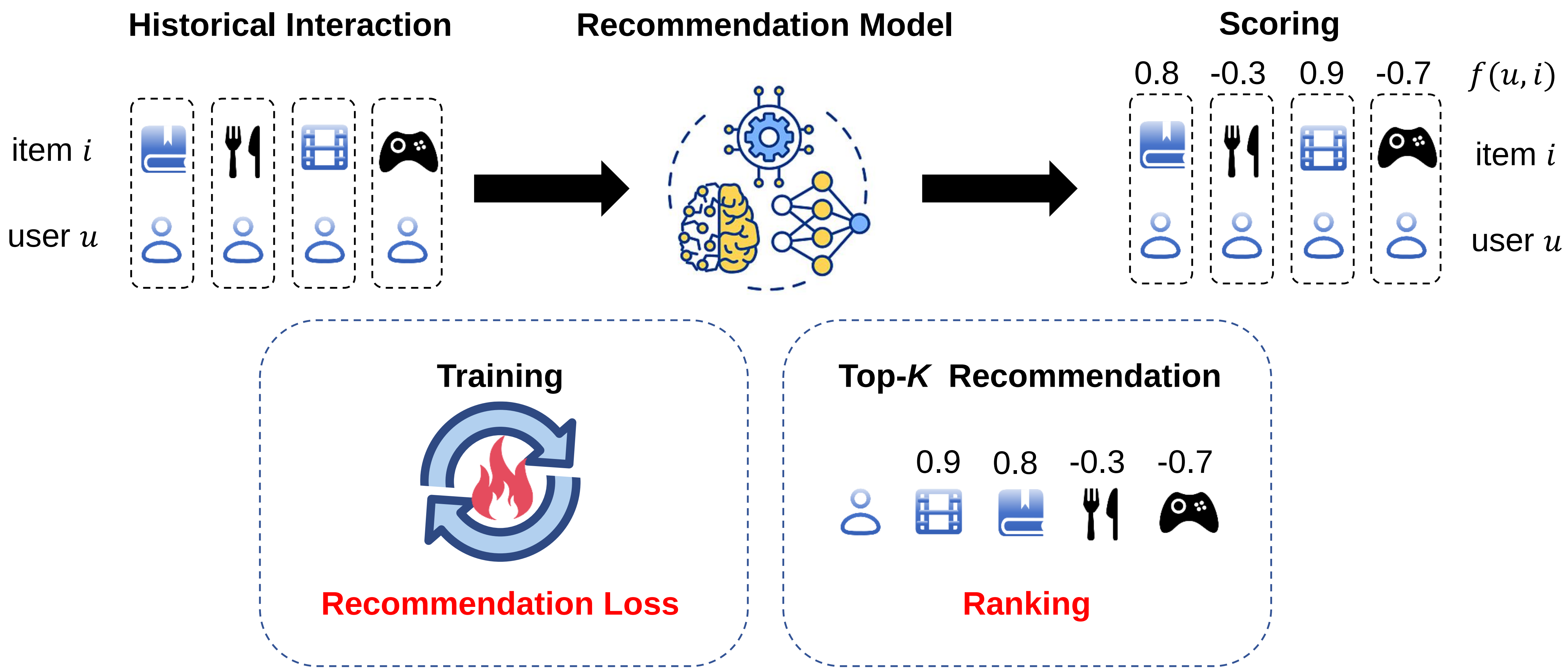
NEURAL INFORMATION PROCESSING SYSTEMS

PSL: Rethinking and Improving Softmax Loss from Pairwise Perspective for Recommendation

Weiqin Yang, Jiawei Chen, Xin Xin,
Sheng Zhou, Binbin Hu, Yan Feng, Chun Chen, Can Wang

Zhejiang University, Shandong University, Ant Group

Recommender Systems



Recommendation Loss

Pointwise Loss (e.g., BCE and MSE) 1: pos. 0: neg.

- Treats recommendation as a **binary classification** or **regression** problem
- Applied to each positive and negative instance separately, i.e., **pointwise score** $f(u,i)$

$$\mathcal{L}_{\text{pointwise}}(u) = -\sum_{i \in \mathcal{P}_u} \log(\varphi^+(f(u,i))) - \sum_{j \in \mathcal{I} \setminus \mathcal{P}_u} \log(\varphi^-(f(u,j)))$$

where $\mathcal{I}$ is the item set, $\mathcal{P}_u$ is the positive item set of user u , φ^+, φ^- are activation functions.

Recommendation Loss

Pointwise Loss (e.g., BCE and MSE) 1: pos. 0: neg.

- Treats recommendation as a **binary classification** or **regression** problem
- Applied to each positive and negative instance separately, i.e., **pointwise score** $f(u,i)$

$$\mathcal{L}_{\text{pointwise}}(u) = -\sum_{i \in \mathcal{P}_u} \log(\varphi^+(f(u,i))) - \sum_{j \in \mathcal{I} \setminus \mathcal{P}_u} \log(\varphi^-(f(u,j)))$$

where $\mathcal{I}$ is the item set, $\mathcal{P}_u$ is the positive item set of user u, φ^+, φ^- are activation functions.

Pairwise Loss (BPR) 0.8: pos. > -0.7: neg.

- Treats recommendation as learning a **partial order** among items
- Applied to each positive-negative item pair, i.e., **pairwise score** $\boxed{d_{uij} = f(u,j) - f(u,i)}$

$$\mathcal{L}_{\text{BPR}}(u) = \sum_{i \in \mathcal{P}_u} \sum_{j \in \mathcal{I} \setminus \mathcal{P}_u} \log \sigma(\boxed{f(u,j) - f(u,i)})$$

where σ is activation function.

ranking loss

Softmax Loss

Softmax Loss (SL)

0.8: pos. > -0.7: neg. & -0.3: neg.

- Maximizes the **probability** of recommending positive items
- **Ranks** positive items higher than all negative items

$$\mathcal{L}_{\mathrm{SL}}(u) = -\sum_{i \in \mathcal{P}_u} \log \left(\frac{\exp(f(u,i)/\tau)}{\sum_{j \in \mathcal{I}} \exp(f(u,j)/\tau)} \right)$$

τ : temperature

ranking loss

Softmax Loss

Softmax Loss (SL)

0.8: pos. > -0.7: neg. & -0.3: neg.

- Maximizes the **probability** of recommending positive items
- **Ranks** positive items higher than all negative items

$$\mathcal{L}_{\mathrm{SL}}(u) = -\sum_{i \in \mathcal{P}_u} \log \left(\frac{\exp(f(u,i)/\tau)}{\sum_{j \in \mathcal{I}} \exp(f(u,j)/\tau)} \right)$$

τ : temperature

ranking loss

Question: Can SL be expressed in **pairwise form**?

- Recall the **pairwise score** $d_{uij} = f(u,j) - f(u,i)$

Softmax Loss

Softmax Loss (SL) 0.8: pos. > -0.7: neg. & -0.3: neg.

- Maximizes the **probability** of recommending positive items
- **Ranks** positive items higher than all negative items

$$\mathcal{L}_{\mathrm{SL}}(u) = -\sum_{i \in \mathcal{P}_u} \log \left(\frac{\exp(f(u,i)/\tau)}{\sum_{j \in \mathcal{I}} \exp(f(u,j)/\tau)} \right)$$

τ : temperature

ranking loss

Question: Can SL be expressed in **pairwise form**?

- Recall the **pairwise score** $d_{uij} = f(u,j) - f(u,i)$

SL (pairwise form)

$$\mathcal{L}_{\mathrm{SL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \exp(d_{uij}/\tau) \right)$$

ranking loss

DCG Surrogate Loss

Q: Why pairwise perspective?

A: Only pairwise loss has the potential to be interpreted as a **surrogate loss for ranking metrics**, such as **DCG** (Discounted Cumulative Gain) and **MRR** (Mean Reciprocal Rank).

In fact, we have the following inequalities (omitting irrelevant constants):

$$-\log \mathrm{DCG}(u) \leq -\log \mathrm{MRR}(u) \leq \sum_{i \in \mathcal{P}_u} \log \pi_u(i) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \delta(d_{uij}) \right)$$

where $\pi_u(i)$ is the ranking position of item i according to user u 's preference, and $\delta(\cdot)$ is the Heaviside step function.

DCG Surrogate Loss

Q: Why pairwise perspective?

A: Only pairwise loss has the potential to be interpreted as a **surrogate loss for ranking metrics**, such as **DCG** (Discounted Cumulative Gain) and **MRR** (Mean Reciprocal Rank).

In fact, we have the following inequalities (omitting irrelevant constants):

$$-\log \mathrm{DCG}(u) \le -\log \mathrm{MRR}(u) \le \sum_{i \in \mathcal{P}_u} \log \pi_u(i) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \boxed{\delta(d_{uij})} \right)$$

where $\pi_u(i)$ is the ranking position of item i according to user u 's preference, and $\delta(\cdot)$ is the Heaviside step function.

SL (pairwise form)

$$-\log \mathrm{DCG}(u) \le \mathcal{L}_{\mathrm{SL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \boxed{\exp(d_{uij}/\tau)} \right)$$

surrogate:

$$\delta(d_{uij}) \le \exp\left(d_{uij}/\tau\right)$$

Limitations of Softmax Loss

Limitation 1: SL is not tight enough as a DCG surrogate loss.

- The gap between $\delta(d_{uij})$ in DCG and its surrogate activation $\exp(d_{uij}/\tau)$ in SL is significant when d_{uij} increases from 0.
- Leading to **suboptimal accuracy**.

Limitation 2: SL is highly sensitive to false negative noise.

- The gradient is exponential w.r.t. d_{uij} , while false negatives often have large d_{uij}
- Leading to **poor noise resistance and robustness**.

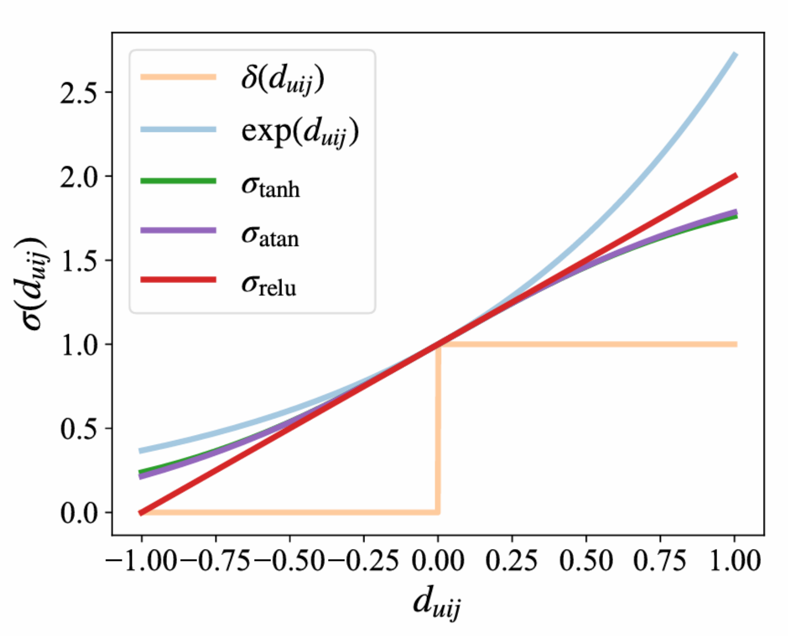
exp is NOT suitable for SL !!!

Pairwise Softmax Loss

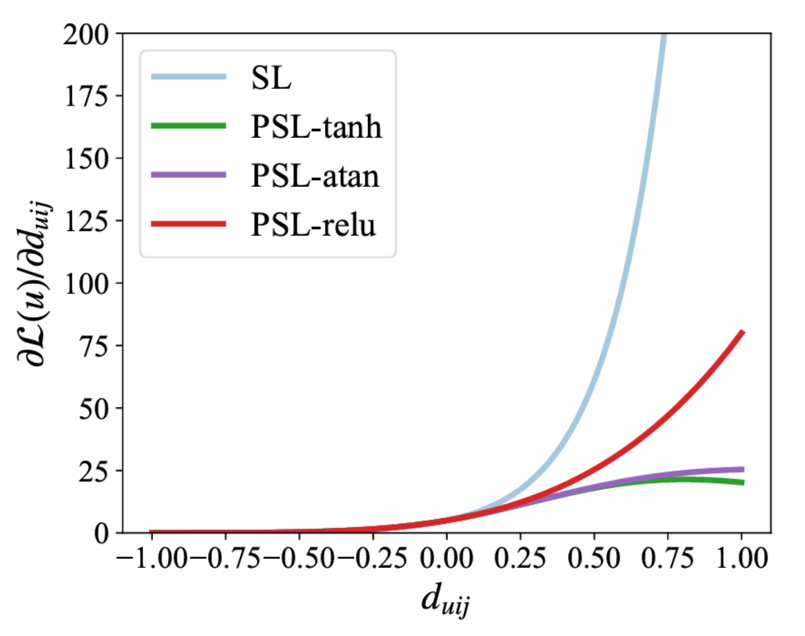
Our work: **Pairwise Softmax Loss (PSL)**, A general family of losses, which **replace exp in SL with other surrogate activations** σ, and **adjust the position of temperature** τ :

$$\mathcal{L}_{\mathrm{SL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \exp(d_{uij}/\tau) \right) \quad \Longrightarrow \quad \mathcal{L}_{\mathrm{PSL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \sigma(d_{uij})^{1/\tau} \right)$$

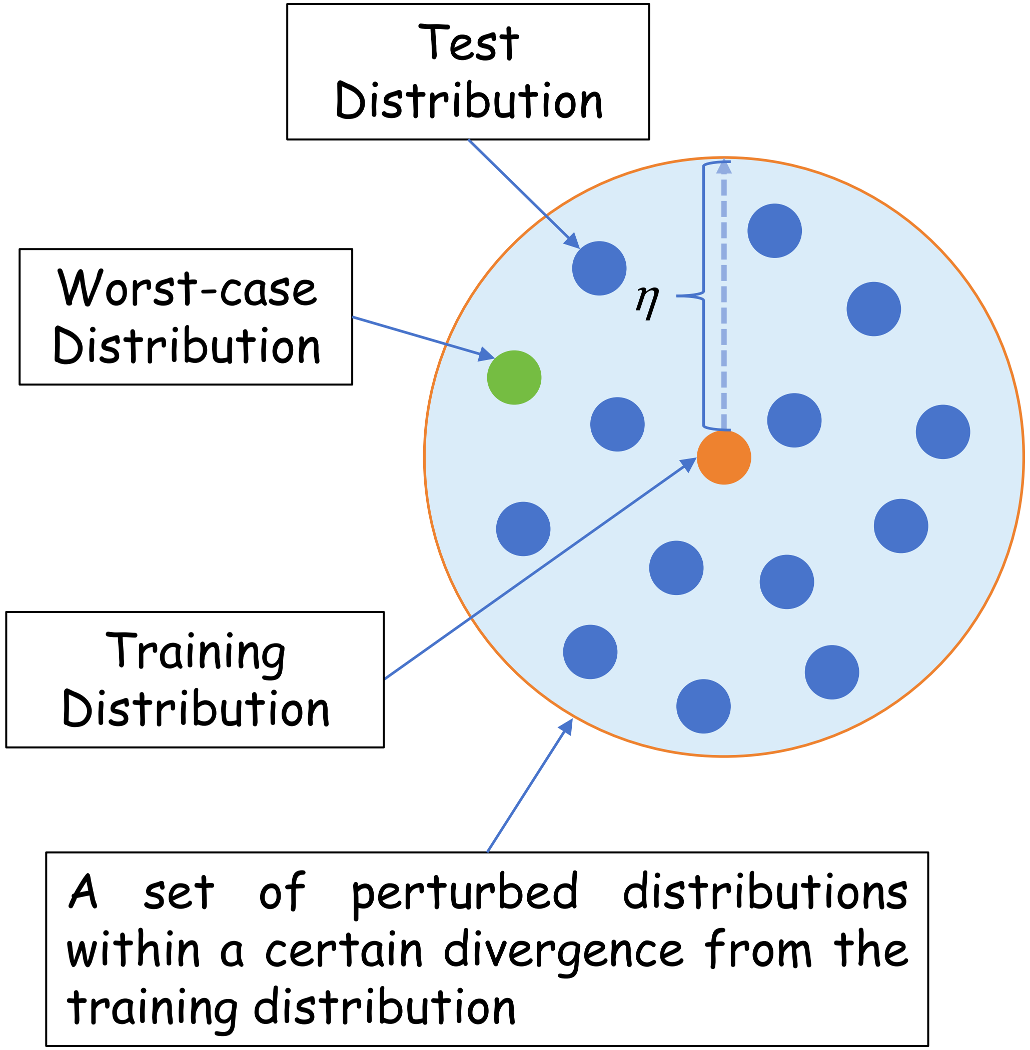
① Accuracy: Tighter surrogate for ranking metrics



② Noise Robustness More moderate gradient



PSL = BPR+DRO



DRO (Distributionally Robust Optimization):

A **robust optimization framework** against distribution shifts in out-of-distribution (**OOD**) scenarios.

- DRO optimizes for the worst-case perturbed distributions.

③ PSL is a DRO-empowered BPR loss

- PSL has better **OOD robustness** compared to BPR.
- This theorem establishes a **theoretical connection** among **pairwise losses**.

Experiments

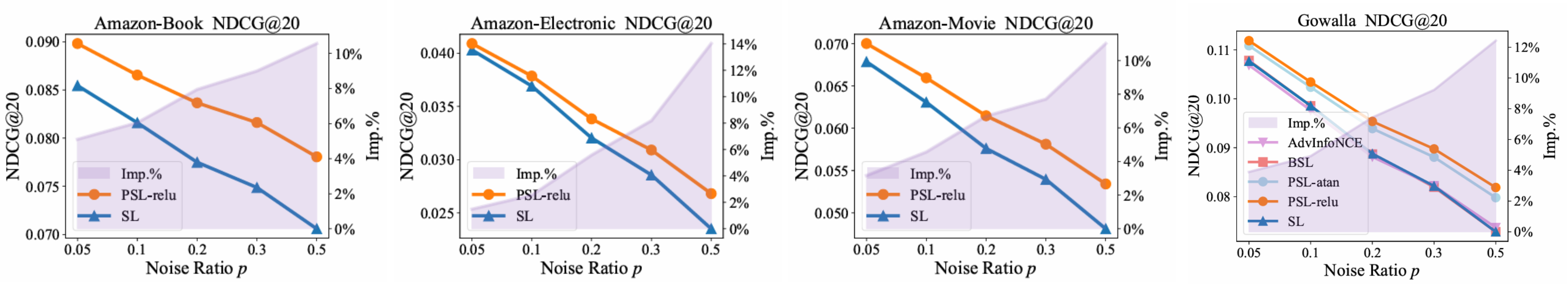
IID setting (Accuracy)

Loss	Amazon-Book		Amazon-Electronic		Amazon-Movie		Gowalla	
	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
BPR [10]	0.0665	0.0453	0.0816	0.0527	0.0916	0.0608	0.1355	0.1111
LLPAUC [44]	0.1150	0.0811	0.0821	0.0499	0.1271	0.0883	0.1610	0.1189
SL [11]	0.1559	0.1210	0.0821	0.0529	0.1286	0.0929	0.2064	0.1624
AdvInfoNCE [38]	0.1557	0.1172	0.0829	0.0527	0.1293	0.0934	0.2067	0.1627
BSL [15]	0.1563	0.1212	0.0834	0.0530	0.1288	0.0931	0.2071	0.1630
PSL-tanh	0.1567	0.1225	0.0832	0.0535	0.1297	0.0941	0.2088	0.1646
PSL-atan	0.1567	0.1226	0.0832	0.0535	0.1296	0.0941	0.2087	0.1646
PSL-relu	**0.1569**	**0.1227**	**0.0838**	**0.0541**	**0.1299**	**0.0945**	**0.2089**	**0.1647**
Imp.%	+1.40%*		+2.31%*		+1.72%*		+1.42%*	

OOD setting (OOD Robustness)

Loss	Amazon-CD		Amazon-Electronic		Gowalla		Yelp2018	
	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
BPR [10]	0.0518	0.0318	0.0132	0.0069	0.0382	0.0273	0.0118	0.0072
LLPAUC [44]	0.1103	0.0764	0.0225	0.0134	0.0729	0.0522	0.0324	0.0210
SL [11]	0.1184	0.0815	0.0230	0.0142	0.1006	0.0737	0.0349	0.0224
AdvInfoNCE [38]	0.1189	0.0818	0.0228	0.0139	0.0927	0.0676	0.0348	0.0223
BSL [15]	0.1184	0.0815	0.0231	0.0142	0.1006	0.0738	0.0351	0.0225
PSL-tanh	0.1202	0.0834	0.0239	0.0146	0.1013	0.0748	0.0357	0.0228
PSL-atan	0.1202	0.0835	0.0239	0.0146	0.1013	0.0748	**0.0358**	0.0228
PSL-relu	**0.1203**	**0.0839**	**0.0241**	**0.0149**	**0.1014**	**0.0752**	**0.0358**	**0.0229**
Imp.%	+3.01%*		+5.02%*		+2.02%*		+2.05%*	

Noise setting (Noise Resistance)



Metrics: NDCG@20, Recall@20

Thank You!

浙江大學
ZHEJIANG UNIVERSITY

arXiv

OpenReview

Weiqin Yang
(1st auth.)

Jiawei Chen
(corr. auth.)