \beta}} \max_{\pi \in \Pi_{\beta \times k}} \Big\{ \Delta_{AUC}(\pi^*, \pi) - \frac{1}{\beta k} w^T (\varphi(Z_+, Z_-, \pi^*) - \varphi(Z_+, Z_-, \pi)) \Big\}. \tag{6}$$

Note that $\widehat{R}_{\text{pAp@k}}^{\text{ramp}}(w; S)$ is similar to the "ramp" losses for binary classification (Do et al., 2008). Next we show that the ramp surrogate $\widehat{R}_{\text{pAp@k}}^{\text{ramp}}(w; S)$ upper bounds pAp@k and is consistent under the following margin condition.

Definition 2. ***Weak* (β, δ)*-margin*** *(Kar et al., 2015): A dataset S satisfies the weak (β, δ)-margin condition if for some scoring function f and a set $\widetilde{S}_+ \subseteq S_+$ of size β,*

$$\min_{i \in \widetilde{S}_+} f_i - \max_{j \in S_-} f_j \geq \delta. \tag{7}$$

Moreover, we state that the function f realizes this margin. We refer the weak $(\beta, 1)$-margin condition as simply the weak β-margin condition.

Proposition 1. *For any scoring function $w^T x$, we have $\widehat{R}_{\text{pAp@k}}^{\text{ramp}}(w; S) \geq \widehat{R}_{\text{pAp@k}}(w; S)$. Moreover, if the scoring function $w^T x$ realizes the weak β-margin condition over a dataset S, then $\widehat{R}_{\text{pAp@k}}^{\text{ramp}}(w; S) = \widehat{R}_{\text{pAp@k}}(w; S) = 0$.*

Proposition 1 establishes that $R^{\text{ramp}}_{\text{pAp@k}}(w;S)$ upper bounds pAp@k and is indeed consistent w.r.t. pAp@k under the weak β-margin condition, then when there exist some β positive points that substantially outrank all the negatives. For $\beta = k < n_+$, this notion of margin is weaker than the standard notion of margin for binary classification, but the ramp surrogate (6) turns out to be non-convex in this case.

Proposition 2. *The ramp surrogate* (6) *for* pAp@k *is non-convex in general. It is convex when* $n_+ \leq k$.

4.2. The Average Surrogate for pAp@k

Let us expand the ramp surrogate as:

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_-\subseteq S_-\\|Z_-|=k}} \min_{\substack{Z_+\subseteq S_+\\|Z_+|=\beta}} \max_{\pi\in\Pi_{\beta\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} - \sum_{i=1}^{\beta} p_i w^T z_i^+ + \sum_{j=1}^{k} q_j w^T z_j^-\right], \tag{8}$$

where $p_i = \sum_{j=1}^{k}\pi_{i,j} \geq 0$ and $q_j = \sum_{i=1}^{\beta}\pi_{i,j} \geq 0$. 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:

$$\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_-\subseteq S_-\\|Z_-|=k}} \max_{\pi\in\Pi_{\beta\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} + \sum_{j=1}^{k} q_j w^T z_j^- - \max_{\substack{Z_+\subseteq S_+\\|Z_+|=\beta}} \sum_{i=1}^{\beta} p_i w^T z_i^+\right]. \tag{9}$$

Given a π, the maximum in the third term of the ramp surrogate as defined in (9) is lower bounded by average of $\sum_{i=1}^{\beta} p_i w^t z_i$ over all the subsets of size β. Formally,

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

where $\mathcal{Z}$ is the set of all ordered sets of size β. Now notice that the right hand side is the sum of p_i's weighted by the average score of the positive instances, i.e. $(\frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+)\sum_{i=1}^{\beta} p_i$. Using this in (9) allows us to define the average surrogate as follows:

$$\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_-\subseteq S_-\\|Z_-|=k}} \max_{\pi\in\Pi_{\beta\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} + \sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j} w^T z_j^- - \frac{1}{n_+}\sum_{l=1}^{n_+} w^T x_l^+ \sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{i,j}\right]. \tag{10}$$

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

Definition 3. (β,δ)***-margin*** *(Kar et al., 2015): A dataset* S *satisfy the* (β,δ) *margin condition if for some scoring function* f*, we have, for all sets* $\widetilde{S}_+ \subseteq S_+$ *of size* β*,*

$$\frac{1}{\beta}\sum_{i\in\widetilde{S}_+} f_i - \max_{j\in S_-} f_j \geq \delta. \tag{11}$$

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

Proposition 3. *For any scoring function* $w^T x$*, we have* $\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) \geq \widehat{R}_{\text{pAp@k}}(w;S)$*. Moreover, if the scoring function* $w^T x$ *realizes the* β*-margin condition over a dataset* S*, then* $\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) = \widehat{R}_{\text{pAp@k}}(w;S) = 0$.

A note on Max surrogate for pAp@k (Appendix C): We relax the ramp surrogate by replacing the maximum over Z_+ in (9) by a minimum (which becomes max when pushed outside) and thus construct another surrogate, namely the max surrogate (24). Due to space constraints and its similarity with the average surrogate in construction, we defer the details to Appendix C. Notice that the max surrogate is further loose than the average surrogate since the maximum in (9) is replaced by a minimum. Moreover, it is consistent under the strong (β,δ)-margin condition (Definition 5), which is the standard notion of binary classification margin, i.e, where all positives are separated by negatives by a margin δ, and hence is much stronger than the (β,δ)-margin condition.

Notice that slight variants of the weak-β, β, and strong-β margin conditions were proposed by Kar et al. (2015) for prec@k. Despite the apparent similarity, we note that the "natural" origin of these conditions and the consistency proofs for pAp@k follow an entirely different path, because pAp@k and its surrogates, by definition, deal with pairwise comparisons of positives and negatives; whereas, prec@k and its surrogates are not pairwise. We next propose our fourth surrogate for pAp@k and its consistency condition.

4.3. The tight-struct surrogate for pAp@k

The following surrogate was developed by noting close links between pAp@k and pAUC$(0,k/n_-]$. By subtracting the pairwise ranking loss for $n_+ - \beta$ positives from loss for all the positives in (5), we may write $\widehat{R}_{\text{pAp@k}}(w;S) =$

$$\max_{\substack{Z_-\subseteq S_-\\|Z_-|=k}} \min_{\substack{Z_+\subseteq S_+\\|Z_+|=n_+-\beta}} \frac{1}{\beta k}\left[\sum_{i=1}^{n_+}\sum_{j=1}^{k}\mathbb{1}(w^T x_i^+ \leq w^T z_j^-) - \sum_{i=1}^{n_+-\beta}\sum_{j=1}^{k}\mathbb{1}(w^T z_i^+ \leq w^T z_j^-)\right]. \tag{12}$$

The first term is AUC considering all the positives and the set of negatives $Z_- \subseteq S_-$ w.r.t to w. We already have a convex upper bound for this term from Section 3.1. The second term in (12) is relative ordering of the bottom $n_+ - \beta$ positives w.r.t $w^T x$ and negatives in the set $Z_- \subseteq S_-$, i.e.:

$$\max_{\substack{Z_+\subseteq S_+\\ |Z_+|=n_+-\beta}} \sum_{i=1}^{n_+-\beta}\sum_{j=1}^{k}\pi_{i,j} = \sum_{i=1}^{n_+-\beta}\sum_{j=1}^{k}\pi_{(\beta+i)_w,j}, \quad (13)$$

where $(i)_w$ is the index of the i-th ranked positive instance in S_+, when the instances are sorted in descending order by $w^T x$. Combining the above upper bound and (13) together in (12), we obtain the following tight-struct (TS) surrogate:

$$\begin{aligned}\widehat{R}_{\text{pAp@k}}(w,S) &\le \max_{\substack{Z_-\subseteq S_-\\|Z_-|=k}} \max_{\pi\in\Pi_{n_+\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{n_+}\sum_{j=1}^{k}\pi_{i,j}\right.\\ &\left. - \sum_{i=1}^{n_+-\beta}\sum_{j=1}^{k}\pi_{(\beta+i)_w,j} - \sum_{i=1}^{n_+} p_i w^T x_i^+ + \sum_{j=1}^{k} q_j w^T z_j^-\right]\\ &\le \max_{\substack{Z_-\subseteq S_-\\|Z_-|=k}} \max_{\pi\in\Pi_{n_+\times k}} \frac{1}{\beta k}\left[\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{(i)_\pi,j}\right.\\ &\left. - \sum_{i=1}^{n_+} p_i w^T x_i^+ + \sum_{j=1}^{k} q_j w^T z_j^-\right] =: \widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S), \quad (14)\end{aligned}$$

where $p_i = \sum_{j=1}^{k}\pi_{i,j}$, $q_j = \sum_{i=1}^{n_+}\pi_{i,j}$, and $(i)_\pi$ is the index of the i-th ranked positive by any fixed ordering consistent with π. The second inequality follows from $\sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{(i)_w,j} \le \sum_{i=1}^{\beta}\sum_{j=1}^{k}\pi_{(i)_\pi,j}$, since the maximum is taken over π. This is a convex surrogate similar to the structural surrogate $\widehat{R}^{\text{tight}}_{\text{pAUC}}$ provided for pAUC$(0, k/n_-]$ (Narasimhan & Agarwal, 2017), with the difference in the first term, as the π term only includes the top-β positives in the rank ordering by π. Interestingly, unlike other proposed surrogates, this surrogate takes features from all the positives into account. By construction, TS surrogate is an upper bound to the metric pAp@k and is consistent under the following interpolation of the weak and the strong (β,δ)-margin conditions.

Definition 4. ***Moderate*** (β,δ)***-margin:*** *A dataset S satisfies the moderate (β,δ)-margin condition if for some scoring function f and a set $\widetilde{S}_+ \subseteq S_+$ of size β,*

$$\min_{i\in S_+} f_i - \max_{j\in S_-} f_j \ge 0, \quad \min_{i\in \widetilde{S}_+} f_i - \max_{j\in S_-} f_j \ge \delta. \quad (15)$$

Moreover, we state that the function f realizes this margin. We refer the moderate $(\beta,1)$-margin condition as simply the moderate β-margin condition.

The above condition not only requires positives to be separated from negatives but also requires the top-β positives to be further separated from negatives by a margin δ.

Proposition 4. *For any scoring function $w^T x$, we have $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S) \ge \widehat{R}_{\text{pAp@k}}(w;S)$. Moreover, if the scoring function $w^T x$ realizes the moderate β-margin condition over a dataset S, then $\widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S) = \widehat{R}_{\text{pAp@k}}(w;S) = 0$.*

Algorithm 1 GD-pAp@k-*surr*

Input: Step lengths η_t, feasible set $\mathcal{W}$, surrogate 'surr', k
Output: A model $\overline{w} \in \mathcal{W}$
$w_0 \leftarrow 0, t \leftarrow 0$
while not converged **do**
  Set $g_t \in \partial_w \widehat{R}^{\text{surr}}_{\text{pAp@k}}(w_t; X, y, k)$ {*See Algorithm 2*}
  $w_{t+1} \leftarrow \Pi_{\mathcal{W}}[w_t - \eta_t \cdot g_t]$, $t \leftarrow t+1$ {*project onto $\mathcal{W}$*}
return $\overline{w} = w_{t+1}$

Algorithm 2 Subgradient calculation for pAp@k surrogates

Input: A model w, data X, y, surrogate 'surr', k
Output: A subgradient $g \in \partial_w \widehat{R}^{\text{surr}}_{\text{pAp@k}}(w; X, y, k)$
Obtain top-k neg. $\overline{Z}_-$ ordered in dec. order of scores by w
if 'surr'=='avg' **then**
  Set $\overline{\pi}_{i,j} = \mathbb{1}(1 - w^T(\frac{1}{n_+}\sum_{l=1}^{n_+} x_l^+ - \overline{z}_j^-) \ge 0)$
  Set $g = \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\overline{\pi}_{i,j}[\overline{z}_j^- - \frac{1}{n_+}\sum_{l=1}^{n_+} x_l^+]$
else if 'surr'=='max' **then**
  Get bottom-β pos. $\underline{Z}_+$ based on scores by w
  Set $\overline{\pi}_{i,j} = \mathbb{1}(1 - w^T(\underline{z}_i^+ - \overline{z}_j^-) \ge 0)$
  Set $g = \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k}\overline{\pi}_{i,j}[\overline{z}_j^- - \underline{z}_i^+]$
else
  Sort pos. in dec. order of scores assigned by w
  Set $\overline{\pi}_{(i)_w,j} = \mathbb{1}\{w^T(x^+_{(i)_w} - \overline{z}_j^-) \le \mathbb{1}(i \le \beta)\}$
  Set $g = \frac{1}{\beta k}\sum_{i=1}^{n_+}\sum_{j=1}^{k}\overline{\pi}_{i,j}[\overline{z}_j^- - x_i^+]$
return g

4.4. Comparing the Margins and Surrogates

Recall that the β-margin condition is stronger than the weak β-margin condition and weaker than the strong β-margin condition. Similar is the case with the moderate β-margin condition. However, there is no concrete relation between the β-margin and the moderate β-margin conditions and their respective consistent surrogates. We show this in our simulated experiments in Section 7.2. We conclude this section by summarizing the hierarchy in the surrogates.

Proposition 5. *For a dataset S and a scoring function $w^T x$, $\widehat{R}_{\text{pAp@k}}(w;S) \le \widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) \le \widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) \le \widehat{R}^{\text{max}}_{\text{pAp@k}}(w;S)$ and $\widehat{R}^{\text{ramp}}_{\text{pAp@k}}(w;S) \le \widehat{R}^{\text{TS}}_{\text{pAp@k}}(w;S)$.*

5. Gradient Descent Algorithms for pAp@k

We now present sub-gradient descent (GD) based algorithms for maximizing the pAp@k performance measure. The proposed algorithms can be readily modified for stochastic GD (Kar et al., 2015), or cutting plane based (Joachims, 2005) methods as well. In particular, we discuss optimization routines for the three proposed surrogates i.e. average surrogate (10), max surrogate (24), and TS surrogate (14) in Algorithm 1. The algorithms for optimizing the surrogates follow a common routine with an exception of computing their respective non-trivial subgradients which are specified in Algorithm 2; see Appendix D for more details. Furthermore, since the algorithms use subgradient descent for optimizing convex (except ramp surrogate) but non-smooth

functions, the resulting method converges to an at most ϵ-sub optimal solution in $O(1/\epsilon^2)$ steps (Bubeck, 2014).

6. Generalization

We now derive a uniform convergence generalization bound for the pAp@k metric. This will establish that good training performance w.r.t pAp@k will also imply good generalization performance. Let us first define the population version of pAp@k for a general scoring function $f : X \to \mathbb{R}$. Let $\gamma_- \in (0, 1]$ (equiv. to k/n_- in the empirical pAp@k (1)), then the population pAp@k is defined as:

$$R_{\text{pAp@k}}(f;\mathcal{D}) = \frac{1}{\gamma_+\gamma_-}\mathbb{E}_{\substack{x^+\sim\mathcal{D}_+\\ x^-\sim\mathcal{D}_-}}\left[\mathbb{1}\{f(x^+) \le f(x^-)\}\times \right.$$
$$\left. \times T_{\gamma^+}(f,x^+)T_{\gamma^-}(f,x^-)\right], \quad (16)$$

where γ_+ is 1 if $\mathbb{P}[x \sim \mathcal{D}_+] \le \gamma_-$, and γ_- otherwise, $T_{\gamma^-}(f, x^-)$ 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. Next, we provide a bound on the generalization performance of any learned scoring function chosen from a function class $\mathcal{F}$ of reasonably bounded capacity in terms of its empirical risk. We measure the capacity of such a function class $\mathcal{F}$ using the VC dimension of the class of thresholded classifiers obtained from scoring functions in the class: $\tau_{\mathcal{F}} = \{\text{sign}(f - t) \mid f \in \mathcal{F}, t \in \mathbb{R}\}$. We have the following convergence bound for pAp@k:

Theorem 2. *Let $\mathcal{F}$ be a class of real-valued functions on X, and $\tau_{\mathcal{F}} = \{sign(f - t) \mid f \in \mathcal{F}, t \in \mathbb{R}\}$. Let $\rho > 0$. Then with probability at least $1 - \rho$ (over draw of sample $S = (S_+, S_-)$ from $\mathcal{D}_+^{n_+} \times \mathcal{D}_-^{n_-}$), we have for all $f \in \mathcal{F}$,*

$$R_{\text{pAp@k}}(f;\mathcal{D}) \le \widehat{R}_{\text{pAp@k}}(f;S) +$$
$$C\left(\frac{1}{\gamma_+}\sqrt{\frac{d\ln n_+ + \ln 1/\rho}{n_+}} + \frac{1}{\gamma_-}\sqrt{\frac{d\ln n_- + \ln 1/\rho}{n_-}}\right),$$

where d is the VC-dimension of $\tau_{\mathcal{F}}$, and $C > 0$ is a distribution-independent constant.

The tightness of this bound depends on k (via $\gamma_- = k/n_-$). In particular, the smaller the value for k, looser is the bound. Moreover, the proof of this result bounds eight terms in tandem and differs substantially from the existing literature (Agarwal et al., 2005; Narasimhan & Agarwal, 2017).

7. Experiments

In this section, we present evaluation of our methods on simulated and real data.[4] Section 7.1 highlights the advantages of pAp@k in evaluating modern recommender/notifications systems via simulations. In Section 7.2, we discuss the behavior of the proposed surrogates for different margin conditions. Finally, in Section 7.3, we compare our methods to baselines for optimizing pAp@k on real-world datasets.

[4] Source code: https://github.com/gaurush-hiranandani/pap-k

7.1. pAp@k Interwining pAUC and prec@k

In this section, we demonstrate that pAp@k is a more useful evaluation criterion in varied per-user fraction of positives than prec@k and pAUC and thus optimizing pAp@k maybe advantageous. To this end, we simulate a bipartite ranking dataset. The positives and negatives are generated from the multivariate Gaussian distributions $\mathcal{N}(-1_d, I_{d\times d})$ and $\mathcal{N}(0_d, I_{d\times d})$, respectively. 1_d, 0_d, and $I_{d\times d}$ denote the vector of ones, vector of zeros, and identity matrix of dimension d, respectively. We fix $d = 5$. In this experiment, GD-pAp@k-avg is used for optimizing pAp@k. We also use the algorithms SGD@k-avg (Kar et al., 2015) and SVM-pAUC (Narasimhan & Agarwal, 2017) that directly optimize prec@k and pAUC$(0, k/n_-]$, respectively. The reported results are averaged over 300 random runs.

Case 1 ($n_+ < k$): We sample 10 positives and 160 negatives, and fix $k = 20$. We then apply GD-pAp@k-avg and SGD@k-avg methods until they converge. In Table 2, we report prec@k and AUC@k when the number of positives in top-k is same for both methods, i.e. AUC@k when prec@k is same. The number of runs over which the mean and standard deviation are computed is in parenthesis. We observe that GD-pAp@k-avg achieves not only better prec@k, but also significantly higher AUC@k when prec@k is the same for both methods. This suggests that when the number of positives is same in top-k for both methods, GD-pAp@k-avg pushes positives above negatives thus getting a better solution than optimizing prec@k. Furthermore, we see that GD-pAp@k-avg achieves higher prec@k and higher AUC@k (when prec@k is same) in the majority of the runs implying that the data settings which are tough for GD-pAp@k-avg are tougher for SGD@k-avg.

Case 2 ($n_+ > k$): We sample 20 positives and 160 negatives, and fix $k = 10$. We then apply SVM-pAUC and GD-pAp@k-avg methods until they converge. In this case, we seek which method puts more positives in top-k i.e. whose prec@k is higher. In Table 2, we see that GD-pAp@k-avg has higher prec@k than SVM-pAUC . This shows that SVM-pAUC tries to improve ranking beyond top-k; whereas, GD-pAp@k-avg has a higher focus on ranking at the top. Furthermore, GD-pAp@k-avg achieves higher prec@k in the majority of the runs in this case as well.

7.2. Behavior of Surrogates

We simulate ranking for one user for $d = 5$. We generate 250 positives from $\mathcal{N}(0_d, I_{d\times d})$. We fix $k = 30$, $w = 1_d$, and margin to be 1. We then generate 2000 negatives from $\mathcal{N}(0.5 \times 1_d, I_{d\times d})$ while maintaining three data regularity conditions, i.e., β, moderate β, and strong β-margin conditions. We do not consider the weak β-margin condition in this experiment since we lack an exact optimization method for the non-convex ramp surrogate. For the remaining mar-

Table 2: Dual Behavior of pAp@k: When $n_+ < k$ and when the number of positives is same in top-k for both GD-pAp@k-avg and SGD@k-avg, GD-pAp@k-avg pushes positives above negatives. When $n_+ > k$, SVM-pAUC tries to improve ranking beyond top-k; whereas, GD-pAp@k-avg has a higher focus on ranking at the top.

Method \ Metric	prec@k	#runs when prec@k is higher	AUC@k when prec@k is same	#runs when AUC@k is higher and prec@k is same
Case 1: $n_+ = 10, k = 20, n_- = 160$				
SGD@k-avg	0.20 ±0.14 (300)	5	0.59 ±0.34 (88)	30
GD-pAp@k-avg	**0.27** ±**0.13** (300)	**207**	**0.68** ±**0.34** (88)	**58**
Case 2: $n_+ = 20, k = 10, n_- = 160$				
SVM-pAUC	0.62 ±0.29 (300)	15	0.66 ±0.31 (156)	**82**
GD-pAp@k-avg	**0.68** ±**0.28** (300)	**129**	**0.71** **0.30** (156)	74

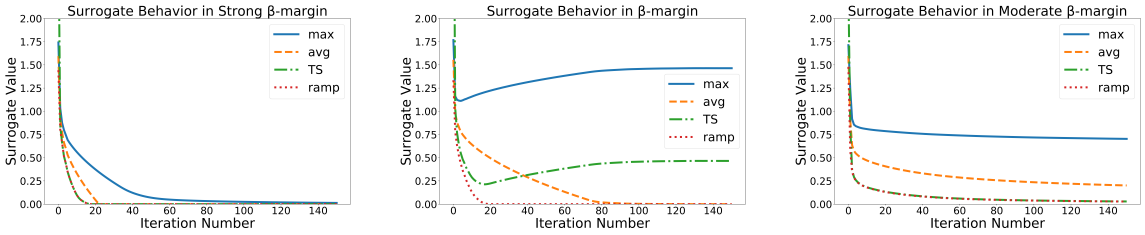

Figure 1: Behavior of surrogates under different margin conditions. See Section 7.2 for details.

gin conditions, we optimize their respective consistent surrogates and observe the behavior of all the surrogates.

First, in Figure 1(a), we see that all the surrogates converge to zero when the max surrogate is optimized in the strong β-margin condition. This validates Proposition 6 (Appendix C). Also notice that despite no direct connection with the max surrogate (Section 4.4), the TS surrogate still converges to zero as the strong β-margin condition is stricter than the moderate β-margin condition. Second, as we see in Figure 1(b), the ramp and average surrogates converge to zero in the β-margin condition validating Proposition 3; whereas, max and TS surrogates do not, and in fact, they increase in the later half of the optimization. Thus, the proposed surrogates might not be consistent with each other in general. Third, while optimizing TS surrogate in the moderate β-margin condition, we see in Figure 1(c) that the ramp and TS surrogates converge to zero validating Proposition 4.

7.3. Real-World Experiments

7.3.1. Datasets

We take three publicly available datasets and process them to reflect data imbalance and heterogeneity in per-user fraction of positives. Moreover, we focus on recommending 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. We summarize data properties below and defer detailed statistics to Appendix F.

Movie Recommendation (70K instances, 15.5K positives, 638 users, $d = 90, k = 8 - 24$): We use the Movielens 100K dataset (Harper & Konstan, 2015), where the task is to recommend movies (items) to users. We create a rating matrix by considering the first 20 ratings by timestamp of the users. Then we apply matrix factorization (Lee & Seung, 2001) to construct 30-dimensional user and item features. The rest of the ratings are used for constructing our dataset, where label 1 denotes a movie with rating 5, and 0 otherwise.

Citation Recommendation (142K instances, 21K positives, 2477 users, $d = 157, k = 6 - 18$): The task in the citation dataset (Budhiraja et al., 2020) is to recommend relevant research papers (items) for a paper in-progress (user). The 50-dimensional Glove embedding (Pennington et al., 2014) of the papers and the binary labels for relevance are given. The other 7 features denote author-conference interactions.

Image Recommendation (670K instances, 111K positives, 2498 users, $d = 150, k = 5 - 25$): We take the Behance dataset (He et al., 2016), where the task is to recommend images (items) to users. We first apply UMAP (McInnes et al., 2018) to reduce image dimensions from 4096 to 50. We then define user features by averaging features of randomly selected 50 liked images by that user. For the rest of the images, label 1 is given when a user has liked an image. Instances for label 0 are generated by random sampling.

7.3.2. Evaluation Metric and Baselines

Since pAp@k, by construction, is flexible to varied engagement level across users, our metric of interest with multiple users is the *micro* version of pAp@k (in gain form):

$$\text{Micro-pAp@k}(f) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{pAp@k}_u(f), \qquad (17)$$

Table 3: Micro-pAp@k gain (in %) for different methods on Movielens, Citation, and Behance datasets. Higher values are better. The performance of the proposed methods, especially GD-pAp@k-avg is better than the baselines.

Datasets→	**Movielens**					**Citation**					**Behance**				
↓ Methods, k →	**8**	**12**	**16**	**20**	**24**	**6**	**9**	**12**	**15**	**18**	**5**	**10**	**15**	**20**	**25**
GD-pAp@k-max	32.6	**35.1**	**37.6**	40.5	43.7	17.5	21.4	28.6	**34.8**	**35.1**	19.2	24.3	28.4	28.6	31.8
GD-pAp@k-avg	**35.5**	34.4	36.1	**42.5**	**46.5**	**20.7**	**25.6**	26.7	33.6	33.4	21.6	26.5	29.7	**30.8**	33.7
GD-pAp@k-TS	33.5	33.3	35.6	42.2	46.0	15.0	19.3	**31.6**	31.0	34.3	**22.8**	**26.9**	28.6	30.5	33.0
SVM-pAUC	34.9	33.7	35.7	41.1	46.3	14.5	24.4	29.5	32.3	30.8	19.7	24.4	27.8	30.7	32.8
Greedy-pAp@k	29.3	31.5	34.1	37.4	40.0	18.5	19.6	29.9	30.1	29.4	20.1	24.5	27.5	28.8	31.4
SGD@k-avg	30.7	32.3	32.8	35.0	36.3	20.4	23.0	24.2	28.1	30.5	19.6	26.6	**31.7**	30.6	**35.3**
Select-pAp@k	31.0	34.9	36.9	39.8	41.5	20.0	23.7	27.0	30.4	31.9	15.6	18.9	22.0	24.3	25.9

where f is the scoring function, $\mathcal{U}$ is the set of users, and pAp@k$_u(f)$ is pAp@k (gain) for the user u's ranked list.

We compare our proposed methods to (a) SVM-pAUC, an optimization method for pAUC$(0, k/n_-]$ (Narasimhan & Agarwal, 2017), (b) SGD@K-avg, a method for optimizing prec@k (Kar et al., 2015), (c) a greedy heuristic method (Ricamato & Tortorella, 2011) extended so to optimize pAp@k, denoted by Greedy-pAp@k, and (d) a procedure to optimize pAp@k from Algorithm 1 of (Budhiraja et al., 2020), denoted by Select-pAp@k. The baseline Select-pAp@k is a meta-procedure where in each iteration, the procedure creates a pool of datapoints by selecting top-β positives and top-k negatives for each user and then fully optimize a standard classification loss on the pooled sample. We choose to optimize cross entropy with gradient descent in each iteration of Select-pAp@k. All the methods are compared using linear models. 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 all the methods, including baselines, the learning rate and regularization parameters are cross validated on the set $\{10^{-4}, 2\times10^{-4}, 5\times10^{-4}, 10^{-3}, \ldots, 0.5\}$ and $10^{\{-3,\ldots,1\}}$, respectively. For fair comparisons, baseline methods are also cross-validated on Micro-pAp@k (17) instead of the metrics for which they were introduced.

7.3.3. RESULTS

Table 3 compares pAp@k accuracy of our methods against baselines. Overall, we see that our methods consistently outperform baselines especially for small k. On Movielens dataset, the proposed GD-pAp@k-max and GD-pAp@k-avg methods provide $1.5-2\%$ gains, while the relative improvement ranges from $1.8-5\%$, e.g. for $k=16$, GD-pAp@k-max is relatively 5.3% better than SVM-pAUC . Moreover, our best method has a relative improvement of 7.1% over Select-pAp@k on an average. Similarly, for citation dataset, we observe that GD-pAp@k-avg performs better than the other methods, especially for smaller values of k. For larger values of k, all the three surrogates perform comparable but much better than the baselines. For example, GD-pAp@k-max provides 10% relative improvement in comparison to the closest baseline for $k=18$. On the Behance dataset, we see that all the three proposed methods perform comparably to each other. However, they are closely followed by SGD@k-avg for smaller values of k, and beaten by it for larger values of $k=15,25$. So, in conclusion, our methods perform better in two out of three real-world datasets for optimizing pAp@k. Among the proposed methods, we find that methods based on tighter surrogates such as GD-pAp@k-avg are indeed beneficial.

8. Conclusion

In this paper, we investigated the bipartite ranking metric pAp@k. We found that the pAp@k metric possesses a dual behavior with respect to both partial-AUC and precision@k, and is particularly useful in evaluating large-scale recommender systems with heterogeneous user-engagement profiles. We then provided four novel, conditionally consistent surrogates for pAp@k and developed algorithms to optimize the surrogates directly. With a variety of simulated and real-world experiments, we demonstrated effectiveness of our proposed surrogates and algorithms in optimizing the pAp@k metric.

Acknowledgements

We thank the anonymous reviewers for providing helpful and constructive feedback on the paper. We also thank Harikrishna Narasimhan for helpful discussions. Oluwasanmi Koyejo acknowledges partial support by NSF IIS 1909577 and NSF IIS 1934986.

References

Agarwal, S. The infinite push: A new support vector ranking algorithm that directly optimizes accuracy at the absolute top of the list. In *SDM*, pp. 839–850. SIAM, 2011.

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. Generalization bounds for the area under the roc curve. *Journal of Machine Learning Research*, 6(Apr): 393–425, 2005.

Agichtein, E., Brill, E., and Dumais, S. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th annual international*

ACM SIGIR conference on Research and development in information retrieval, pp. 19–26. ACM, 2006.

Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Bubeck, S. Convex optimization: Algorithms and complexity, 2014.

Budhiraja, A., Hiranandani, G., Yarrabelly, N., Choure, A., Koyejo, O., and Jain, P. Rich-item recommendations for rich-users via gcnn: Exploiting dynamic and static side information. *arXiv preprint arXiv:2001.10495*, 2020.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. N. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*, pp. 89–96, 2005.

Clémençcon, S. J. and Vayatis, N. Empirical performance maximization for linear rank statistics. In *Advances in neural information processing systems*, pp. 305–312, 2009.

Cortes, C. and Mohri, M. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, pp. 313–320, 2004.

Do, C. B., Le, Q., Teo, C. H., Chapelle, O., and Smola, A. Tighter bounds for structured estimation. In *NeurIPS*, pp. 281–288. Curran Associates Inc., 2008.

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

He, R., Fang, C., Wang, Z., and McAuley, J. Vista: a visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 309–316, 2016.

Herbrich, R. Large margin rank boundaries for ordinal regression. *Advances in large margin classifiers*, pp. 115–132, 2000.

Jiang, Y., Metz, C. E., and Nishikawa, R. M. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201(3):745–750, 1996.

Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142. ACM, 2002.

Joachims, T. A support vector method for multivariate performance measures. In *ICML*, pp. 377–384. ACM, 2005.

Kar, P., Narasimhan, H., and Jain, P. Online and stochastic gradient methods for non-decomposable loss functions. In *Advances in Neural Information Processing Systems*, pp. 694–702, 2014.

Kar, P., Narasimhan, H., and Jain, P. Surrogate functions for maximizing precision at the top. In *ICML*, pp. 189–198, 2015.

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

Lee, D. D. and Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556–562, 2001.

McClish, D. K. Analyzing a portion of the roc curve. *Medical Decision Making*, 9(3):190–195, 1989.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Menon, A. K. and Williamson, R. C. Bipartite ranking: a risk-theoretic perspective. *The Journal of Machine Learning Research*, 17(1):6766–6867, 2016.

Narasimhan, H. and Agarwal, S. Support vector algorithms for optimizing the partial area under the roc curve. *Neural computation*, 29(7):1919–1963, 2017.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Prabhu, Y. and Varma, M. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 263–272. ACM, 2014.

Rakotomamonjy, A. Optimizing area under roc curve with svms. In *ROCAI*, pp. 71–80, 2004.

Rakotomamonjy, A. Sparse support vector infinite push. *arXiv preprint arXiv:1206.6432*, 2012.

Ricamato, M. T. and Tortorella, F. Partial auc maximization in a linear combination of dichotomizers. *Pattern Recognition*, 44(10-11):2669–2677, 2011.

Wang, Y., Wang, L., Li, Y., He, D., Chen, W., and Liu, T.-Y. A theoretical analysis of ndcg ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, pp. 6, 2013.

Yang, H., Lu, K., Lyu, X., and Hu, F. Two-way partial auc and its properties. *Statistical methods in medical research*, 28(1):184–195, 2019.

Yilmaz, E. and Aslam, J. A. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 102–111. ACM, 2006.

Yu, C.-N. J. and Joachims, T. Training structural svms with kernels using sampled cuts. In *KDD*, pp. 794–802. ACM, 2008.