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

Gaurush Hiranandani (UIUC),
Warut Vijitbenjaronk (UIUC),
Sanmi Koyejo (UIUC),
Prateek Jain (Microsoft Research)

NUANCES OF MODERN RECOMMENDERS/NOTIFIERS

- Three key challenges:
  - Data imbalance, i.e., high fraction of irrelevant items
  - Space constraints, i.e., recommending only top-*k* items
  - Heterogeneous user engagement profiles, i.e, varied fraction of relevant items across users

MANY EVALUATION METRICS, BUT…

Can be framed as bipartite ranking problems

Data Imbalance AUC W-ranking Measure precision@k

Space constraints (accuracy at the top) p-AUC map@k ndcg@k

Heterogeneous user engagement profiles!!???

Accommodating different engagement profiles of users or data imbalance per user has largely been ignored

INTRODUCING 'partial AUC + precision@k (pAp@k)'

We [Budhiraja et al. 2020] propose pAp@k, which measures the probability of correctly ranking a top-ranked positive instance over top-ranked negative instances

$$\hat{R}_{pAp@k}(f;S) = \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k} 1\left[f\left(x^{+}_{(i)_f}\right) \leq f\left(x^{-}_{(j)_f}\right)\right]$$

- $\hat{R}_{pAp@k}$ is pAp@k risk
- f is any scoring function
- S is finite data in $\mathcal{X} \times \{0,1\}$
- $x^{+}_{(i)_f}$ is the i-th positive when positives are sorted in decreasing order of scores by f
- $x^{-}_{(j)_f}$ is the j-th negative when negatives are sorted in decreasing order of scores by f
- $\beta = \min(n_+, k)$, where $n_+ = |S^+|$ is the number of positives in S

INTRODUCING 'partial AUC + precision@k (pAp@k)'

$$\hat{R}_{AUC}(f;S) = \frac{1}{n_+ n_-}\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} 1\left[f(x_i^+) \le f(x_j^-)\right]$$

AUC: All positives vs All negatives

$$\hat{R}_{pAUC}(f;S) = \frac{1}{n_+ k}\sum_{i=1}^{n_+}\sum_{j=1}^{k} 1\left[f(x_i^+) \le f\left(x_{(j)_f}^-\right)\right]$$

partial-AUC: All positives vs Top-k negatives

$$\hat{R}_{pAp@k}(f;S) = \frac{1}{\beta k}\sum_{i=1}^{\beta}\sum_{j=1}^{k} 1\left[f\left(x_{(i)_f}^+\right) \le f\left(x_{(j)_f}^-\right)\right]$$

pAp@k: Top β positives vs Top k negatives

$$\hat{R}_{prec@k}(f;S) = \frac{1}{k}\sum_{i=1}^{n} 1\left[x_{(i)_f} \in S^+\right]$$

prec@k: Counts positives in Top-k. No pairwise comparisons

	$k=2$				$k=6$	
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	-	-	-

CONTRIBUTIONS

- Analyze the pAp@k metric, discuss its utility, and further motivate its use to evaluate recommender systems

- Four novel surrogates for pAp@k that are consistent under certain data regularity conditions
- Procedures to compute sub-gradients that enable sub-gradient descent optimization methods
- Uniform convergence generalization bound

- Illustrate how pAp@k is advantageous compared to pAUC and prec@k through various simulated studies
- Extensive experiments show that the proposed methods optimize pAp@k better than a range of baselines in disparate recommendation applications

SURROGATES – RAMP SURROGATE

Let f(x) be of the form $w^T x$ (linear model)

- Rewriting the pAp@k risk

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

- The ramp surrogate

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

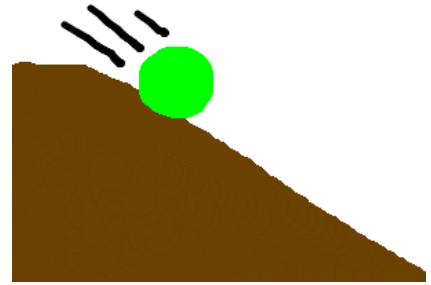
Structural Surrogate of AUC [Jaochims, 2005]

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

- Consistent under the Weak β-margin condition

  (a set of β positives are separated by all negatives by a margin)

- Non-convex



SURROGATES – AVG SURROGATE

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

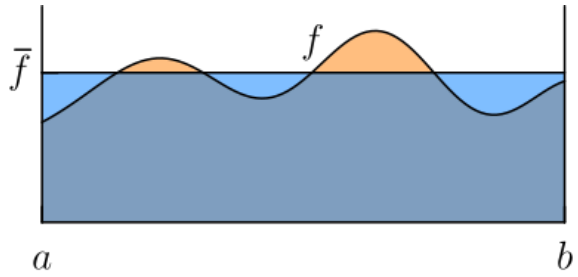
- The avg surrogate $$\widehat{R}^{\text{avg}}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_- \subseteq S_- \\ |Z_-|=k}} \max_{\pi \in \Pi_{\beta \times k}} \frac{1}{\beta k} \left[\sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} + \sum_{i=1}^{\beta} \sum_{j=1}^{k} \pi_{i,j} w^T z_j^- - \frac{1}{n_+} \sum_{l=1}^{n_+}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

The inside Max is replaced by average over all sets

- Consistent under the β-margin condition

  (the average score of positives is separated by scores of all negatives by a margin)

- Convex as it is point-wise maximum over convex functions in w



SURROGATES – MAX SURROGATE

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

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

The inside max is replaced by min and taken outside

- Consistent under the Strong β-margin condition

  (all positives are separated by negatives by a margin)

- Convex as it is point-wise maximum over convex functions in w

SURROGATES – TIGHT-STRUCT (TS) SURROGATE

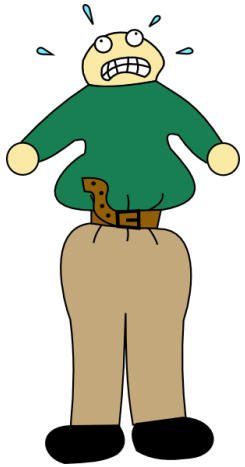
Previous margin conditions were proposed by [Kar et al., 2015] for prec@k (which is not pairwise); however,

the "natural" origin and consistency proofs for pAp@k (which is pairwise) follow an entirely different path

- Rewriting the pAp@k metric $$\widehat{R}_{\text{pAp@k}}(w;S) = \max_{\substack{Z_- \subseteq S_- \\ |Z_-|=k}} \min_{\substack{Z_+ \subseteq S_+ \\ |Z_+|=n_+-\beta}} \frac{1}{\beta k} \left[\sum_{i=1}^{n_+} \sum_{j=1}^{k} \mathbb{1}(w^T x_i^+ \leq w^T z_j^-) - \sum_{i=1}^{n_+-\beta} \sum_{j=1}^{k} \mathbb{1}(w^T z_i^+ \leq w^T z_j^-) \right]$$

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

Similar to structural surrogate for p-AUC
[Narasimhan et al., 2016] except for the first term



- Consistent under the Moderate β-margin condition (all positives are separated by negatives and a set of β positives are further separated by negatives by a margin)

- Convex as it is point-wise maximum over convex functions in w

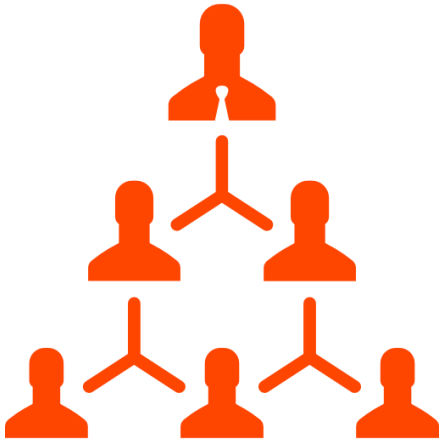
HIERARCHY

Weak β-Margin $\subseteq$ β-Margin $\subseteq$ Strong β-Margin

Weak β-Margin $\subseteq$ Moderate β-Margin $\subseteq$ Strong β-Margin

Moderate β-Margin ? β-Margin (shown in experiments)

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



GD ALGORITHM AND GENERALIZATION

Algorithm:

While not converged do:

1. $g_t \in \partial_w \hat{R}^{surr}_{pAp@k}(w_t; X, y, k)$
2. $w_{t+1} \leftarrow \Pi_{\mathcal{W}}[w_t - \eta_t g_t]$

Non-trivial sub-gradients of the surrogates derived in the paper

Convergence: converges to an ϵ-sub optimal solution in $O\left(\frac{1}{\epsilon^2}\right)$ steps

Generalization:

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

where $\gamma_- \in (0,1]$ (equivalent to k/n_- in the empirical setting)

γ_+ is 1 if $\mathbb{P}[x \sim D_+] \leq \gamma_-$ and γ_- otherwise

The smaller the value for *k*, looser is the bound

EXPERIMENTS: pAp@k INTERWINING pAUC AND prec@k

Simulate 1 user in two cases with positives and negatives generated from Gaussian with mean separation 1 **(300 trials)**

Algorithms SGD@k-avg and SVM-pAUC directly optimize prec@k and pAUC, respectively

Case 1 ($n_+ < k$): sample 10 positives, 160 negatives, and fix $k = 20$

Suggests GD-pAp@k-avg pushes positives above negatives more than SGD@k-avg

↓ Method, Metric →	prec@k	#trials prec@k >	#trials prec@k same	AUC@k when prec@k is same	#trials AUC@k > when prec@k is same
SGD@k-avg	0.20 ± 0.14	5	88	0.59 ± 0.34	30
GD-pAp@k-avg	**0.27 ± 0.13**	**207**	88	**0.68 ± 0.34**	**58**

Case 1 ($n_+ > k$): sample 20 positives, 160 negatives, and fix $k = 10$

Suggests SVMpAUC improves ranking beyond top-k; whereas, GD-pAp@k-avg focuses at the top

↓ Method, Metric →	prec@k	#trials prec@k >	#trials prec@k same	AUC@k when prec@k is same	#trials AUC@k > when prec@k is same
SVM-pAUC	0.62 ± 0.29	15	156	0.66 ± 0.31	82
GD-pAp@k-avg	**0.68 ± 0.28**	**129**	156	0.71 ± 0.30	74

EXPERIMENTS: pAp@k INTERWINING pAUC AND prec@k

Only a few positives are further separated then

Case 1 ($n_+ < k$): sample 10 positives, 160 negatives, and fix $k = 20$

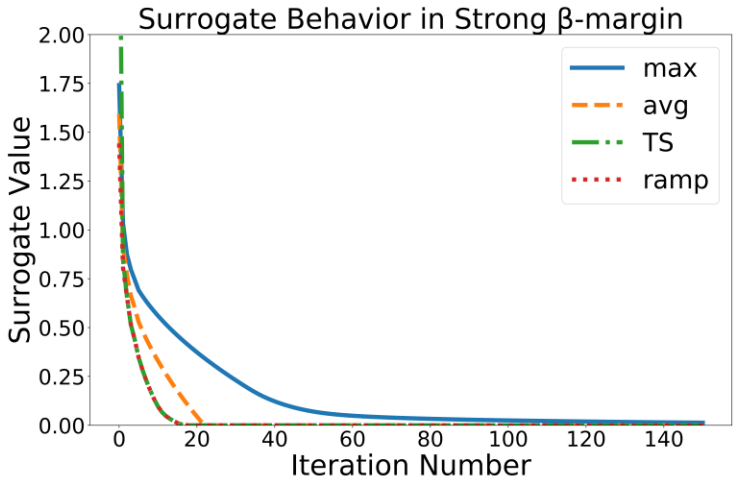
↓ Method, Metric →	prec@k	#trials prec@k >	#trials prec@k same	AUC@k when prec@k is same	#trials AUC@k > when prec@k is same
SGD@k-avg	0.45 ± 0.10	0	192	0.93 ± 0.07	75
GD-pAp@k-avg	**0.49 ± 0.02**	**108**	192	**0.98 ± 0.02**	**117**

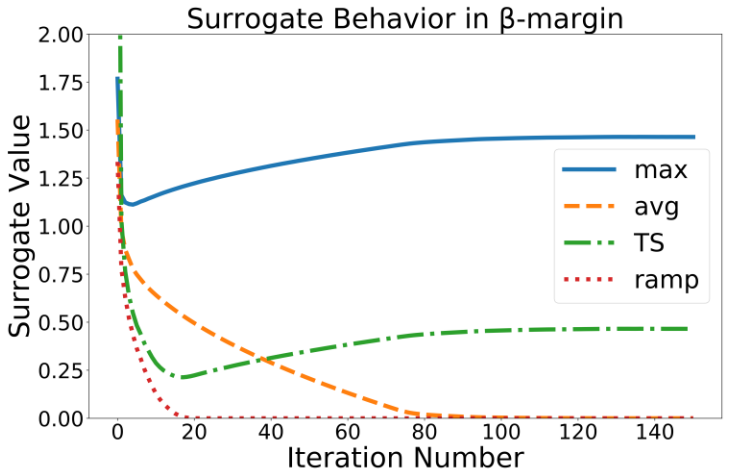
Case 1 ($n_+ > k$): sample 20 positives, 160 negatives, and fix $k = 10$

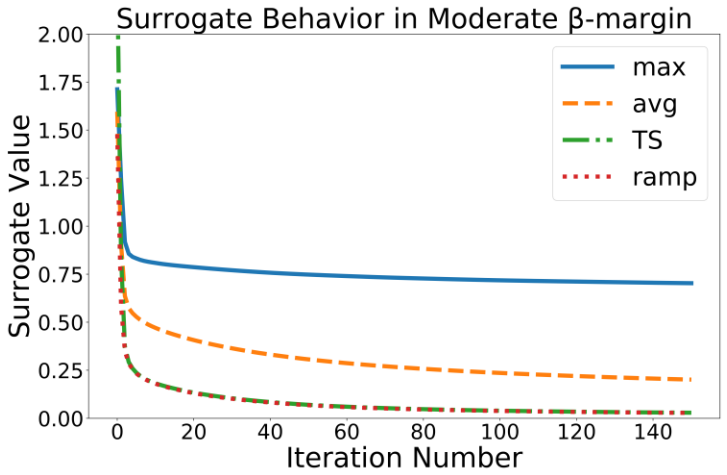
↓ Method, Metric →	prec@k	#trials prec@k >	#trials prec@k same	AUC@k when prec@k is same	#trials AUC@k > when prec@k is same
SVM-pAUC	0.85 ± 0.17	12	170	0.80 ± 0.20	117
GD-pAp@k-avg	**0.89 ± 0.14**	**118**	170	0.86 ± 0.17	53

EXPERIMENTS: BEHAVIOR OF SURROGATES

- Simulate 1 user with $d = 5$ features, fix $k = 30, n_+ = 250$ from $\mathcal{N}(0_d, I_{d\times d}), n_- = 2000$ from $\mathcal{N}(2 \times 1_d, I_{d\times d})$
- Maintain the margin conditions, optimize their respective consistent surrogates, and observe behaviour of all surrogates







- All surrogates converge to zero when max surrogate is optimized in strong β-margin condition. Despite no direct connection, TS surrogate converges to zero as strong β-margin condition is stricter than moderate β-margin condition
- Ramp and average surrogates converge to zero in the β-margin condition; whereas, max and TS surrogates do not
- While optimizing TS surrogate in the moderate β-margin condition, the ramp and TS surrogates converge to zero

EXPERIMENTS: REAL-WORLD DATA, COMPARING SURROGATES

Datasets: Movielens (latent features), Citation (text features), Behance (image features)

Dataset schema: <*user-feat, item-feat, prod-feat, label*>, where *prod-feat* is Hadamard product of user-feat and item-feat

Baselines:

(a) SVM-pAUC, an optimization method for pAUC

(b) SGD@K-avg, a method for optimizing prec@k

(c) greedy-pAp@k, a greedy heuristic extended so to optimize pAp@k

Evaluation: Micro-pAp@k (in gain %) – higher values are better

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**

CONCLUSIONS

- Analyze the learning-theoretic properties of the novel bipartite ranking metric pAp@k
- pAp@k indeed exhibits a certain dual behavior wrt p-AUC and prec@k (both in theory and in applications)
- Propose novel surrogates that are consistent under certain data regularity conditions
- Provide gradient descent based algorithms to optimize the surrogates directly
- Provide a generalization bound, thus establishing good training performance implies good generalization performance
- Analysis and experimental evaluation reveal that pAp@k is a more useful evaluation measure in data imbalanced, top-k constrained, and heterogeneous user engagement profile-based recommender and notification systems
- **Overall, our results motivate the use of pAp@k for large-scale recommender systems**

Thank You!