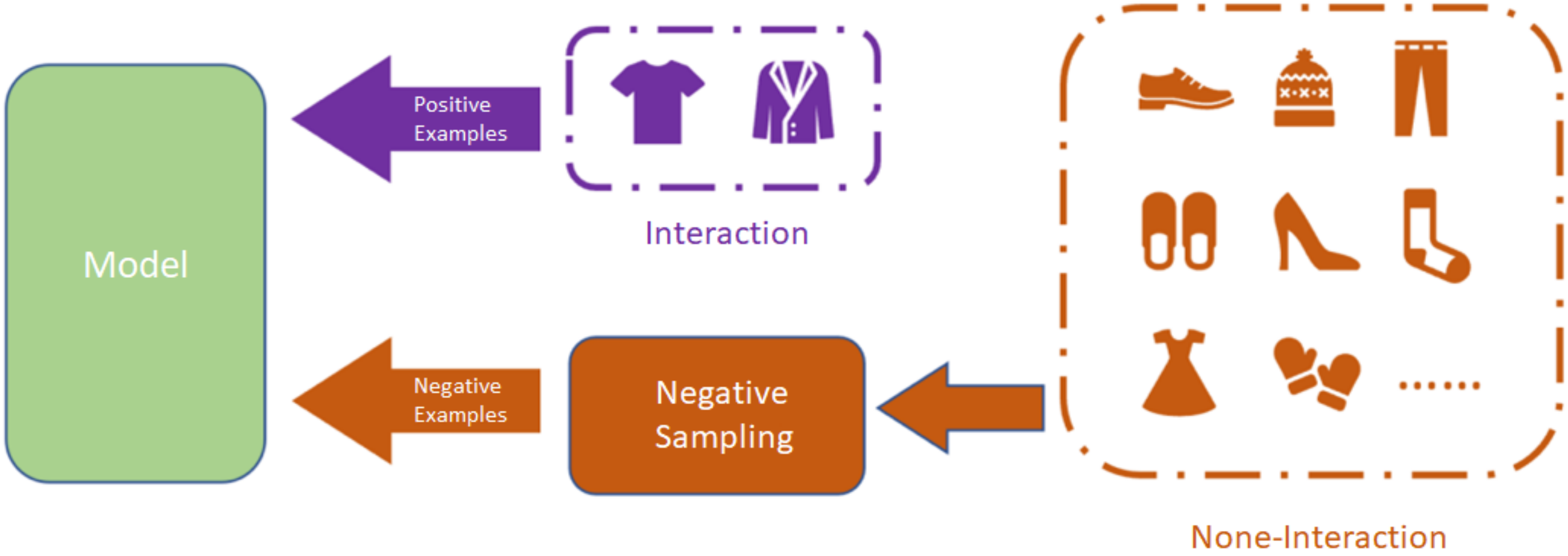
❖ Outline

1. Introduction
    - Negative Sampling in Recommendation
    - Revisite Hard Negative Sampling
    - Overview of Theoretical Structure
2. Preliminary
3. Theoretical Analysis and Guidelines
4. Experiments

1.1 Negative Sampling in Recommendation

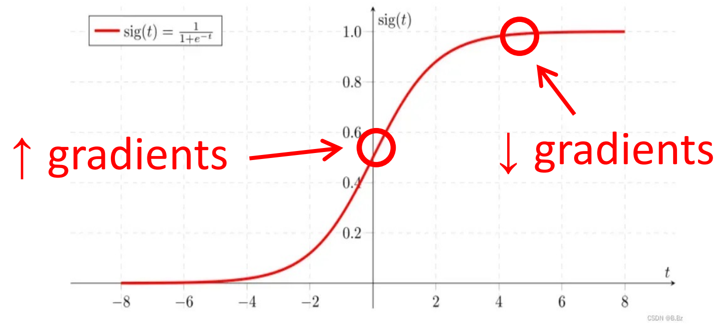
- Negative sampling
  - ✓ In implict feedback, pair-wise loss functions require to select negative sample from large scale of non-interacted items.
  - ✓ The effectiveness and efficiency of learning are limited.



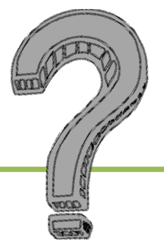
Negative samping in implict feedback

1.1 Negative Sampling in Recommendation

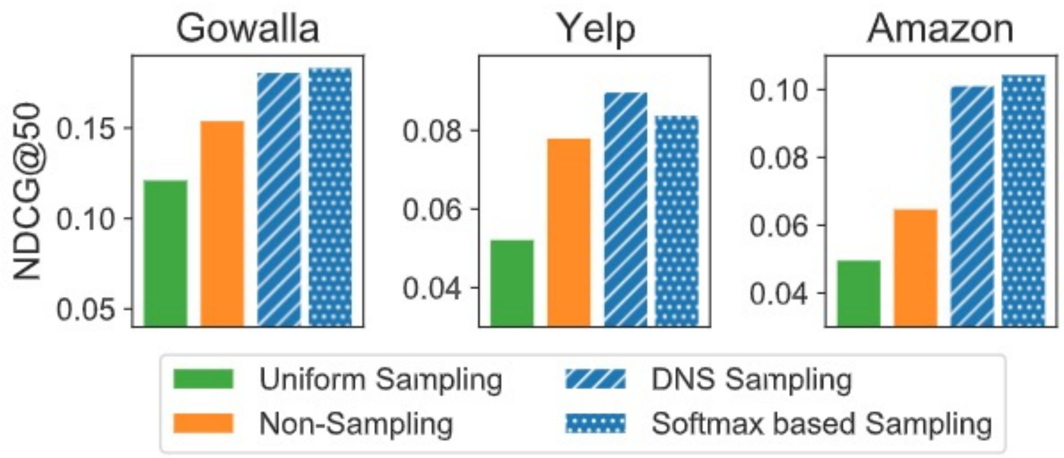
- **Classical sampling strategy**
  - ✓ Static Sampler: RNS, PNS
  - ✓ Adaptive Sampler: DNS, Softmax based Sampler



Well Convergence is always at the heart of understanding of hard negative sampling !



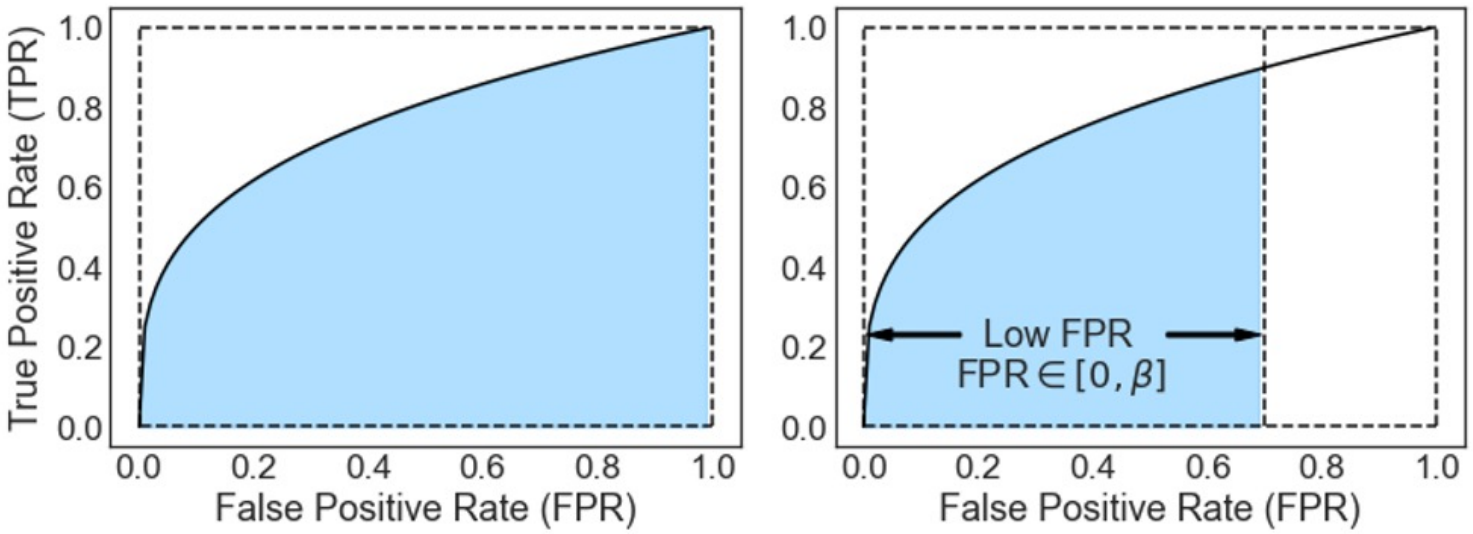
- **Empirical Experiments analysis**



- ✓ Full sampling: use all negative items to accelerate convergence.
- ✓ Well convergence may not be the only justification !

1.2 Revisit Hard Negative Sampling

- Different Opimization Objective
  - ✓ BPR optimizes AUC measure.
  - ✓ BPR equipped with hard negative sampling optimizes One way Partial AUC (OPAUC) measure. [Theoretical Analysis]



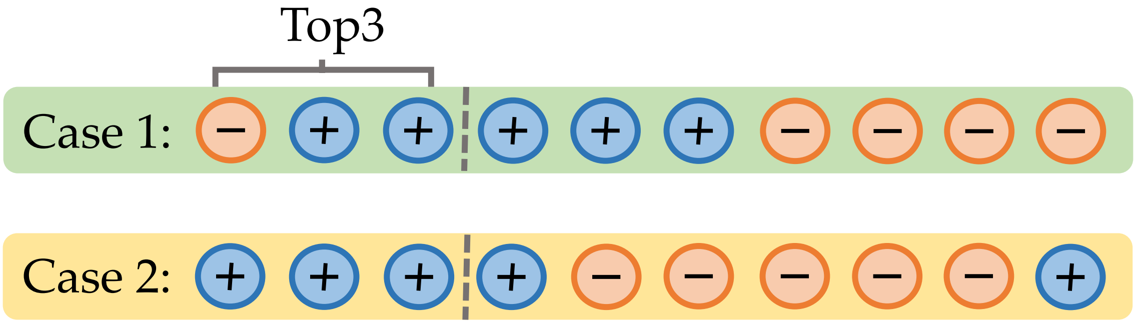
(a) AUC (b) One-Way Partial AUC (OPAUC)

		Label	
		True	False
Prediction	Positive	TP	FP
	Negative	FN	TN

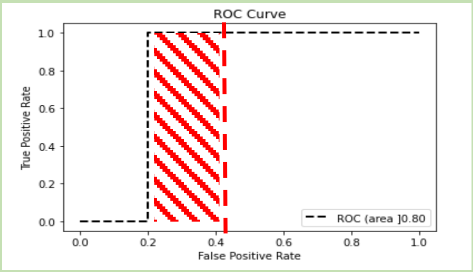
$$TPR = TP / (TP + FN)$$
$$FPR = FP / (FP + TN)$$

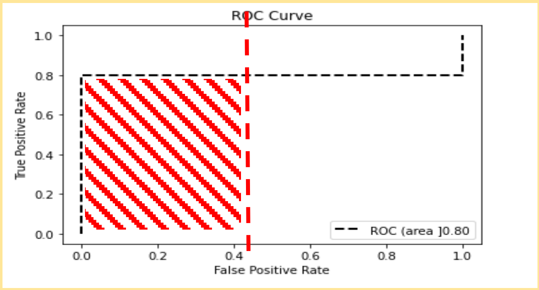
1.2 Revisit Hard Negative Sampling

◻ OPAUC and TopK evaluation measures



	AUC	OPAUC(0.4)	NDCG@3
Case 1	0.800	0.200	0.531
Case 2	0.800	0.320	1.000

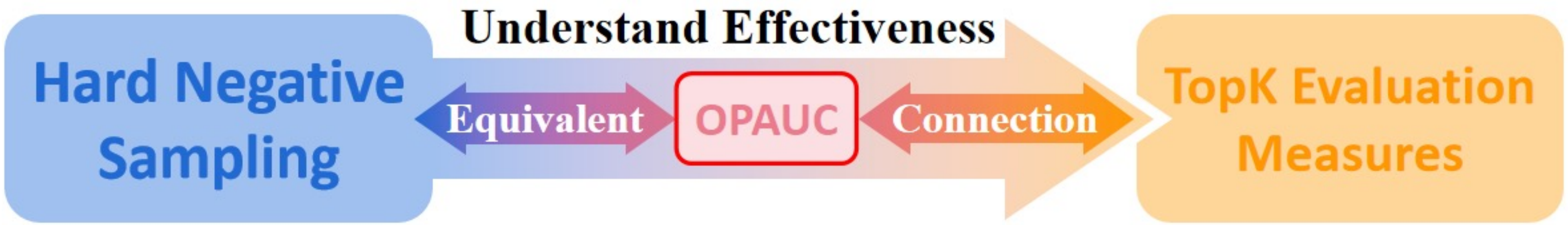




- ✓ Two simple cases that have the same overall ranking performance but quite different top ranking performance.
- ✓ OPAUC has stronger connection with TopK evaluation measures.

1.3 Overview of Theoretical Structure

- Why Hard Negative Sampling is effective?



- Future guidelines

  ✓ Hard negative sampling strategy should have hyper parameter to adjust the level of sampling hardness.

  ✓ The smaller the K we considered in Top$K$ evaluation measures, the harder the negative samples we should draw.

❖ Outline

1. Introduction
2. **Preliminary**
    - Hard Negative Sampling Strategies
    - Distributionally Robust Optimization (DRO)
3. Theoretical Analysis and Guidelines
4. Experiments

2.1 Hard Negative Sampling Strategies

- Implicit Feedback

$$\min_{\theta} \sum_{c \in C} \sum_{i \in I_c^+} E_{j \sim P_{ns}(j|c)} \left[\ell \left(r(c, i|\theta) - r(c, j|\theta) \right) \right],$$

$P_{ns}(j|c)$ denotes the negative sampling probability that a negative item $j \in I_c^-$ in the context c is drawn

- Dynamic Negative Sampling (DNS)

$$P_{ns}^{DNS}(j|c) = \begin{cases} \frac{1}{M}, & r_{cj} \in S_{I_c^-}^{\downarrow}[1, M] \\ 0, & r_{cj} \in others \end{cases},$$

$S_{I_c^-}^{\downarrow}[1, M] \subset I_c^-$ denotes the subset of negative samples who rank in topM.

- Softmax based Sampling

$$P_{ns}^{Softmax}(j|c) = \frac{\exp(r_{cj}/\tau)}{\sum_{k \in I_c^-} \exp(r_{ck}/\tau)} = \frac{\exp(r_{cij}/\tau)}{\sum_{k \in I_c^-} \exp(r_{cik}/\tau)},$$

2.2 Distributionally Robust Optimization (DRO)

☐ Formally Definition

✓ DRO aims to minimize the expected risk over the worst-case distribution Q, where Q is in a divergence ball around training distribution P.

$$\min_{\theta} \sup_{Q} E_Q \left[\mathcal{L}(f_\theta(\mathbf{x}), y)\right]$$
$$s.t.\ D_\phi(Q||P) \le \rho,$$

☐ Commen Divergence

✓ KL divergence $D_{KL}(Q||P) = \int \log(\frac{dQ}{dP})dQ,$

✓ CVaR divergence $D_{CVaR}(Q||P) = \sup \log(\frac{dQ}{dP}).$

❖ Outline

1. Introduction
2. Preliminary
3. **Theoretical Analysis and Guidelines**
   - Hard Negative Sampling Meets OPAUC
   - OPAUC Meets TopK Evaluation Measures
   - Hard Negative Sampling Understanding
4. Experiments

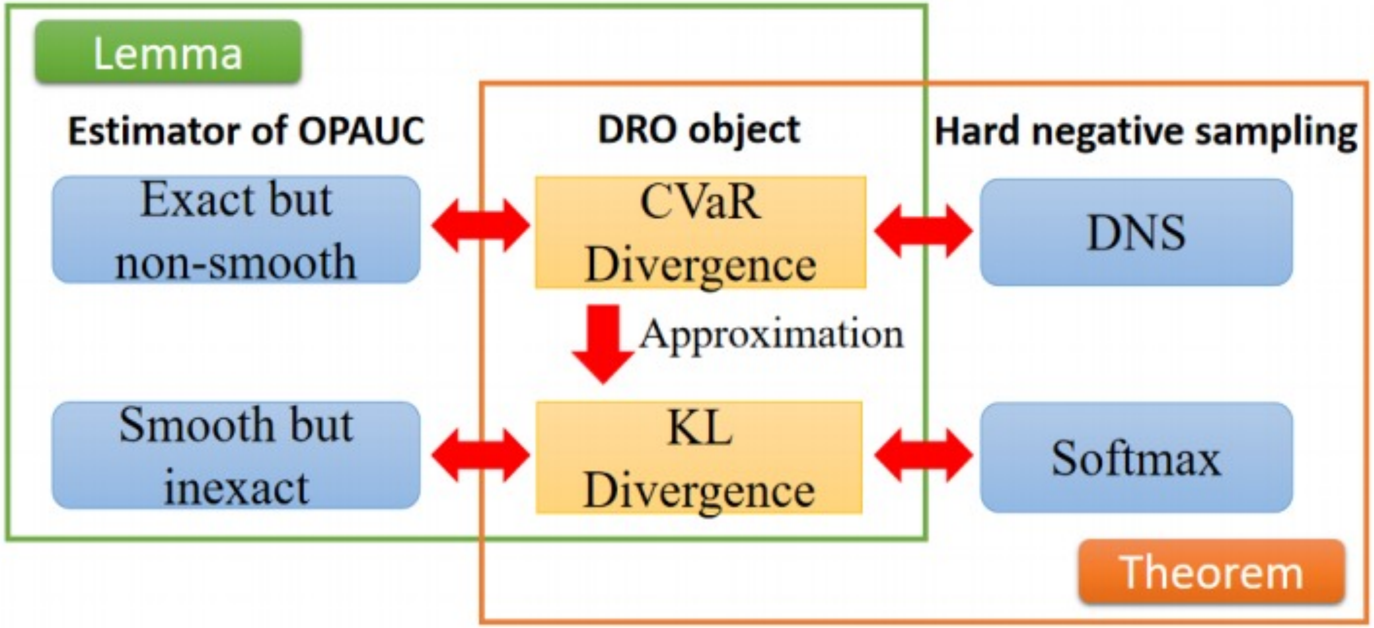
3.1 Hard Negative Sampling Meets OPAUC

☐ Arguement:

✓ The model equipped with hard negative sampling approximately optimizes OPAUC.

3.1 Hard Negative Sampling Meets OPAUC

Theoretical Structure

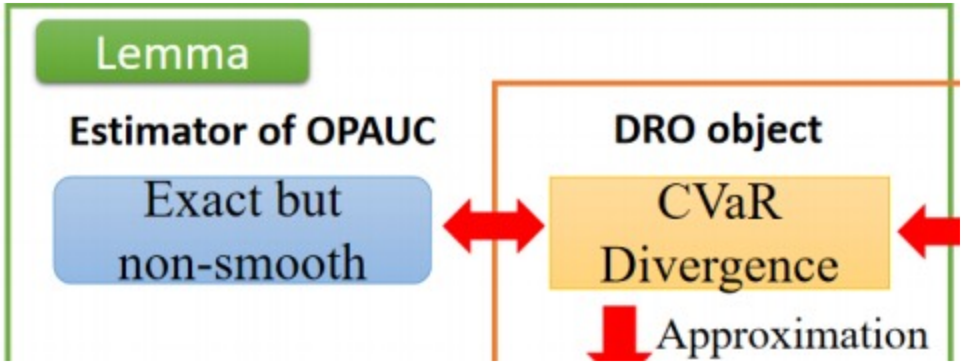


- ✓ The Lemma 1[1] shows the equivalence between estimator of OPAUC and DRO object.
- ✓ Based on DRO object, we prove the equivalence between hard negative sampling and OPAUC in Theorem 1 and Theorem 2.

[1] *Dixian Zhu et al. 2022, When AUC meets DRO: Optimizing Partial AUC for Deep Learning with Non-Convex Convergence Guarantee*

3.1 Hard Negative Sampling Meets OPAUC

Lemma 1



✓ OPAUC Estimator

$$\widehat{OPAUC}(\beta) = 1 - \frac{1}{|C|}\sum_{c\in C}\frac{1}{n_+}\frac{1}{n_-}\sum_{i\in\mathcal{I}_c^+}\sum_{j\in S^{\downarrow}_{\mathcal{I}_c^-}[1,n_-\beta]}\mathbb{I}(r_{ci} < r_{cj}),$$

$$\min_{\theta}\frac{1}{|C|}\sum_{c\in C}\frac{1}{n_+}\sum_{i\in\mathcal{I}_c^+}\frac{1}{n_-\beta}\sum_{j\in S^{\downarrow}_{\mathcal{I}_c^-}[1,n_-\beta]} L(c,i,j). \qquad (12)$$

✓ DRO object

$$\min_{\theta}\frac{1}{|C|}\sum_{c\in C}\frac{1}{n_+}\sum_{i\in\mathcal{I}_c^+}\max_{Q} E_Q\left[L(c,i,j)\right] \qquad (13)$$
$$s.t.\ D_\phi(Q||P_0) \le \rho,$$

- DRO over negative distribution P_0

Lemma 1 (Theorem 1 of [25]). *By choosing CVaR divergence* $D_\phi = D_{CVaR}(Q||P_0) = \sup\log(\frac{\mathrm{d}Q}{\mathrm{d}P_0})$, *then problem (13) reduces to*

$$\min_{\theta}\min_{\eta\ge 0}\frac{1}{|C|}\sum_{c\in C}\frac{1}{n_+}\sum_{i\in\mathcal{I}_c^+}\left\{\frac{1}{\exp(-\rho)}\cdot E_{j\sim P_0}\left[(L(c,i,j)-\eta_i)_+\right]+\eta_i\right\}, \qquad (14)$$

where P_0 denotes uniform distribution over $\mathcal{I}_c^-$. If ℓ is a monotonically decreasing function for $\ell(\cdot)\ge 0$ and setting $\beta = \exp(-\rho)$, the objective in (13) is equivalent to (12) of OPUAC(β).

3.1 Hard Negative Sampling Meets OPAUC

- Theorem 1

Theorem 1. *By choosing $P_{ns} = P_{ns}^{DNS}$ and $\boxed{M = n_{-}\beta,}$ then the DNS sampling based problem (1) is equivalent to (12) of $OPUAC(\beta)$.*

$$\min_{\theta} \sum_{c \in C} \sum_{i \in \mathcal{I}_c^+} E_{j \sim P_{ns}(j|c)} \left[\ell \left(r(c, i|\theta) - r(c, j|\theta) \right) \right],$$

$$P_{ns}^{DNS}(j|c) = \begin{cases} \frac{1}{M}, & r_{cj} \in S^{\downarrow}_{\mathcal{I}_c^-}[1, M] \\ 0, & r_{cj} \in others \end{cases}, \quad (1)$$

$$\longleftrightarrow$$

$$\min_{\theta} \frac{1}{|C|} \sum_{c \in C} \frac{1}{n_+} \sum_{i \in \mathcal{I}_c^+} \frac{1}{n_{-}\beta} \sum_{j \in S^{\downarrow}_{\mathcal{I}_c^-}[1, n_{-}\beta]} L(c, i, j). \quad (12)$$

Remark:

1) The DNS sampling based problem is an exact but non smooth estimator of OPUAC(β).

2) The hyperparameter M in DNS strategy directly determines βin OPAUC objective.

3.1 Hard Negative Sampling Meets OPAUC

☐ Theorem 2

Theorem 2. *By choosing* $P_{ns} = P_{ns}^{Softmax}$,

$$\boxed{\tau = \sqrt{\frac{\mathrm{Var}_j(L(c,i,j))}{-2\log\beta}},} \qquad (15)$$

$$\mathrm{Var}_j(L(c,i,j)) = E_{j\sim P_0}\left[\left(L(c,i,j) - E_{j\sim P_0}[L(c,i,j)]\right)^2\right], \qquad (16)$$

the softmax sampling based problem (1) is a surrogate version of (12) of $OPUAC(\beta)$.

$$\min_{\theta} \sum_{c\in C}\sum_{i\in I_c^+} E_{j\sim P_{ns}(j|c)}\left[\ell\left(r(c,i|\theta) - r(c,j|\theta)\right)\right],$$

$$P_{ns}^{Softmax}(j|c) = \frac{\exp(r_{cj}/\tau)}{\sum_{k\in I_c^-}\exp(r_{ck}/\tau)} = \frac{\exp(r_{cij}/\tau)}{\sum_{k\in I_c^-}\exp(r_{cik}/\tau)}, \qquad (1)$$

Remark:

1) **The softmax distribution based sampling problem is a smooth but inexact estimator of OPAUC(β).**
2) **We propose to use an adaptive τ, instead of a fixed τ in softmax distribution. This ensures the optimization objective OPAUC(β) remains the same during training**

3.1 Hard Negative Sampling Meets OPAUC

◻ Arguement:

- ✓ Compared to AUC measure, OPAUC(β) has stronger correlation with Top$K$ evaluation measures.
- ✓ A smaller K in Top$K$ evaluation measures has stronger correlation with a smaller β in OPAUC(β).

3.2 OPAUC Meets TopK Evaluation Measures

Theoretical Analysis

Theorem 3. *Suppose there are N_+ positive items and N_- negative items, where $N_+ > K$ and $N_- > K$. For any permutation of all items in descending order, we have*

$$\frac{1}{N_+}\left[\frac{N_+ + K - \sqrt{(N_+ + K)^2 - 4N_+N_- \times OPAUC(\beta)}}{2}\right] \leq Recall@K \leq \frac{1}{N_+}\left[\sqrt{N_+N_- \times OPAUC(\beta)}\right], \tag{17}$$

$$\frac{1}{K}\left[\frac{N_+ + K - \sqrt{(N_+ + K)^2 - 4N_+N_- \times OPAUC(\beta)}}{2}\right] \leq Precision@K \leq \frac{1}{K}\left[\sqrt{N_+N_- \times OPAUC(\beta)}\right], \tag{18}$$

where $\boxed{\beta = \frac{K}{N_-}}$.

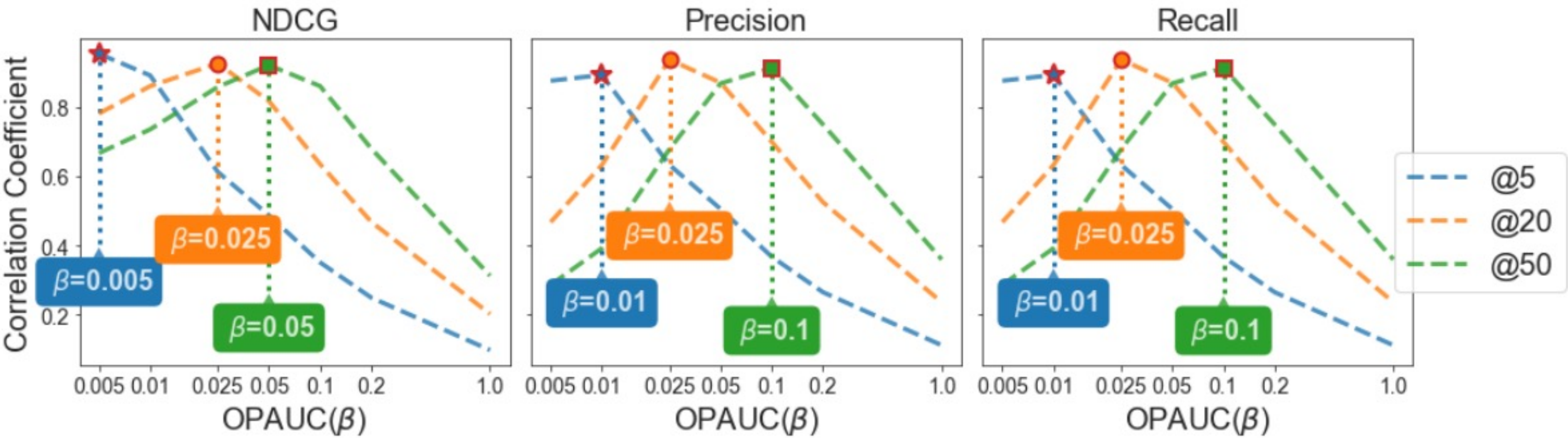
Remark:

1) The top$K$ evaluation measure Precision@K and Recall@K are higher and lower bounded by specific OPAUC(β), where $\beta = \frac{K}{N_-}$.

2) The smaller the K is, the smaller the β ($= \frac{K}{N_-}$) should be considered.

3.2 OPAUC Meets TopK Evaluation Measures

Simulation Experiments



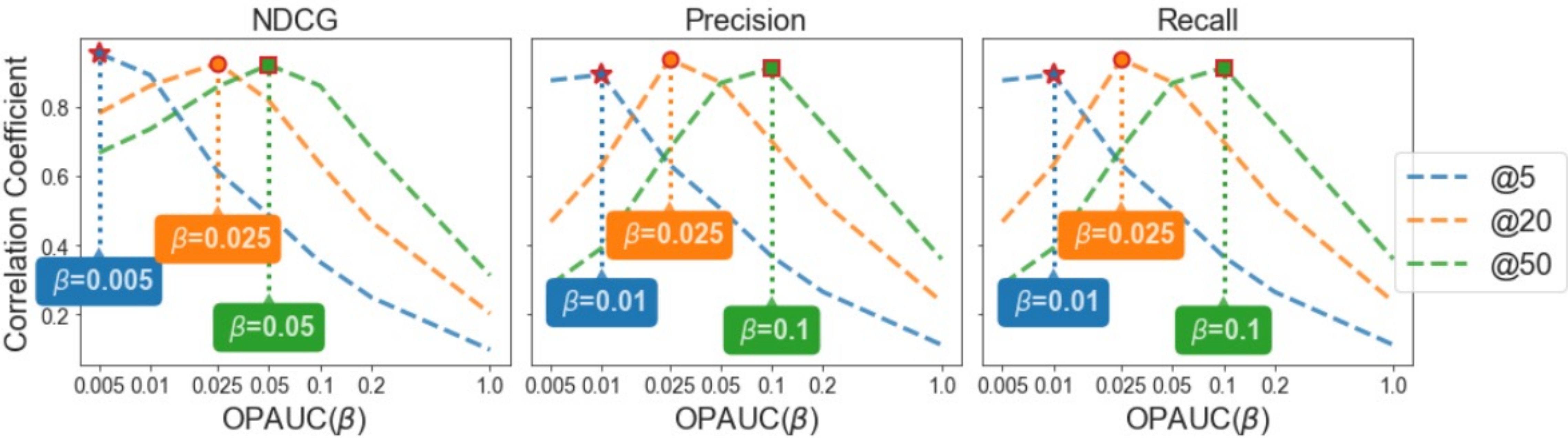
Remark:

1) The correlation coefficient of the highest point of the curve is much larger than the correlation coefficient when β is equal to 1.

2) Given a specific K in Top$K$ measure, the correlation coefficient with Norm_OPAUC(β) get the maximum value at a specific β.

3.2 OPAUC Meets TopK Evaluation Measures

☐ Simulation Experiments



Remark:

1) For different K, the peak of the curve varies according to β.

2) On the left side of the peak of the curve, we find that the correlation coefficient of NDCG@K descend more slowly than other two measures.

3.3 Hard Negative Sampling Understanding

- Corollary:

> **COROLLARY 1.** *Hard negative sampling approximately optimizes TopK evaluation measures while the parameter K is determined by the level of sampling hardness.*

- Guidelines:

  ✓ To adapt to different Top$K$ evaluation measures and datasets, hard negative sampling strategy should have hyperparameter to adjust the level of sampling hardness.

  ✓ The smaller the K we considered in Top$K$ evaluation measures, the harder the negative samples we should draw.

3.3 Hard Negative Sampling Understanding

- Converted Algorithms:

Algorithm 1 DNS (M, N)

Initialize θ
for $t = 1, \ldots, T$ **do**
  Sample a mini-batch $\mathcal{S} \in \mathcal{D}$
  for $(c, i) \in \mathcal{S}$ **do**
    Uniformly sample a mini-batch $\mathcal{S}'_c \in \mathcal{I}^-_c$, $|\mathcal{S}'_c| = N$.
    Let $\boxed{p_{ij} = \begin{cases} \frac{1}{M}, & r_{cj} \in \mathcal{S}^{\downarrow}_{S'} \\ 0, & r_{cj} \in others. \end{cases}}$
  end for
  Compute a gradient estimator ∇_t by

$$\nabla_t = \frac{1}{|\mathcal{S}|} \sum_{(c,i) \in \mathcal{D}} \sum_{j \in \mathcal{S}'} p_{ij} \nabla_\theta L(c, i, j).$$

  Update $\theta_{t+1} = \theta_t - \eta \nabla_t$.
end for

Algorithm 2 Softmax-v (ρ, N)

Initialize θ
for $t = 1, \ldots, T$ **do**
  Sample a mini-batch $\mathcal{S} \in \mathcal{D}$
  for $(c, i) \in \mathcal{S}$ **do**
    Uniformly sample a mini-batch $\mathcal{S}'_c \in \mathcal{I}^-_c$, $|\mathcal{S}'_c| = N$.
    Let $\boxed{p_{ij} = \frac{e^{\ell(r_{cij})/\tau}}{\sum_{k \in \mathcal{S}'} e^{\ell(r_{cik})/\tau}} \text{ and } \tau = \sqrt{\frac{\mathrm{Var}_j(L(c,i,j))}{2\rho}}}, j \in \mathcal{S}'_c$
  end for
  Compute a gradient estimator ∇_t by

$$\nabla_t = \frac{1}{|\mathcal{S}|} \sum_{(c,i) \in \mathcal{D}} \sum_{j \in \mathcal{S}'} p_{ij} \nabla_\theta L(c, i, j).$$

  Update $\theta_{t+1} = \theta_t - \eta \nabla_t$.
end for

❖ Outline

1. Introduction
2. Preliminary
3. Theoretical Analysis and Guidelines
4. Experiments
   - Q1: Can experiment results validate our guidelines?
   - Q2: How do fine tuned parameters benefit models?
   - Q3: Can the converted model outperform baselines?

(RQ1) Performance with Different Sampling Distributions

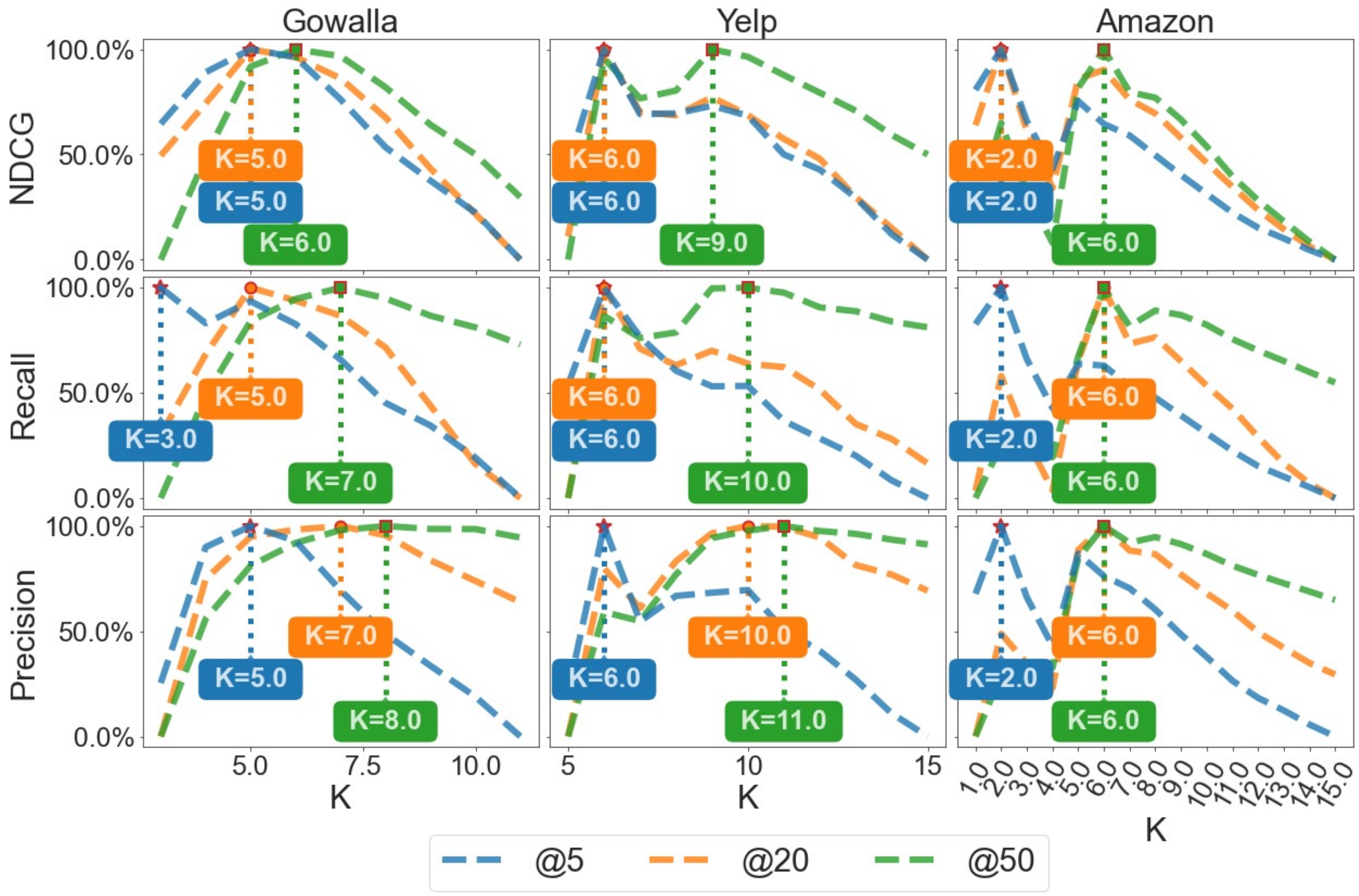
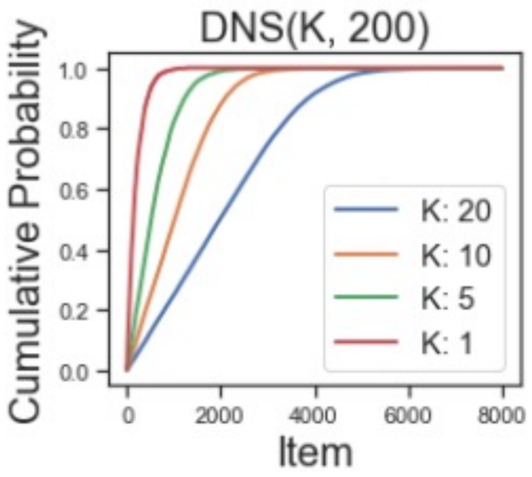


Figure 8: The effect of M in DNS(M, N), where N is set to 200, 200, 500 for Gowalla, Yelp and Amazon respectively.



Smaller M in DNS(M,N) means higher sampling hardness.

- ✓ For all datasets and all measures, the lower the *K* in Top*K* measures is, the smaller the M in DNS(M, N) when the curve achieve its maximum performance.

(RQ1) Performance with Different Sampling Distributions

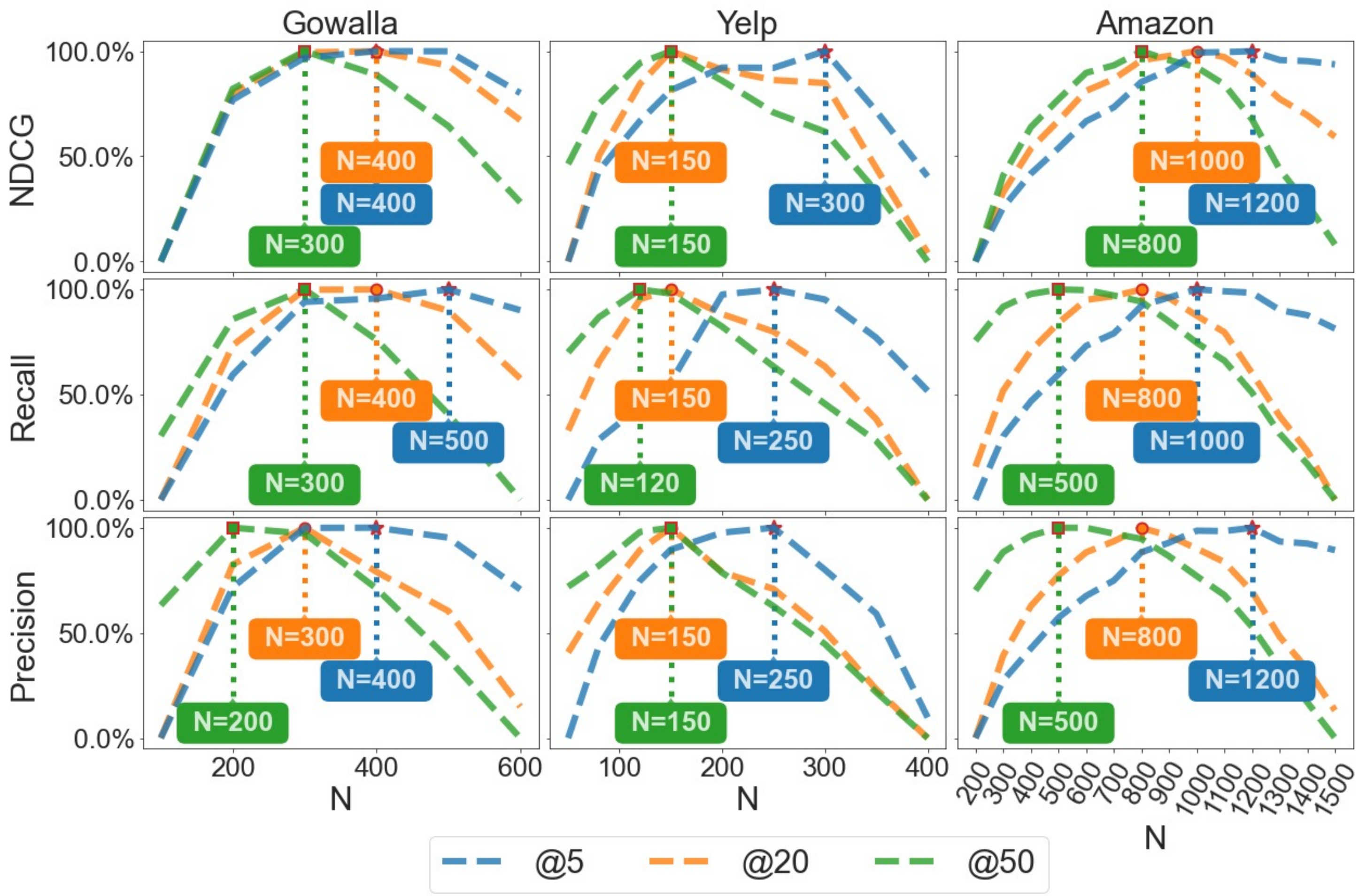
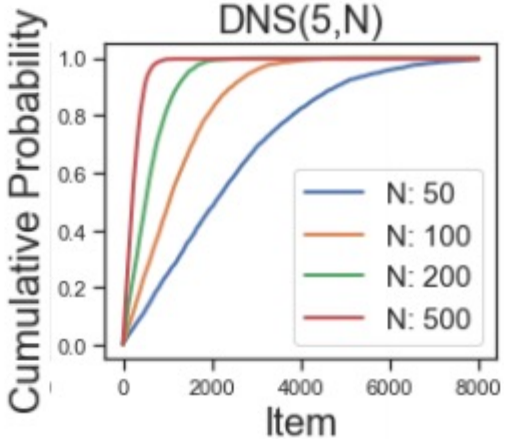


Figure 9: The effect of N in DNS(M, N), where K is set to 5 for all three datasets.



Larger N in DNS(M,N) means higher sampling hardness.

- ✓ For all datasets and all measures, the lower the K in Top$K$ measures is, the larger the N in DNS(M, N) when the curve achieve its maximum performance.

(RQ1) Performance with Different Sampling Distributions

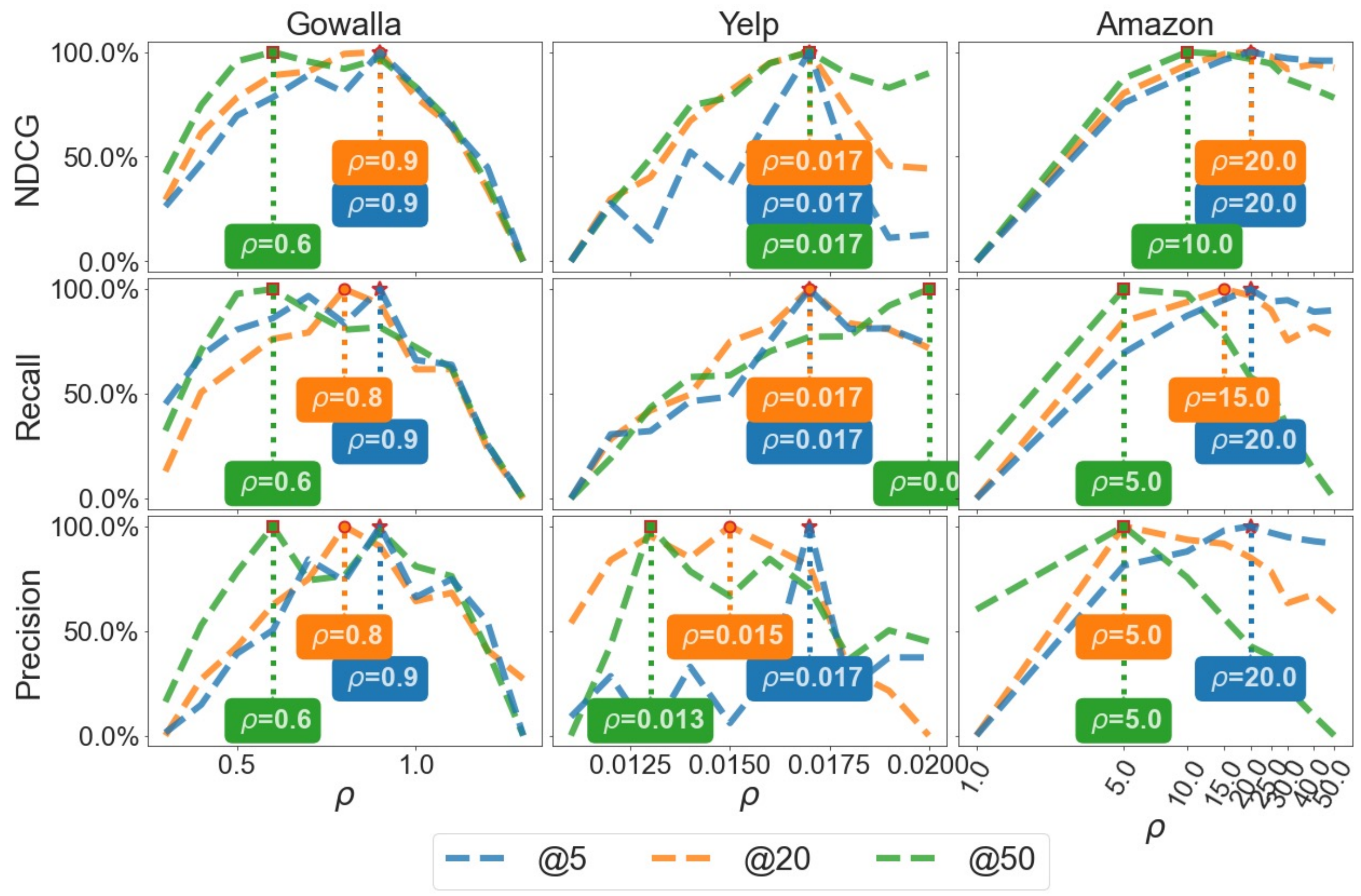
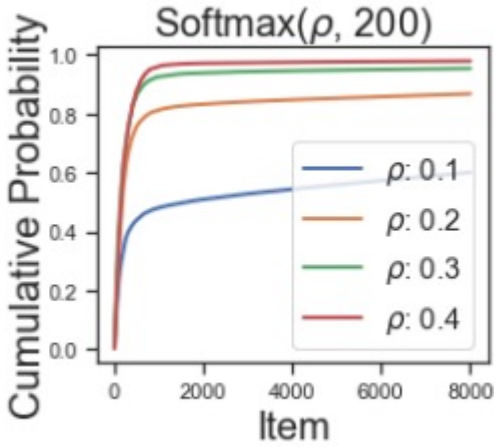


Figure 10: The effect of ρ in Softmax-v(ρ, N), where N is set to 200, 200, 500 for Gowalla, Yelp and Amazon respectively.



Larger ρ in Softmax-v(ρ,N) means higher sampling hardness.

- ✓ For all datasets and all measures, the lower the K in Top$K$ measures is, the larger the ρ in Softmax-v(ρ, N) when the curve achieve its maximum performance.

(RQ2, RQ3) Performance Comparison

Method	Gowalla		Yelp		Amazon	
	NDCG@50	Recall@50	NDCG@50	Recall@50	NDCG@50	Recall@50
BPR	0.1216	0.2048	0.0524	0.1083	0.0499	0.1171
AOBPR	0.1385	0.2369	0.0675	0.1354	0.0563	0.1303
WARP	0.1555	0.2650	0.0636	0.1334	0.0560	0.1217
IRGAN	0.1443	0.2242	0.0695	0.1367	0.0627	0.1395
DNS	0.1412	0.1839	0.0693	0.1425	0.0615	0.1378
PRIS(U)	0.1567	0.2625	0.0789	0.1603	0.0793	0.1695
PRIS(P)	0.1521	0.2449	0.0780	0.1597	0.0797	0.1700
AdaSIR(U)	0.1545	0.2565	0.0756	0.1512	0.0744	0.1562
AdaSIR(P)	0.1553	0.2567	0.0764	0.1514	0.0781	0.1627
Kernel	0.1399	0.2264	0.0658	0.1315	0.0700	0.1495
DNS(*)	0.1811	0.2989	**0.0899**	**0.1774**	0.1014	0.1833
Softmax-v	**0.1837**	**0.2993**	0.0840	0.1690	**0.1046**	**0.1937**

Remark:

1) **Benefited from the adjustable sampling hardness, the converted DNS(M*, N) and Softmax-v significantly outperform their original versions.**

2) **The converted hard negative sampling methods perform state-of-the-art baselines.**

5. Conclusion

1. We prove that the model equipped with hard negative sampling approximately optimizes OPAUC, where DNS is an exact estimator and softmax based sampling is a soft estimator.
2. We conduct theoretical analysis, simulation studies, and real world experiments to validate the stronger correlation between OPAUC and Top$K$ evaluation measures.
3. We provide two important guidelines on how to design hard negative sampling strategies. Through theoretical analysis and experiments analysis, we conclude that the smaller the K in Top$K$ measure is, the harder the negative items we should sample.