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

Weiqin Yang [†‡]
Zhejiang University
tinysnow@zju.edu.cn

Jiawei Chen [*†‡§]
Zhejiang University
sleepyhunt@zju.edu.cn

Xin Xin
Shandong University
xinxin@sdu.edu.cn

Sheng Zhou
Zhejiang University
zhousheng_zju@zju.edu.cn

Binbin Hu
Ant Group
bin.hbb@antfin.com

Yan Feng [†‡]
Zhejiang University
fengyan@zju.edu.cn

Chun Chen [†‡]
Zhejiang University
chenc@zju.edu.cn

Can Wang [†§]
Zhejiang University
wcan@zju.edu.cn

Abstract

Softmax Loss (SL) is widely applied in recommender systems (RS) and has demonstrated effectiveness. This work analyzes SL from a pairwise perspective, revealing two significant limitations: 1) the relationship between SL and conventional ranking metrics like DCG is not sufficiently tight; 2) SL is highly sensitive to false negative instances. Our analysis indicates that these limitations are primarily due to the use of the exponential function. To address these issues, this work extends SL to a new family of loss functions, termed Pairwise Softmax Loss (PSL), which replaces the exponential function in SL with other appropriate activation functions. While the revision is minimal, we highlight three merits of PSL: 1) it serves as a tighter surrogate for DCG with suitable activation functions; 2) it better balances data contributions; and 3) it acts as a specific BPR loss enhanced by Distributionally Robust Optimization (DRO). We further validate the effectiveness and robustness of PSL through empirical experiments. The code is available at https://github.com/Tiny-Snow/IR-Benchmark.

1 Introduction

Nowadays, recommender systems (RS) have permeated various personalized services [1–4]. What sets recommendation apart from other machine learning tasks is its distinctive emphasis on ranking [5]. Specifically, RS aims to retrieve positive items in higher ranking positions (i.e., giving larger prediction scores) over others and adopts specific ranking metrics (e.g., DCG [6] and MRR [7]) to evaluate its performance.

The emphasis on ranking inspires a surge of research on loss functions in RS. Initial studies treated recommendation primarily as a classification problem, utilizing pointwise loss functions (e.g., BCE [8], MSE [9]) to optimize models. Recognizing the inherent ranking nature of RS, pairwise loss

[*]Corresponding author.
[†]State Key Laboratory of Blockchain and Data Security, Zhejiang University.
[‡]College of Computer Science and Technology, Zhejiang University.
[§]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3844

functions (e.g., BPR [10]) were introduced to learn a partial ordering among items. More recently, Softmax Loss (SL) [11] has integrated contrastive learning paradigms [12, 13], augmenting positive items as compared with negative ones, achieving state-of-the-art (SOTA) performance.

While SL has proven effective, it still suffers from **two limitations**: 1) SL can be used to approximate ranking metrics, e.g., DCG and MRR [11, 14], but their relationships are not sufficiently tight. Specifically, SL uses the exponential function $\exp(\cdot)$ as the *surrogate activation* to approximate the Heaviside step function in DCG, resulting in a notable gap, especially when the surrogate activation takes larger values. 2) SL is sensitive to noise (e.g., false negatives [15]). Gradient analysis reveals that SL assigns higher weights to negative instances with large prediction scores, while the weights are rather skewed and governed by the exponential function. This characteristic renders the model highly sensitive to false negative noise. Specifically, false negative instances are common in RS, as a user's lack of interaction with an item might stem from unawareness rather than disinterest [16–18]. These instances would receive disproportionate emphasis, potentially dominating the training direction, leading to performance degradation and training instability.

To address these challenges, we propose a new family of loss functions, termed **Pairwise Softmax Loss (PSL)**. PSL first reformulates SL in a pairwise manner, where the loss is applied to the score gap between positive-negative pairs. Such pairwise perspective is more fundamental to recommendation as the ranking metrics are also pairwise dependent. Recognizing that the primary weakness of SL lies in its use of the exponential function, PSL replaces this with other surrogate activations. While this extension is straightforward, it brings significant theoretical merits:

- **Tighter surrogate for ranking metrics.** We establish theoretical connections between PSL and conventional ranking metrics, e.g., DCG. By choosing appropriate surrogate activations, such as ReLU or Tanh, we demonstrate that PSL achieves a tighter DCG surrogate loss than SL.
- **Control over the weight distribution.** PSL provides flexibility in choosing surrogate activations that control the weight distribution of training instances. By substituting the exponential function with an appropriate surrogate activation, e.g., ReLU or Tanh, PSL can mitigate the excessive impact of false negatives, thus enhancing robustness to noise.
- **Theoretical connections with BPR loss.** Our analyses reveal that optimizing PSL is equivalent to performing Distributionally Robust Optimization (DRO) [19] over the conventional pairwise loss BPR [10]. DRO is a theoretically sound framework where the optimization is not only on a fixed empirical distribution but also across a set of distributions with adversarial perturbations. This DRO characteristic endows PSL with stronger generalization and robustness against out-of-distribution (OOD), especially given that such distribution shifts are common in RS, e.g., shifts in user preference and item popularity [16, 20, 21].

Our analyses underscore the theoretical effectiveness and robustness of PSL. To empirically validate these advantages, we implement PSL with typical surrogate activations (Tanh, Atan, ReLU) and conduct extensive experiments on four real-world datasets across three experimental settings: 1) IID setting [22] where training and test distributions are identically distributed [23]; 2) OOD setting [24] with distribution shifts in item popularity; 3) Noise setting [15] with a certain ratio of false negatives. Experimental results demonstrate the superiority of PSL over existing losses in terms of recommendation accuracy, OOD robustness, and noise resistance.

2 Preliminaries

Task formulation. We will conduct our discussion in the scope of collaborative filtering (CF) [25], a widely-used recommendation scenario. Given the user set $\mathcal{U}$ and item set $\mathcal{I}$, CF dataset $\mathcal{D} \subset \mathcal{U} \times \mathcal{I}$ is a collection of observed interactions, where each instance $(u, i) \in \mathcal{D}$ means that user u has interacted with item i (e.g., clicks, reviews, etc). For each user u, we denote $\mathcal{P}_u = \{i \in \mathcal{I} : (u, i) \in \mathcal{D}\}$ as the set of positive items of u, while $\mathcal{I} \setminus \mathcal{P}_u$ represents the negative items.

The goal of recommendation is to learn a recommendation model, or essentially a scoring function $f(u, i) : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ that quantifies the preference of user u on item i accurately. Modern RS often adopts an embedding-based paradigm [26]. Specifically, the model maps user u and item i into d-dim embeddings $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, and predicts their preference score $f(u, i)$ based on embedding similarity. The cosine similarity is commonly utilized in RS and has demonstrated particular effectiveness [27]. Here we set $f(u, i) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} \cdot \frac{1}{2}$, where the scaling factor $\frac{1}{2}$ is introduced for faciliating analyses

and can be absorbed into the temperature hyperparameter (τ). The scores $f(u,i)$ are subsequently utilized to rank items for generating recommendations.

Ranking metrics. The Discounted Cumulative Gain (DCG) [6] is a prominent ranking metric for evaluating the recommendation quality. Formally, for each user u, DCG is calculated as follows:

$$\mathrm{DCG}(u) = \sum_{i \in \mathcal{P}_u} \frac{1}{\log_2(1 + \pi_u(i))} \tag{2.1}$$

where $\pi_u(i)$ is the ranking position of item i in the ranking list sorted by the scores $f(u,i)$. DCG quantifies the cumulative gain of positive items, discounted by their ranking positions. Similarly, the Mean Reciprocal Rank (MRR) [7, 28] is another popular ranking metric using the reciprocal of the ranking position as the gain, i.e., $\mathrm{MRR}(u) = \sum_{i \in \mathcal{P}_u} 1/\pi_u(i)$. Additionally, other metrics such as Recall [29], Precision [29], and AUC [30] are also utilized in RS [29]. Compared to these metrics, DCG and MRR focus more on the top-ranked recommendations, thus attracting increasing attention in RS [11, 31]. In this work, we aim to explore the surrogate loss for DCG and MRR.

Recommendation losses. To train recommendation models effectively, a series of recommendation losses has been developed. Recent work on loss functions can mainly be classified into three types:

- **Pointwise loss** (e.g., BCE [8], MSE [9], etc.) formulates recommendation as a specific classification or regression task, and the loss is applied to each positive and negative instance separately. Specifically, for each user u, the pointwise loss is defined as

$$\mathcal{L}_{\text{pointwise}}(u) = -\sum_{i \in \mathcal{P}_u} \log(\varphi^+(f(u,i))) - \sum_{j \in \mathcal{I} \setminus \mathcal{P}_u} \log(\varphi^-(f(u,j))) \tag{2.2}$$

  where $\varphi^+(\cdot)$ and $\varphi^-(\cdot)$ are the activation functions adapted for different loss choices.
- **Pairwise loss** (e.g., BPR [10], etc.) optimizes partial ordering among items, which is applied to the score gap between negative-positive pairs. BPR [10] is a representative pairwise loss, which is defined as

$$\mathcal{L}_{\text{BPR}}(u) = \sum_{i \in \mathcal{P}_u} \sum_{j \in \mathcal{I} \setminus \mathcal{P}_u} \log \sigma(f(u,j) - f(u,i)) \tag{2.3}$$

  where σ denotes the activation function that approximates the Heaviside step function. The basic intuition behind BPR loss is to let the positive instances have higher scores than negative instances. In practice, there are various choices of the activation function. For instance, Rendle et al. [10] originally uses the sigmoid function, and the resultant BPR loss can approximate AUC metric.
- **Softmax Loss** (i.e., SL [11]) normalizes the predicted scores into a multinomial distribution [32] and optimizes the probability of positive instances over negative ones [33], which is defined as

$$\mathcal{L}_{\text{SL}}(u) = -\sum_{i \in \mathcal{P}_u} \log \left(\frac{\exp(f(u,i)/\tau)}{\sum_{j \in \mathcal{I}} \exp(f(u,j)/\tau)} \right) \tag{2.4}$$

  where τ is the temperature hyperparameter. SL can also be understood as a specific contrastive loss, which draws positive instances (u,i) closer and pushes negative instances (u,j) away [13].

3 Analyses on Softmax Loss from Pairwise Perspective

In this section, we aim to first represent the Softmax Loss (SL) in a pairwise form, followed by an analysis of its relationship with the DCG metric, where two limitations of SL are exposed.

Pairwise form of SL. To facilitate the analysis of SL and to build its relationship with the DCG metric, we rewrite SL (cf. Equation (2.4)) in the following pairwise form:

$$\mathcal{L}_{\text{SL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \exp(d_{uij}/\tau) \right), \quad \text{where } d_{uij} = f(u,j) - f(u,i) \tag{3.1}$$

Equation (3.1) indicates that SL is penalized based on the score gap between negative-positive pairs, i.e., $d_{uij} = f(u,j) - f(u,i)$. This concise expression is fundamental for ranking, as it optimizes the relative order of instances rather than their absolute values.

https://doi.org/10.52202/079017-3844

Connections between SL and DCG. We now analyze the connections between SL and the DCG metric (cf. Equations (2.1) and (3.1)), which could enhance our understanding of the advantages and disadvantages of SL. Our analysis follows previous work [11, 14], which begins by relaxing the negative logarithm of DCG with

$$-\log \mathrm{DCG}(u) + \log |\mathcal{P}_u| \leq -\log \left(\frac{1}{|\mathcal{P}_u|} \sum_{i \in \mathcal{P}_u} \frac{1}{\pi_u(i)} \right) \leq \frac{1}{|\mathcal{P}_u|} \sum_{i \in \mathcal{P}_u} \log \pi_u(i) \tag{3.2}$$

where the first inequality holds due to $\log_2(1 + \pi_u(i)) \leq \pi_u(i)$, and the second inequality holds due to Jensen's inequality [34]. Note that the ranking position $\pi_u(i)$ of item i can be expressed as

$$\pi_u(i) = \sum_{j \in \mathcal{I}} \mathbb{I}(f(u, j) \geq f(u, i)) = \sum_{j \in \mathcal{I}} \delta(d_{uij}) \tag{3.3}$$

where $\delta(\cdot)$ denotes the Heaviside step function, with $\delta(x) = 1$ for $x \geq 0$ and $\delta(x) = 0$ for $x < 0$. Since $\delta(d_{uij}) \leq \exp(d_{uij}/\tau)$ holds for all $\tau > 0$, we deduce that SL is a smooth upper bound of Equation (3.2), and thus serves as a reasonable surrogate loss for DCG and MRR metrics[1].

However, our analysis also reveals **two limitations of SL**:

- **Limitation 1: SL is not tight enough as a DCG surrogate loss.** There remains a significant gap between the Heaviside step function $\delta(\cdot)$ and the exponential function $\exp(\cdot)$, especially when d_{uij} reaches a relatively large value, where $\exp(\cdot)$ becomes substantially larger than $\delta(\cdot)$. This gap is further exacerbated by the temperature τ. Practically, we find that the optimal τ is usually chosen to be less than 0.2 (cf. Appendix B.5.2). Given the explosive nature of $\exp(\cdot)$, the gap becomes extremely large, potentially leading to suboptimal performance of SL in optimizing DCG.
- **Limitation 2: SL is highly sensitive to noise (e.g., false negative instances).** False negative instances [15] are common in the typical RS. This is often due to the exposure bias [16], where a user's lack of interaction with an item might stem from unawareness rather than disinterest. Unfortunately, SL is highly sensitive to these false negative instances. On one hand, these instances (u, j), which may exhibit patterns similar to true positive ones, are difficult for the model to differentiate and often receive larger predicted scores, thus bringing potentially larger d_{uij} for positive items i. As analyzed in Limitation 1, these instances can significantly enlarge the gap between SL and DCG due to the exponential function, causing the optimization to deviate from the DCG metric.

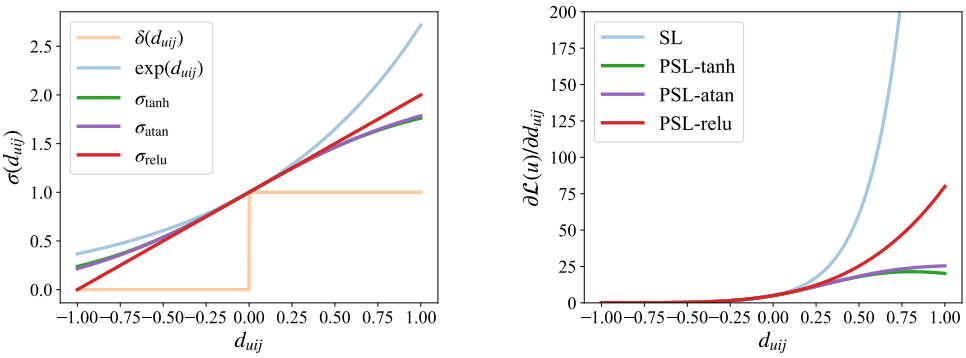
Gradient analysis of SL. Another perspective to support the view of Limitation 2 comes from the gradient analysis. Specifically, the gradient of SL w.r.t. d_{uij} is

$$\frac{\partial \mathcal{L}_{\mathrm{SL}}(u)}{\partial d_{uij}} = \frac{\exp(d_{uij}/\tau)/\tau}{|\mathcal{I}| \mathbb{E}_{j' \sim \mathcal{I}}[\exp(d_{uij'}/\tau)]} \quad \propto \exp(d_{uij}/\tau)/\tau \tag{3.4}$$

As can be seen, SL implicitly assigns a weight to the gradient of each negative-positive pair, where the weight is proportional to $\exp(d_{uij}/\tau)$. This suggests that instances with larger d_{uij} will receive larger weights. While this property may be desirable for hard mining [11], which can accelerate convergence, it also means that false negative instances, which typically have larger d_{uij}, will obtain disproportionately large weights, as shown in the weight distribution of SL in Figure 1b. Therefore, the optimization of SL can be easily dominated by false negative instances, leading to performance drops and training instability.

Discussions on DRO robustness and noise sensitivity. Recent work [15] claims that SL exhibits robustness to noisy data through Distributionally Robust Optimization (DRO) [19]. However, we argue that this is not the case. DRO indeed can enhance model robustness to distribution shifts, but it also increases the risk of noise sensitivity, as demonstrated by many studies on DRO [35, 36]. Intuitively, DRO emphasizes hard instances with larger losses, making noisy data contribute more rather than less to the optimization. This is also demonstrated from the experiments with false negative instances (cf. Figure 8 in [15]), where the improvements of SL over other baselines in Noise setting do not increase significantly but sometimes decay.

[1] Note that the middle term in Equation (3.2), i.e., $-\log\left(\frac{1}{|\mathcal{P}_u|}\sum_{i \in \mathcal{P}_u} 1/\pi_u(i)\right)$, is exactly $-\log \mathrm{MRR}(u)$. Therefore, SL serves as an upper bound of the negative logarithm of DCG and MRR, and minimizing SL leads to the improvement of these ranking metrics.



(a) Surrogate activations.

(b) Weight distributions.

Figure 1: (a) Illustration of different surrogate activations. (b) The weight distribution of SL as compared with PSL using three different surrogate activations. Here we set $\tau = 0.2$, which typically achieves optimal results in practice.

4 Methodology

4.1 Pairwise Softmax Loss

Recognizing the limitations of SL, particularly its reliance on the unsatisfactory exponential function, we propose to extend SL with a more general family of losses, termed **Pairwise Softmax Loss (PSL)**. In PSL, the exponential function $\exp(\cdot)$ is replaced by other *surrogate activations* $\sigma(\cdot)$ approximating the Heaviside step function $\delta(\cdot)$. For each user u, the PSL is defined as

$$\mathcal{L}_{\text{PSL}}(u) = \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \sigma(d_{uij})^{1/\tau} \right) \tag{4.1}$$

One might wonder why we apply the temperature outside the activation function (i.e., extending $\exp(d_{uij})^{1/\tau}$ to $\sigma(d_{uij})^{1/\tau}$)[2] rather than within it (i.e., extending $\exp(d_{uij}/\tau)$ to $\sigma(d_{uij}/\tau)$). This subtlety will be elucidated later as we demonstrate that the form in Equation (4.1) offers superior properties over the alternative.

Our PSL provides a flexible framework for selecting better activation functions, allowing the loss to exhibit improved properties compared to SL. We advocate for three activations, including **PSL-tanh**: $\sigma_{\text{tanh}} = \tanh(d_{uij})+1$, **PSL-atan**: $\sigma_{\text{atan}} = \arctan(d_{uij})+1$, and **PSL-relu**: $\sigma_{\text{relu}} = \text{ReLU}(d_{uij}+1)$. In the following, we will discuss the advantages of PSL and provide evidence for the selection of these surrogate activations.

Advantage 1: PSL is a better surrogate for ranking metrics. To highlight the advantages of replacing $\exp(\cdot)$ with alternative surrogate activations, we present the following lemma:

Lemma 4.1. *If the condition*

$$\delta(d_{uij}) \leq \sigma(d_{uij}) \leq \exp(d_{uij}) \tag{4.2}$$

is satisfied for any $d_{uij} \in [-1, 1]$, then PSL serves as a tighter DCG surrogate loss compared to SL.

The proof is presented in Appendix A.1. This lemma reveals that PSL could be a tighter surrogate loss for DCG compared to SL. Additionally, it provides guidance on the selection of a proper surrogate activation — we may choose the activation that lies between $\exp(\cdot)$ and $\delta(\cdot)$. As demonstrated in Figure 1a, our chosen surrogate activations σ_{tanh}, σ_{atan}, and σ_{relu} adhere to this principle.

Advantage 2: PSL controls the weight distribution. The gradient of PSL w.r.t. d_{uij} is

$$\frac{\partial \mathcal{L}_{\text{PSL}}(u)}{\partial d_{uij}} = \frac{\sigma'(d_{uij}) \cdot \sigma(d_{uij})^{1/\tau-1}/\tau}{|\mathcal{I}|\mathbb{E}_{j' \sim \mathcal{I}}[\sigma(d_{uij'})^{1/\tau}]} \quad \propto \sigma'(d_{uij}) \cdot \sigma(d_{uij})^{1/\tau-1}/\tau \tag{4.3}$$

[2]Note that the equation $\exp(d_{uij}/\tau) = \exp(d_{uij})^{1/\tau}$ holds.

https://doi.org/10.52202/079017-3844

This implies that the shape of the weight distribution is determined by the choice of surrogate activation. By selecting appropriate activations, PSL can better balance the contributions of instances during training. For example, the three activations advocated before can explicitly mitigate the explosive issue on larger d_{uij} (cf. Figure 1b), bringing better robustness to false negative instances.

One might argue that adjusting τ in SL could improve noise resistance. However, such adjustments do not alter the fundamental shape of the weight distribution, which remains exponential. Furthermore, as we discuss subsequently, τ plays a crucial role in controlling robustness against distribution shifts. Thus, indiscriminate adjustments to τ may compromise out-of-distribution (OOD) robustness.

Advantage 3: PSL is a DRO-empowered BPR loss. We establish a connection between PSL and BPR [10] based on Distributionally Robust Optimization (DRO) [19, 37]. Specifically, optimizing PSL is equivalent to applying a KL divergence DRO on negative item distribution over BPR loss (cf. Equation (2.3)), as demonstrated in the following theorem[3]:

Theorem 4.2. *For each user u and its positive item i, let $P = P(j|u,i)$ be the uniform distribution over $\mathcal{I}$. Given a robustness radius $\eta > 0$, consider the uncertainty set $\mathcal{Q}$ consisting of all perturbed distributions $Q = Q(j|u,i)$ satisfying: (i) Q is absolutely continuous w.r.t. P, i.e., $Q \ll P$; (ii) the KL divergence between Q and P is constrained by η, i.e., $D_{\mathrm{KL}}(Q\|P) \leq \eta$. Then, optimizing PSL is equivalent to performing DRO over BPR loss, i.e.,*

$$\min \underbrace{\left\{ \mathbb{E}_{i\sim\mathcal{P}_u} \left[\log \mathbb{E}_{j\sim\mathcal{I}} \left[e^{\log(\sigma(d_{uij}))/\tau} \right] \right] \right\}}_{\mathcal{L}_{\text{PSL}}(u)} \Leftrightarrow \min \underbrace{\left\{ \mathbb{E}_{i\sim\mathcal{P}_u} \left[\sup_{Q\in\mathcal{Q}} \mathbb{E}_{j\sim Q(j|u,i)} \left[\log \sigma(d_{uij}) \right] \right] \right\}}_{\mathcal{L}_{\text{BPR-DRO}}(u)} \tag{4.4}$$

where $\tau = \tau(\eta)$ is a temperature parameter controlled by η.

The proof is presented in Appendix A.2. Theorem 4.2 demonstrates how PSL, based on the DRO framework, is inherently robust to distribution shifts. This robustness is particularly valuable in RS, where user preference and item popularity may shift significantly. Therefore, PSL can be regarded as a robust generalization of BPR loss, offering better performance in OOD scenarios.

In addition, Theorem 4.2 also gives insights into the **rationality of PSL** that differs from serving as a DCG surrogate loss, but rather as a DRO-empowered BPR loss:

- **Rationality of surrogate activations:** The activation function in BPR is originally chosen as an approximation to the Heaviside step function [10]. Since PSL is a generalization of BPR as stated in Theorem 4.2, it is reasonable to select the activations in PSL that aligns with the ones in BPR. Interestingly, this principle coincides with our analysis from the perspective of DCG surrogate loss.
- **Rationality of the position of temperature:** Theorem 4.2 also rationalizes the extension form that places the temperature on the outside rather than inside. For the outside form (i.e., $\sigma(d_{uij})^{1/\tau}$), Theorem 4.2 holds, and the temperature τ can be interpreted as a Lagrange multiplier in DRO optimization, which controls the extent of distribution perturbation. However, for the inside form (i.e., $\sigma(d_{uij}/\tau)$), Theorem 4.2 no longer holds, and it would be challenging to establish the relationship between PSL and BPR.
- **Rationality of pairwise perspective:** Recent work such as BSL [15] also reveals the DRO property of SL (cf. Lemma 1 in [15]). However, we wish to highlight the distinctions between Theorem 4.2 and Wu et al. [15]'s analyses: 1) Wu et al. [15] views SL from a pointwise perspective and associates it with a specific, less commonly used pointwise loss. In contrast, our analyses adopt a pairwise perspective and establish a relationship between PSL and the widely used BPR loss. 2) We construct a link between two families of losses with flexible activation selections, and Wu et al. [15]'s analyses can be regarded as a special case within our broader framework.

The above analyses underscore the advantages of PSL and provide the principles to select surrogate activations. Remarkably, PSL is easily implemented and can be integrated into various recommendation scenarios. This can be achieved by merely replacing the exponential function $\exp(\cdot)$ in SL with another activation $\sigma(\cdot)$ surrogating the Heaviside step function, requiring minimal code modifications.

[3]Note that $e^{\log(\sigma(d_{uij}))/\tau} = \sigma(d_{uij})^{1/\tau}$ holds, thus the PSL in Equation (4.4) is identical to the one in Equation (4.1).

4.2 Discussions

Comparisons of two extension forms. In previous discussions, we highlight the advantages of the form that positions the temperature outside (i.e., $\sigma(d_{uij})^{1/\tau}$) over the inside (i.e., $\sigma(d_{uij}/\tau)$). As discussed in the analyses of Theorem 4.2, the outside form can be regarded as a DRO-empowered BPR, while the inside form cannot, which ensures the robustness of PSL against distribution shifts.

Here we provide an additional perspective on the advantages of the outside form. In fact, the outside form facilitates the selection of surrogate activations. For instance, to ensure that PSL serves as a tighter DCG surrogate loss compared to SL (i.e., ensure Lemma 4.1 holds), the outside form only need to consider the condition (4.2) on the range of $d_{uij} \in [-1, 1]$. However, for the inside form, this condition should be satisfied on the entire domain of the activation $\sigma(\cdot)$, which complicates the selection of activation functions. Therefore, the outside form is more flexible and easier to implement. We further provide empirical evidence in Appendix C.3, demonstrating that the inside form will lose the advantages of achieving tighter DCG surrogate loss, leading to compromised performance.

Connections with other losses. We further discuss the connections between PSL and other losses:

- **Connection with AdvInfoNCE [38]:** According to Theorem 3.1 in Zhang et al. [38], AdvInfoNCE can indeed be considered as a special case of PSL with $\sigma(\cdot) = \exp(\exp(\cdot))$. We argue that this activation is not a good choice as it would enlarge the gap between the loss and DCG. In fact, we have $-\log \mathrm{DCG} \leq \mathcal{L}_{\mathrm{PSL}} \leq \mathcal{L}_{\mathrm{SL}} \leq \mathcal{L}_{\mathrm{AdvInfoNCE}}$ (cf. Appendix A.3 for proof). While AdvInfoNCE may achieve good performance in some specific OOD scenarios as tested in Zhang et al. [38], we argue that AdvInfoNCE is a looser DCG surrogate loss and would be highly sensitive to noise (cf. Table 1 and Figure 2 in Section 5.2 for empirical validation).
- **Connection with BPR [10]:** Besides the DRO relation stated in Theorem 4.2, we also derive the bound relation between BPR and PSL with the same activation, i.e., $-\log \mathrm{DCG} \leq \mathcal{L}_{\mathrm{PSL}} \leq \log \mathcal{L}_{\mathrm{BPR}}$ (cf. Appendix A.3 for proof). This relation clearly demonstrates the effectiveness of PSL over BPR — performing DRO over BPR results robustness to distribution shifts, while also achieving a tighter surrogate of DCG, which is interesting (cf. Tables 1 and 2 in Section 5.2 for empirical validation). An intuitive explanation is that DCG focuses more on the higher-ranked items. Given that DRO would give more weight to the hard negative instances with larger prediction scores and higher positions, it would naturally narrow the gap between BPR and DCG.

5 Experiments

5.1 Experimental Setup

Testing scenarios. We adopt three representative testing scenarios to comprehensively evaluate model accuracy and robustness, including: 1) **IID setting:** the conventional testing scenario where training and test data are randomly split and identically distributed; 2) **OOD setting:** to assess the model's robustness on the out-of-distribution (OOD) data, we adopt a debiasing testing paradigm where the item popularity distribution shifts. We closely refer to Zhang et al. [20], Wang et al. [24], and Wei et al. [39], sampling a test set where items are uniformly distributed while maintaining the long-tail nature of the training dataset; 3) **Noise setting:** to evaluate the model's sensitivity to noise, following Wu et al. [15], we manually impute a certain proportion of false negative items in the training data. The details of the above testing scenarios are provided in Appendix B.1.

Datasets. Four widely-used datasets including Amazon-Book, Amazon-Electronic, Amazon-Movie [40, 41], and Gowalla [42] are used in our experiments. Considering the item popularity is not heavily skewed in the Amazon-Book and Amazon-Movie datasets, we turn to other conventional datasets, Amazon-CD [40, 41] and Yelp2018 [43], as replacements for OOD testing. All datasets are split into 80% training set and 20% test set, with 10% of the training set further treated as the validation set. The details of the above datasets are summarized in Appendix B.1.

Metrics. We closely refer to Wu et al. [15] and Zhang et al. [38], adopting Top-K metrics including $\mathrm{NDCG@}K$ [6] and $\mathrm{Recall@}K$ [29] for performance evaluation, where NDCG is the normalized DCG, i.e., dividing DCG by the ideal value. Here we simply set $K = 20$ as in recent work [15, 38] while observing similar results with other choices. For more details, please refer to Appendix B.2.

https://doi.org/10.52202/079017-3844

Table 1: Performance comparison in terms of Recall@20 and NDCG@20 under the IID setting. The best result is bolded, and the blue-colored zone indicates that PSL is better than SL. Imp.% denotes the NDCG@20 improvement of PSL over SL. The marker "*" indicates that the improvement is statistically significant (p-value < 0.05).

Model	Loss	Amazon-Book		Amazon-Electronic		Amazon-Movie		Gowalla	
		Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
MF [26]	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%***	
LightGCN [22]	BPR [10]	0.0984	0.0678	0.0813	0.0524	0.1006	0.0681	0.1745	0.1402
	LLPAUC [44]	0.1147	0.0810	**0.0831**	0.0507	0.1272	0.0886	0.1616	0.1192
	SL [11]	0.1567	0.1220	0.0823	0.0526	0.1304	0.0941	0.2068	0.1628
	AdvInfoNCE [38]	0.1568	0.1177	0.0823	0.0528	0.1292	0.0936	0.2066	0.1625
	BSL [15]	0.1568	0.1220	0.0823	0.0526	**0.1306**	0.0943	0.2069	0.1628
	PSL-tanh	**0.1575**	**0.1233**	0.0825	0.0532	0.1300	0.0947	**0.2091**	**0.1648**
	PSL-atan	**0.1575**	**0.1233**	0.0825	0.0532	0.1300	0.0948	**0.2091**	**0.1648**
	PSL-relu	**0.1575**	**0.1233**	0.0830	**0.0536**	0.1300	**0.0953**	0.2086	**0.1648**
	Imp.%	**+1.12%***		**+1.98%***		**+1.22%***		**+1.24%***	
XSimGCL [45]	BPR [10]	0.1269	0.0905	0.0777	**0.0508**	0.1236	0.0857	0.1966	0.1570
	LLPAUC [44]	0.1363	0.1008	0.0781	0.0481	0.1184	0.0828	0.1632	0.1200
	SL [11]	0.1549	0.1207	0.0772	0.0490	0.1255	0.0905	0.2005	0.1570
	AdvInfoNCE [38]	0.1568	0.1179	0.0776	0.0489	0.1252	0.0906	0.2010	0.1564
	BSL [15]	0.1550	0.1207	0.0800	0.0507	0.1267	0.0918	**0.2037**	**0.1597**
	PSL-tanh	0.1567	0.1225	0.0790	0.0501	0.1308	0.0926	0.2034	0.1591
	PSL-atan	0.1565	0.1225	0.0792	0.0502	0.1253	0.0917	0.2035	0.1591
	PSL-relu	**0.1571**	**0.1228**	**0.0801**	0.0507	**0.1313**	**0.0935**	**0.2037**	0.1593
	Imp.%	**+1.72%***		**+3.39%***		**+3.42%***		**+1.48%***	

Compared methods. Five representative loss functions are compared in our experiments, including 1) the representative pairwise loss **BPR** (UAI'09 [10]); 2) the SOTA recommendation loss **Softmax Loss (SL)** (TOIS'24 [11]) and its two DRO-enhancements **AdvInfoNCE** (NIPS'23 [38]) and **BSL** (ICDE'24 [15]); 3) another SOTA loss **LLPAUC** (WWW'24 [44]) that optimizes the Lower-Left Partial AUC. Refer to Appendix B.3 for more details about these baselines.

Backbones. We also adopt three representative backbone models to evaluate the effectiveness of loss, including MF [26], LightGCN [22], and XSimGCL [45], see Appendix B.4 for more details.

Hyperparameter settings. A grid search is utilized to find the optimal hyperparameters. For all compared methods, we closely refer to the configurations provided in their respective publications to ensure their optimal performance. As we also carefully finetune SL, the improvements of existing methods over it are not as significant as those presented in their papers. The hyperparameter settings are provided in Appendix B.5, where the detailed optimal hyperparameters for each method on each dataset and backbone are reported.

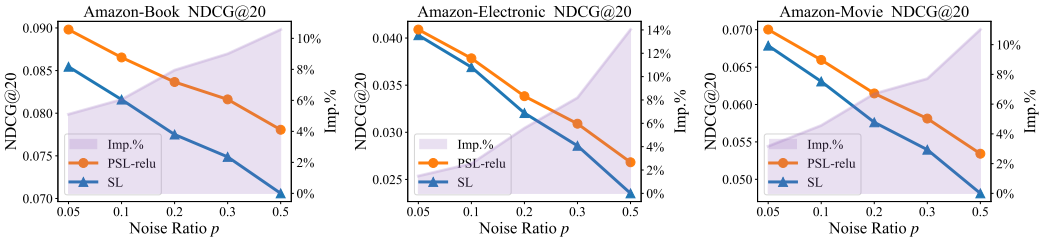
5.2 Performance Comparisons

Results under IID setting. Table 1 presents the performance of our PSL compared with baselines.

- **PSL outperforms SL and other baselines.** Experimental results demonstrate that PSL, with three carefully selected surrogate activations, consistently outperforms SL across all datasets and backbones, with only a few exceptions. For instance, on the MF backbone, compared to the marginal improvements or sometimes even degradation of AdvInfoNCE (-3%~0.5%) and BSL (0.0%~0.5%), PSL shows a significant enhancement over SL (1%~3%). Moreover, our PSL surpasses all compared baselines in most cases, clearly demonstrating its effectiveness.
- **PSL achieves tighter connections with ranking metrics.** We observe that the results align well with our theoretical analyses of PSL's Advantage 1 in Section 4. By replacing the exponential function with other suitable surrogate activations, PSL establishes a tighter relationship with ranking

Table 2: Performance comparison in terms of Recall@20 and NDCG@20 under the OOD setting with popularity shift (on MF backbone). The best result is bolded, and the blue-colored zone indicates that PSL is better than SL. Imp.% denotes the NDCG@20 improvement of PSL over SL. The marker "*" indicates that the improvement is statistically significant (p-value < 0.05).

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%***	



(a) Amazon-Book (b) Amazon-Electronic (c) Amazon-Movie

Figure 2: Performance comparison of SL and PSL in terms of NDCG@20 with different false negative noise ratio (on MF backbone). We also present the relative improvements (i.e., Imp.%) achieved by PSL over SL. The complete results of other baselines are provided in Appendix C.1.

metrics, thus achieving better NDCG performance (cf. Lemma 4.1). This is also empirically evident from the larger improvements in NDCG compared to Recall. In contrast, as discussed in Section 4.2, other baselines like AdvInfoNCE and BSL either widen the gap or fail to connect with the ranking metrics, resulting in slight improvements or even performance drops.

Results under OOD setting. Table 2 presents the results in OOD scenarios with popularity shift. Given the consistent behavior across the three backbones, here we only report the results on MF.

- **PSL is robust to distribution shifts.** Experimental results indicate that PSL has a strong robustness against distribution shifts, which is consistent with PSL's Advantage 3 in Section 4. As can be seen, PSL not only outperforms all baselines (2%~5%), but also achieves more pronounced improvements than in IID setting, like on Amazon-Electronic (2.31% $\rightarrow$ 5.02%) and Gowalla (1.42% $\rightarrow$ 2.02%). This demonstrates the superior robustness of PSL to distribution shifts, as shown in Theorem 4.2.
- **PSL is a DRO-enhancement of more reasonable loss.** Although both PSL and SL can be considered as DRO-enhanced losses (cf. Theorem 4.2), the original loss of our three PSLs before DRO-enhancement is more reasonable than that of SL, which degenerates from BPR loss to a linear triplet loss [46]. Therefore, we observe significant improvements of PSL over SL.

Results under Noise setting. Figure 2 and Appendix C.1 presents the results with a certain ratio of imputed false negative noise. Specifically, we regard 10% of the positive items in the training set as false negative noise and allow the negative sampling procedure to have a certain probability p of sampling those items. We test the model performance with varying noise ratios $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$.

- **PSL has strong noise resistance.** Experimental results demonstrate that as the noise ratio p increases, both the performance of SL and PSL decline. The performance decline rate of PSL is significantly smaller than that of other baselines, resulting in higher performance enhancement($>$ 10% when $p = 0.5$). These results indicate that PSL possesses stronger noise resistance than SL, which stems from our rational activation design, as discussed in PSL's Advantage 2 in Section 4.

https://doi.org/10.52202/079017-3844

However, for DRO-enhanced losses such as AdvInfoNCE, the performance declines similarly to or even more quickly than SL (cf. Appendix C.1), which coincides with our theoretical analyses.

6 Related Work

Model-related recommendation research. Recent years have witnessed flourishing publications on collaborative filtering (CF) models. The earliest works are mainly extensions of Matrix Factorization [26], building more complex interactions between embeddings [47], such as MF [26], LRML [48], SVD [49, 50], SVD++ [51], NCF [8], etc. In recent years, given the effectiveness of Graph Neural Networks (GNNs) [52–58] in capturing high-order relations, which align well with CF assumptions, GNN-based models have emerged and achieved great success, such as LightGCN [22], NGCF [55], LCF [59], APDA [60], etc. Building upon LightGCN, some works attempt to introduce contrastive learning [12, 61] for graph data augmentation, such as SGL [62] and XSimGCL [45], achieving SOTA performance in recommendation.

Loss-related recommendation research. Existing recommendation losses can be primarily categorized into pointwise loss [8, 9], pairwise loss [10], and Softmax Loss (SL) [11], as discussed in Section 2. Given the effectiveness of SL, recently some researchers have proposed to enhance SL from different perspectives. For instance, BSL [15] aims to enhance the positive distribution robustness by leveraging Distributionally Robust Optimization (DRO); AdvInfoNCE [38] employs adversarial learning to enhance SL's robustness; Zhang et al. [20] suggests incorporating bias-aware margins in SL to tackle popularity bias. Beyond these three types of losses, other approaches have also been explored in recent years. For example, Zhao et al. [63] introduces auto-loss, which utilizes automated machine learning techniques to search the optimal loss; Shi et al. [44] proposes LLPAUC to approximate Recall@K metric. The main concerns with these losses are their lack of theoretical connections to ranking metrics like DCG, which may result in them not consistently outperforming the basic SL. Moreover, both auto-loss and LLPAUC require iterative learning, leading to additional computational time and increased instability.

7 Conclusion and Limitations

In this work, we introduce a new family of loss functions, termed Pairwise Softmax Loss (PSL). PSL theoretically offers three advantages: 1) it serves as a better surrogate for ranking metrics with appropriate surrogate activations; 2) it allows flexible control over the distribution of the data contribution; 3) it can be interpreted as a specific BPR loss enhanced by Distributionally Robust Optimization (DRO). These properties demonstrate that PSL has greater effectiveness and robustness compared to Softmax Loss. Our extensive experiments across three testing scenarios validate the superiority of PSL over existing methods.

One limitation of both PSL and SL is inefficiency, as they require sampling a relatively large number of negative instances per iteration. How to address this issue and improve the efficiency of these losses is an interesting direction for future research.

Acknowledgments and Disclosure of Funding

This work is supported by the Zhejiang Province "JianBingLingYan+X" Research and Development Plan (2024C01114).

References

[1] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronics*, 11(1):141, 2022.

[2] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

[3] Feiran Huang, Zefan Wang, Xiao Huang, Yufeng Qian, Zhetao Li, and Hao Chen. Aligning distillation for cold-start item recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1147–1157, 2023.

[4] Feiran Huang, Zhenghang Yang, Junyi Jiang, Yuanchen Bei, Yijie Zhang, and Hao Chen. Large language model interaction simulator for cold-start item recommendation. *arXiv preprint arXiv:2402.09176*, 2024.

[5] Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

[6] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA, 2017.

[7] Xiangkui Lu, Jun Wu, and Jianbo Yuan. Optimizing reciprocal rank with bayesian average for improved next item recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2236–2240, 2023.

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182, 2017.

[9] Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 355–364, 2017.

[10] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461, 2009.

[11] Jiancan Wu, Xiang Wang, Xingyu Gao, Jiawei Chen, Hongcheng Fu, and Tianyu Qiu. On the effectiveness of sampled softmax loss for item recommendation. *ACM Transactions on Information Systems*, 42(4): 1–26, 2024.

[12] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[13] Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Sebastian Bruch, Xuanhui Wang, Michael Bendersky, and Marc Najork. An analysis of the softmax cross entropy loss for learning-to-rank with binary relevance. In *Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval*, pages 75–78, 2019.

[15] Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Jizhi Zhang, and Xiang Wang. Bsl: Understanding and improving softmax loss for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 816–830. IEEE, 2024.

[16] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3): 1–39, 2023.

[17] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. Autodebias: Learning to debias for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–30, 2021.

[18] Bohao Wang, Feng Liu, Jiawei Chen, Yudi Wu, Xingyu Lou, Jun Wang, Yan Feng, Chun Chen, and Can Wang. Llm4dsr: Leveraing large language model for denoising sequential recommendation. *arXiv preprint arXiv:2408.08208*, 2024.

[19] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4): 2258–2275, 2017.

[20] An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat-Seng Chua. Invariant collaborative filtering to popularity distribution shift. In *Proceedings of the ACM Web Conference 2023*, pages 1240–1251, 2023.

[21] Zihao Zhao, Jiawei Chen, Sheng Zhou, Xiangnan He, Xuezhi Cao, Fuzheng Zhang, and Wei Wu. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):9920–9931, 2022.

[22] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648, 2020.

[23] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[24] Bohao Wang, Jiawei Chen, Changdong Li, Sheng Zhou, Qihao Shi, Yang Gao, Yan Feng, Chun Chen, and Can Wang. Distributionally robust graph-based recommendation system. *arXiv preprint arXiv:2402.12994*, 2024.

 https://doi.org/10.52202/079017-3844

[25] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[26] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[27] Jiawei Chen, Junkang Wu, Jiancan Wu, Xuezhi Cao, Sheng Zhou, and Xiangnan He. Adap-τ: Adaptively modulating embedding magnitude for recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 1085–1096, 2023.

[28] Andreas Argyriou, Miguel González-Fierro, and Le Zhang. Microsoft recommenders: best practices for production-ready recommendation systems. In *Companion Proceedings of the Web Conference 2020*, pages 50–51, 2020.

[29] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences*, 10(21):7748, 2020.

[30] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813–831, 2019.

[31] Ahmed Rashed, Josif Grabocka, and Lars Schmidt-Thieme. A guided learning approach for item recommendation via surrogate loss learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 605–613, 2021.

[32] George Casella and Roger Berger. *Statistical inference*. CRC Press, 2024.

[33] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136, 2007.

[34] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[35] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355. PMLR, 2021.

[36] Sloan Nietert, Ziv Goldfeld, and Soroosh Shafiee. Outlier-robust wasserstein dro. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

[38] An Zhang, Leheng Sheng, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. Empowering collaborative filtering with principled adversarial contrastive loss. *Advances in Neural Information Processing Systems*, 36, 2024.

[39] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1791–1800, 2021.

[40] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

[41] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[42] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090, 2011.

[43] Yelp. Yelp dataset. `https://www.yelp.com/dataset`, 2018.

[44] Wentao Shi, Chenxu Wang, Fuli Feng, Yang Zhang, Wenjie Wang, Junkang Wu, and Xiangnan He. Lower-left partial auc: An effective and efficient optimization metric for recommendation. *arXiv preprint arXiv:2403.00844*, 2024.

[45] Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. Xsimgcl: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[46] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

[47] Emile Fiesler and Russell Beale. *Handbook of neural computation*. CRC Press, 2020.

[48] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Latent relational metric learning via memory-based attention for collaborative ranking. In *Proceedings of the 2018 world wide web conference*, pages 729–739, 2018.

[49] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6): 391–407, 1990.

[50] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 95–104, 2007.

[51] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008.

[52] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

[53] Chen Gao, Xiang Wang, Xiangnan He, and Yong Li. Graph neural networks for recommender system. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1623–1625, 2022.

[54] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[55] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[56] Hande Dong, Jiawei Chen, Fuli Feng, Xiangnan He, Shuxian Bi, Zhaolin Ding, and Peng Cui. On the equivalence of decoupled graph convolution network and label propagation. In *Proceedings of the Web Conference 2021*, pages 3651–3662, 2021.

[57] Jiancan Wu, Xiangnan He, Xiang Wang, Qifan Wang, Weijian Chen, Jianxun Lian, and Xing Xie. Graph convolution machine for context-aware recommender system. *Frontiers of Computer Science*, 16(6): 166614, 2022.

[58] Hao Chen, Yuanchen Bei, Qijie Shen, Yue Xu, Sheng Zhou, Wenbing Huang, Feiran Huang, Senzhang Wang, and Xiao Huang. Macro graph neural networks for online billion-scale recommender systems. In *Proceedings of the ACM on Web Conference 2024*, pages 3598–3608, 2024.

[59] Wenhui Yu and Zheng Qin. Graph convolutional network for recommendation with low-pass collaborative filters. In *International Conference on Machine Learning*, pages 10936–10945. PMLR, 2020.

[60] Huachi Zhou, Hao Chen, Junnan Dong, Daochen Zha, Chuang Zhou, and Xiao Huang. Adaptive popularity debiasing aggregator for graph collaborative filtering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 7–17, 2023.

[61] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[62] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[63] Xiangyu Zhao, Haochen Liu, Wenqi Fan, Hui Liu, Jiliang Tang, and Chong Wang. Autoloss: Automated loss function search in recommendations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3959–3967, 2021.

[64] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[65] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[66] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[67] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[68] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[69] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

 https://doi.org/10.52202/079017-3844

A Theoretical Proofs

A.1 Proof of Lemma 4.1

Lemma A.1 (Lemma 4.1)**.** *If the condition*

$$\delta(d_{uij}) \leq \sigma(d_{uij}) \leq \exp(d_{uij}) \tag{4.2}$$

is satisfied for any $d_{uij} \in [-1, 1]$, then PSL serves as a tighter DCG surrogate loss compared to SL.

Proof of Lemma 4.1. For any $\tau > 0$, Equation (4.2) indicates that

$$\delta(d_{uij}) \leq \sigma(d_{uij})^{1/\tau} \leq \exp(d_{uij})^{1/\tau} \tag{A.1}$$

which means $\sigma(\cdot)^{1/\tau}$ is tighter than $\exp(\cdot)^{1/\tau}$ approximating $\delta(\cdot)$. According to Equations (3.2) and (3.3) in Section 3, we conclude that PSL is a tighter surrogate loss for DCG compared to SL. □

A.2 Proof of Theorem 4.2

Theorem A.2 (Theorem 4.2)**.** *For each user u and its positive item i, let $P = P(j|u, i)$ be the uniform distribution over $\mathcal{I}$. Given a robustness radius $\eta > 0$, consider the uncertainty set $\mathcal{Q}$ consisting of all perturbed distributions $Q = Q(j|u, i)$ satisfying: (i) Q is absolutely continuous w.r.t. P, i.e., $Q \ll P$; (ii) the KL divergence between Q and P is constrained by η, i.e., $D_{\text{KL}}(Q\|P) \leq \eta$. Then, optimizing PSL is equivalent to performing DRO over BPR loss, i.e.,*

$$\min \underbrace{\left\{\mathbb{E}_{i\sim\mathcal{P}_u}\left[\log \mathbb{E}_{j\sim\mathcal{I}}\left[e^{\log(\sigma(d_{uij}))/\tau}\right]\right]\right\}}_{\mathcal{L}_{\text{PSL}}(u)} \Leftrightarrow \min \underbrace{\left\{\mathbb{E}_{i\sim\mathcal{P}_u}\left[\sup_{Q\in\mathcal{Q}} \mathbb{E}_{j\sim Q(j|u,i)}\left[\log \sigma(d_{uij})\right]\right]\right\}}_{\mathcal{L}_{\text{BPR-DRO}}(u)} \tag{4.4}$$

where $\tau = \tau(\eta)$ is a temperature parameter controlled by η.

To prove Theorem 4.2, it suffices to prove the following lemma:

Lemma A.3 (DRO under KL divergence)**.** *Given the loss term $\ell(x; \theta)$ of input x and parameters θ, for any robustness radius $\eta > 0$, DRO under KL divergence is equivalent to optimizing a loss in the form of $\log \mathbb{E}[\exp(\cdot)]$, i.e.,*

$$\min_{\theta} \sup_{Q\in\mathcal{Q}} \mathbb{E}_{x\sim Q}[\ell(x; \theta)] \Leftrightarrow \min_{\theta,\tau>0} \{\tau \log \mathbb{E}_{x\sim P}[\exp(\ell(x; \theta)/\tau)] + \tau\eta\} \tag{A.2}$$

where the uncertainty set $\mathcal{Q}$ consists of all perturbed distributions Q constrained by KL divergence w.r.t. the original distribution P, i.e., $\mathcal{Q} = \{Q \ll P : D_{\text{KL}}(Q\|P) \leq \eta\}$.

Lemma A.3, which was first proposed by Hu and Hong [37] with a complex proof, gives a closed-form solution for DRO under KL divergence. Here we provide an elegant proof based on the following general result about the ϕ-divergence DRO, which was first proposed by Shapiro [19].

Theorem A.4 (DRO under ϕ-divergence, [19])**.** *Consider the DRO problem in ϕ-divergence*

$$D_\phi(Q\|P) = \int \phi\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right) \mathrm{d}P \tag{A.3}$$

where $\phi : \mathbb{R} \to \overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\}$ is a convex function such that $\phi(1) = 0$ and $\phi(t) = +\infty$ for any $t < 0$. Then the inner maximization problem in DRO, i.e., $\sup_{Q\in\mathcal{Q}} \mathbb{E}_{x\sim Q}[\ell(x; \theta)]$ with the uncertainty set $\mathcal{Q} = \{Q \ll P : D_\phi(Q\|P) \leq \eta\}$, is equivalent to the following optimization problem:

$$\inf_{\tau>0,\mu} \{\mathbb{E}_{x\sim P}\left[(\tau\phi)^*(\ell(x; \theta) - \mu)\right] + \tau\eta + \mu\} \tag{A.4}$$

where $f^*(y) = \sup_x \{yx - f(x)\}$ is the Fenchel conjugate [64] for any convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$.

Proof of Theorem A.4. Let the likelihood ratio $L(x) = \mathrm{d}Q(x)/\mathrm{d}P(x)$, then the inner maximization problem in DRO can be reformulated as

$$\sup_{L\geq 0} \{\mathbb{E}_{x\sim P}\left[L(x)\ell(x; \theta)\right] \mid \mathbb{E}_{x\sim P}[\phi(L(x))] \leq \eta, \mathbb{E}_{x\sim P}[L(x)] = 1\} \tag{A.5}$$

The Lagrangian of Equation (A.5) is

$$\mathcal{L}(L,\tau,\mu)=\mathbb{E}_{x\sim P}\left[L(x)\ell(x;\theta)-\tau\phi(L(x))-\mu L(x)\right]+\tau\eta+\mu \tag{A.6}$$

where $\tau\geq 0$ and μ are the Lagrange multipliers. Problem (A.5) is a convex optimization problem. One can easily check the Slater's condition [65] by choosing $L(x)\equiv 1$, thus the strong duality [65] holds, and problem (A.5) is equivalent to the dual problem (A.7) of the Lagrangian (A.6):

$$\inf_{\tau\geq 0,\mu}\sup_{L\succeq 0}\mathcal{L}(L,\tau,\mu) \tag{A.7}$$

Consider the inner maximization problem $\sup_{L\succeq 0}\mathcal{L}(L,\tau,\mu)$ in Equation (A.7), $\tau\eta+\mu$ is a constant and can be ignored. By the theorem of interchange of minimization and integration [64], we can interchange $\sup$ and expectation in Equation (A.7).Then $\sup_{L\succeq 0}\mathcal{L}(L,\tau,\mu)$ can be reformulated as

$$\mathbb{E}_{x\sim P}\left[\sup_{L\succeq 0}\left\{L(x)(\ell(x;\theta)-\mu)-\tau\phi(L(x))\right\}\right] \tag{A.8}$$

The above problem can be rewritten by the Fenchel conjugate as

$$\mathbb{E}_{x\sim P}\left[(\tau\phi)^*(\ell(x;\theta)-\mu)\right] \tag{A.9}$$

Thus, problem (A.7) is equivalent to

$$\inf_{\tau\geq 0,\mu}\left\{\mathbb{E}_{x\sim P}\left[(\tau\phi)^*(\ell(x;\theta)-\mu)\right]+\tau\eta+\mu\right\} \tag{A.10}$$

Finally, note that the condition $\tau\geq 0$ in problem (A.10) can be relaxed to $\tau>0$ without affecting the optimal value, thus problem (A.10) is equivalent to problem (A.4), which completes the proof. □

Lemma A.3 can be directly derived from Theorem A.4 as follows:

Proof of Lemma A.3. KL divergence is a special case of ϕ-divergence with $\phi(x)=x\log x$, and the Fenchel conjugate of $\tau\phi$ is

$$(\tau\phi)^*(y)=\sup_x\left\{yx-\tau x\log x\right\}=\tau e^{y/\tau-1} \tag{A.11}$$

By Theorem A.4, the DRO problem under KL divergence is equivalent to

$$\begin{aligned}&\inf_{\tau>0,\mu}\left\{\mathbb{E}_{x\sim P}\left[\tau e^{(\ell(x;\theta)-\mu)/\tau-1}\right]+\tau\eta+\mu\right\}\\ =&\inf_{\tau>0,\mu}\left\{\mathbb{E}_{x\sim P}\left[e^{\ell(x;\theta)/\tau}\right]\tau e^{-\mu/\tau-1}+\tau\eta+\mu\right\}\end{aligned} \tag{A.12}$$

We fix τ and solve the optimal value of μ as

$$\mu^*=\tau\log\mathbb{E}_{x\sim P}\left[e^{\ell(x;\theta)/\tau}\right]-\tau \tag{A.13}$$

Therefore, by substituting the optimal μ^* in Equation (A.13) back to Equation (A.12), the original DRO problem is equivalent to

$$\inf_{\theta,\tau>0}\left\{\tau\log\mathbb{E}_{x\sim P}\left[e^{\ell(x;\theta)/\tau}\right]+\tau\eta\right\} \tag{A.14}$$

This completes the proof. □

Theorem 4.2 is a direct consequence of Lemma A.3, when setting the loss term $\ell(x;\theta)$ as $\log\sigma(d_{uij})$ (i.e., the pairwise loss term in BPR loss), P as the uniform distribution over $\mathcal{I}$, Q as the perturbed distribution constrained by KL divergence w.r.t. P, and $\tau=\tau(\eta)$ as the optimal value of Lagrange multiplier τ in Equation (A.2). This completes the proof of Theorem 4.2.

https://doi.org/10.52202/079017-3844

A.3 Proof of the Bound Connections between PSL and Other Losses in Section 4.2

Proof of the Bound Connections in Section 4.2. We have proved in Lemma 4.1 that

$$-\log \mathrm{DCG}(u) + \log |\mathcal{P}_u| \leq \frac{1}{|\mathcal{P}_u|} \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \sigma(d_{uij})^{1/\tau} \right) \tag{A.15}$$

with any surrogate activation σ satisfying $\delta(d_{uij}) \leq \sigma(d_{uij})$. Furthermore, if two surrogate activations σ_1, σ_2 satisfy $\sigma_1(d_{uij}) \leq \sigma_2(d_{uij})$ for any $d_{uij} \in [-1, 1]$, then the corresponding DCG surrogate losses satisfy the same inequality. Therefore, we have

$$-\log \mathrm{DCG} \leq \mathcal{L}_{\mathrm{PSL}} \leq \mathcal{L}_{\mathrm{SL}} \leq \mathcal{L}_{\mathrm{AdvInfoNCE}} \tag{A.16}$$

where the constant term is omitted for simplicity.

Finally, we prove that BPR serves as a surrogate loss for DCG. Apply Jensen's inequality to the RHS of Equation (A.15), we have

$$\frac{1}{|\mathcal{P}_u|} \sum_{i \in \mathcal{P}_u} \log \left(\sum_{j \in \mathcal{I}} \sigma(d_{uij})^{1/\tau} \right) \leq \log \left(\frac{1}{|\mathcal{P}_u|} \sum_{i \in \mathcal{P}_u} \sum_{j \in \mathcal{I}} \sigma(d_{uij})^{1/\tau} \right) \tag{A.17}$$

The RHS of Equation (A.17) is just $\log \mathcal{L}_{\mathrm{BPR}}(u) - \log |\mathcal{P}_u|$ with the same surrogate activation σ in BPR. Equation (A.17) indicates that for any surrogate activation σ, the general PSL (including SL, BSL, and AdvInfoNCE) is always better than BPR with the same σ, i.e.,

$$-\log \mathrm{DCG} \leq \mathcal{L}_{\mathrm{PSL}} \leq \log \mathcal{L}_{\mathrm{BPR}} \tag{A.18}$$

where the constant term is omitted for simplicity. This completes the proof. □

B Experimental Details

B.1 Datasets

The six benchmark datasets used in our experiments are summarized in Table B.1. In dataset preprocessing, following the standard practice in Wang et al. [55], we use 10-core setting [66], i.e., all users and items have at least 10 interactions. We also remove the low-quality interactions, such as those with ratings (if available) lower than 3. After preprocessing, we split the datasets into 80% training and 20% test sets. In IID and Noise settings, we further randomly split a 10% validation set from training set for hyperparameter tuning.

Table B.1: Statistics of datasets. All datasets are cleaned by 10-core setting. If the dataset is used in both IID and OOD settings, the statistics below are provided for the IID setting.

Dataset	#Users	#Items	#Interactions	Density
Amazon-Electronic [40, 41]	13,455	8,360	234,521	0.00208
Amazon-CD [40, 41]	12,784	13,874	360,763	0.00203
Amazon-Movie [40, 41]	26,968	18,563	762,957	0.00152
Gowalla [42]	29,858	40,988	1,027,464	0.00084
Yelp2018 [43]	55,616	34,945	1,506,777	0.00078
Amazon-Book [40, 41]	135,109	115,172	4,042,382	0.00026

The details of datasets are as follows:

- **Amazon** [40, 41]: The Amazon dataset is a large crawl of product reviews from Amazon[4]. The 2014 version of Amazon dataset contains 142.8 million reviews spanning May 1996 - July 2014. We process four widely-used categories: Electronic, CD, Movie, and Book, with interactions ranging from 200K to 4M.
- **Gowalla** [42]: The Gowalla dataset is a check-in dataset collected from the location-based social network Gowalla[5], including 1M users, 1M locations, and 6M check-ins.
- **Yelp2018** [43]: The Yelp[6] dataset is a subset of Yelp's businesses, reviews, and user data, which was originally used in the Yelp Dataset Challenge. The 2018 version of Yelp dataset contains 5M reviews.

The detailed dataset constructions in IID, OOD and Noise settings are as follows:

- **IID setting** [22]: In the IID setting, the test set is randomly split from the original dataset. Specifically, the positive items of each user are split into 80% training and 20% test sets. Moreover, the training set is further split into 90% training and 10% validation sets for hyperparameter tuning. In the IID setting, the training and test sets are both long-tail.
- **OOD setting** [20, 24, 39]: In the OOD setting, a 20% test set is uniformly sampled (w.r.t. items) from the original dataset, while the 80% training set remains long-tail. The OOD setting is used to simulate real-world online recommender systems. In order to avoid leaking information about the test set distribution, we do not introduce the validation set.
- **Noise setting** [15]: In the Noise setting, the validation and test sets are split in the same way as the IID setting. However, we randomly sample 10% of the training set as the *false negatives*. In Noise training, the negative items will be sampled from the false negatives with a probability of p as the *negative noise*, where $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ is a.k.a. the *noise ratio*.

All experiments are conducted on one NVIDIA GeForce RTX 4090 GPU and one AMD EPYC 7763 64-Core Processor.

B.2 Metrics

This section provides a detailed explanation of the recommendation metrics used or mentioned in our experiments.

[4] https://www.amazon.com/
[5] https://en.wikipedia.org/wiki/Gowalla
[6] https://www.yelp.com/

https://doi.org/10.52202/079017-3844

As stated in Section 5.1, we use Top-K recommendation [5]. It should be noted that for each user, the positive items in the training set will be masked and not included in the Top-K recommendations when evaluating, and the ground-truth positive items $\mathcal{P}_u$ only consist of those in the test set. For convenience, we denote the set of hit items in the Top-K recommendations for user u as $\mathcal{H}_u = \{i \in \mathcal{P}_u : \pi_u(i) \leq K\}$. The recommendation metrics are defined as follows:

- **Recall@K** [29]: The proportion of hit items among $\mathcal{P}_u$ in the Top-K recommendations, i.e., $\text{Recall@}K(u) = |\mathcal{H}_u|/|\mathcal{P}_u|$, and the overall $\text{Recall@}K = \mathbb{E}_{u\sim\mathcal{U}}[\text{Recall@}K(u)]$.
- **NDCG@K** [6]: The Discounted Cumulative Gain in the Top-K recommendations (DCG@K) is defined as $\text{DCG@}K(u) = \sum_{i\in\mathcal{H}_u} 1/\log_2(1+\pi_u(i))$. Since the range of DCG@K will vary with the number of positive items $|\mathcal{P}_u|$, we should consider to normalize DCG@K to $[0, 1]$. The Normalized DCG in the Top-K recommendations (NDCG@K) $= \text{DCG@}K(u)/\text{IDCG@}K(u)$, where IDCG@$K$ is the ideal DCG@K, i.e., $\text{IDCG@}K(u) = \sum_{i=1}^{\min\{K,|\mathcal{P}_u|\}} 1/\log_2(1+i)$. The overall $\text{NDCG@}K = \mathbb{E}_{u\sim\mathcal{U}}[\text{NDCG@}K(u)]$.
- **MRR@K** [7, 28]: The Mean Reciprocal Rank (MRR) is originally defined as the reciprocal of the rank of the first hit item. Here we follow the definition of Argyriou et al. [28]'s to meet the requirements of multi-hit scenarios, i.e., $\text{MRR@}K(u) = \mathbb{E}_{i\sim\mathcal{H}_u}[1/\pi_u(i)]$, and the overall $\text{MRR@}K = \mathbb{E}_{u\sim\mathcal{U}}[\text{MRR@}K(u)]$.

B.3 Baselines

We reproduced the following losses as baselines in our experiments:

- **BPR** [10]: A pairwise loss based on the Bayesian Maximum Likelihood Estimation (MLE). The objective of BPR is to learn a partial order among items, i.e., positive items should be ranked higher than negative items. Furthermore, BPR is a surrogate loss for AUC metric [10, 30]. In our implementation, we follow He et al. [22]'s setting and use the inner product as the similarity function for user and item embeddings.
- **LLPAUC** [44]: A surrogate loss for Recall and Precision. In fact, LLPAUC is a surrogate loss for the lower-left part of AUC. In practice, LLPAUC is a min-max loss.
- **Softmax Loss (SL)** [11]: A SOTA recommendation loss derived from the listwise MLE, i.e., maximizing the probability of the positive items among all items. The effectiveness of SL has been thoroughly reviewed in Sections 2 and 3. In fact, SL is a special case of PSL with surrogate activation $\sigma = \exp(\cdot)$.
- **AdvInfoNCE** [38]: A DRO-based modification of SL. AdvInfoNCE tries to introduce adaptive negative hardness to pairwise score d_{uij} in SL (cf. Equation (3.1)). In Zhang et al. [38]'s original design, AdvInfoNCE can be seen as a failure case of PSL with surrogate activation $\sigma = \exp(\exp(\cdot))$, as discussed in Section 4.2. In practice, AdvInfoNCE is a min-max loss.
- **BSL** [15]: A DRO-based modification of SL. BSL applies additional DRO on the positive term in the pointwise form of SL.

The hyperparameter settings of each method are detailed in Appendix B.5.

B.4 Backbones

We implemented three popular recommendation backbones in our experiments, including

- **MF** [26]: MF is the most basic but still effective recommendation model, which factorizes the user-item interaction matrix into user and item embeddings. All the embedding-based recommendation models use MF as the first layer. Specifically, we set the embedding size $d = 64$ for all settings, following the setting in Wang et al. [55].
- **LightGCN** [22]: LightGCN is an effective GNN-based recommendation model. LightGCN performs graph convolution on the user-item interaction graph, so as to aggregate the high-order interactions. Specifically, LightGCN simplifies NGCF [55] and only retains the non-parameterized graph convolution operator. In our experiments, we set the number of layers as 2, which aligns with the original setting in He et al. [22].
- **XSimGCL** [45]: XSimGCL is a novel recommendation model based on contrastive learning [12, 67]. Based on 3-layers LightGCN, XSimGCL adds a random noise to the output embeddings of each layer, and introduces the contrastive learning between the final layer and the l^*-th layer,

i.e., adding an auxiliary InfoNCE loss [61] between these two layers. Following the original Yu et al. [45]'s setting, the modulus of random noise between each layer is set as 0.1, the contrastive layer $l^* = 1$ (where the embedding layer is 0-th layer), the temperature of InfoNCE is set as 0.1, and the weight of the auxiliary InfoNCE loss is set as 0.2 (except for the Amazon-Electronic dataset, where the weight is set as 0.05).

B.5 Hyperparameters

B.5.1 Hyperparameter Settings

Optimizer. We use Adam [68] optimizer for training. The learning rate (lr) is searched in $\{10^{-1}, 10^{-2}, 10^{-3}\}$, except for BPR, where the lr is searched in $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. The weight decay (wd) is searched in $\{0, 10^{-4}, 10^{-5}, 10^{-6}\}$. The batch size is set as 1024, and the number of epochs is set as 200. Following the negative sampling strategy in Wu et al. [15], we uniformly sample 1000 negative items for each positive instance in training.

Loss. The hyperparameters of each loss are detailed as follows:

- **BPR**: No other hyperparameters.
- **LLPAUC**: Following Shi et al. [44]'s setting, the hyperparameters $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\beta \in \{0.01, 0.1\}$ are searched.
- **Softmax Loss (SL)**: The temperature $\tau \in \{0.005, 0.025, 0.05, 0.1, 0.25\}$ is searched.
- **AdvInfoNCE**: The temperature τ is searched in the same space as SL. The other hyperparameters are fixed as the original setting in Zhang et al. [38]. Specifically, the negative weight is set as 64, the adversarial learning will be performed every 5 epochs, with the adversarial learning rate as 5×10^{-5}.
- **BSL**: The temperatures τ_1, τ_2 for positive and negative terms are searched in the same space as SL, respectively.
- **PSL**: The temperature τ is searched in the same space as SL.

B.5.2 Optimal Hyperparameters

The hyperparameters we search include the learning rate (lr), weight decay (wd), and other hyperparameters: $\{\alpha, \beta\}$ for LLPAUC, $\{\tau\}$ for SL, AdvInfoNCE, and PSL, $\{\tau_1, \tau_2\}$ for BSL.

IID optimal hyperparameters. Table B.2 shows the optimal hyperparameters of IID setting, including four datasets (Amazon-Book, Amazon-Electronic, Amazon-Movie, Gowalla) and three backbones (MF, LightGCN, XSimGCL).

OOD optimal hyperparameters. Table B.3 shows the optimal hyperparameters of OOD setting on MF backbone, including four datasets (Amazon-CD, Amazon-Electronic, Gowalla, Yelp2018).

Noise optimal hyperparameters. The Noise setting uses the optimal hyperparameters of IID setting, as listed in Table B.2. We compare the performance of each method under different noise ratios $p \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ on MF backbone and four IID datasets (Amazon-Book, Amazon-Electronic, Amazon-Movie, Gowalla).

https://doi.org/10.52202/079017-3844

Table B.2: Optimal hyperparameters of IID setting.

Model	Loss	Amazon-Book			Amazon-Electronic		
		lr	wd	others	lr	wd	others
MF	BPR	10^{-4}	0		10^{-3}	10^{-5}	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.5, 0.01}
	AdvInfoNCE	10^{-2}	0	{0.05}	10^{-1}	0	{0.1}
	SL	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	BSL	10^{-1}	0	{0.25, 0.025}	10^{-1}	0	{0.25, 0.1}
	PSL-tanh	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	PSL-atan	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	PSL-relu	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
LightGCN	BPR	10^{-3}	0		10^{-2}	10^{-6}	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.5, 0.01}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-2}	0	{0.1}
	SL	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	BSL	10^{-1}	0	{0.25, 0.025}	10^{-2}	0	{0.1, 0.1}
	PSL-tanh	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	PSL-atan	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	PSL-relu	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
XSimGCL	BPR	10^{-4}	10^{-5}		10^{-2}	0	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.3, 0.01}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-1}	0	{0.1}
	SL	10^{-1}	0	{0.025}	10^{-2}	0	{0.1}
	BSL	10^{-1}	0	{0.025, 0.025}	10^{-1}	0	{0.05, 0.1}
	PSL-tanh	10^{-2}	0	{0.025}	10^{-1}	0	{0.1}
	PSL-atan	10^{-2}	0	{0.025}	10^{-1}	0	{0.1}
	PSL-relu	10^{-1}	0	{0.025}	10^{-1}	0	{0.1}

Model	Loss	Amazon-Movie			Gowalla		
		lr	wd	others	lr	wd	others
MF	BPR	10^{-3}	10^{-6}		10^{-3}	10^{-6}	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.7, 0.01}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	SL	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	BSL	10^{-2}	0	{0.25, 0.05}	10^{-1}	0	{0.1, 0.05}
	PSL-tanh	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-atan	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-relu	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
LightGCN	BPR	10^{-3}	0		10^{-3}	0	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.7, 0.01}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	SL	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	BSL	10^{-1}	0	{0.025, 0.05}	10^{-1}	0	{0.025, 0.05}
	PSL-tanh	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-atan	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-relu	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
XSimGCL	BPR	10^{-4}	10^{-4}		10^{-4}	0	
	LLPAUC	10^{-1}	0	{0.3, 0.01}	10^{-1}	0	{0.7, 0.01}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	SL	10^{-2}	0	{0.05}	10^{-2}	0	{0.05}
	BSL	10^{-1}	0	{0.025, 0.05}	10^{-1}	0	{0.025, 0.05}
	PSL-tanh	10^{-1}	0	{0.1}	10^{-1}	0	{0.05}
	PSL-atan	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-relu	10^{-2}	0	{0.1}	10^{-1}	0	{0.05}

Table B.3: Optimal hyperparameters of OOD setting.

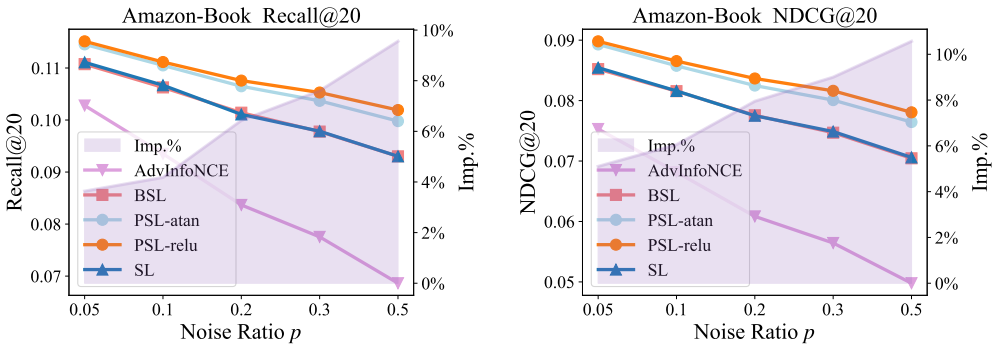
Model	Loss	Amazon-CD			Amazon-Electronic		
		lr	wd	others	lr	wd	others
MF	BPR	10^{-2}	10^{-6}		10^{-2}	10^{-6}	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.7, 0.1}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	SL	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	BSL	10^{-1}	0	{0.05, 0.05}	10^{-1}	0	{0.1, 0.05}
	PSL-tanh	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-atan	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	PSL-relu	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
Model	**Loss**	**Gowalla**			**Yelp2018**		
		lr	**wd**	**others**	**lr**	**wd**	**others**
MF	BPR	10^{-3}	0		10^{-3}	0	
	LLPAUC	10^{-1}	0	{0.7, 0.01}	10^{-1}	0	{0.7, 0.01}
	AdvInfoNCE	10^{-1}	0	{0.05}	10^{-1}	0	{0.05}
	SL	10^{-1}	0	{0.025}	10^{-1}	0	{0.05}
	BSL	10^{-1}	0	{0.25, 0.025}	10^{-1}	0	{0.1, 0.05}
	PSL-tanh	10^{-1}	0	{0.025}	10^{-1}	0	{0.025}
	PSL-atan	10^{-1}	0	{0.025}	10^{-1}	0	{0.025}
	PSL-relu	10^{-1}	0	{0.025}	10^{-1}	0	{0.025}

https://doi.org/10.52202/079017-3844

C Supplementary Experiments

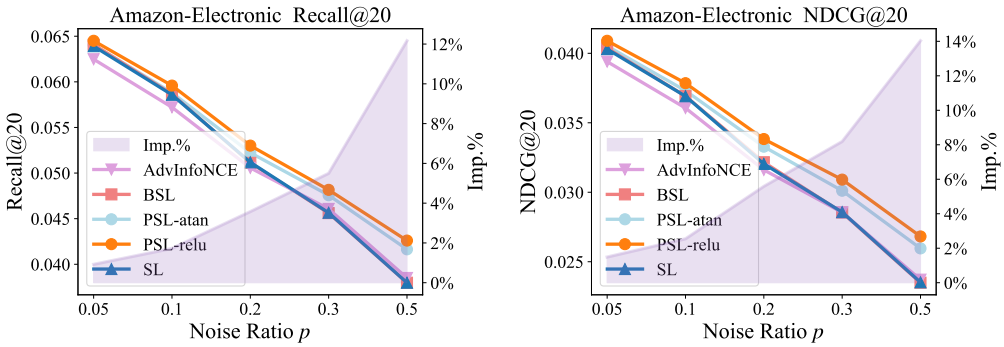
C.1 Noise Results

The Recall@20 and NDCG@20 results under Noise setting on four datasets (Amazon-Book, Amazon-Electronic, Amazon-Movie, Gowalla) are shown in Figures C.1 to C.4.



(a) Amazon-Book (Recall@20) (b) Amazon-Book (NDCG@20)

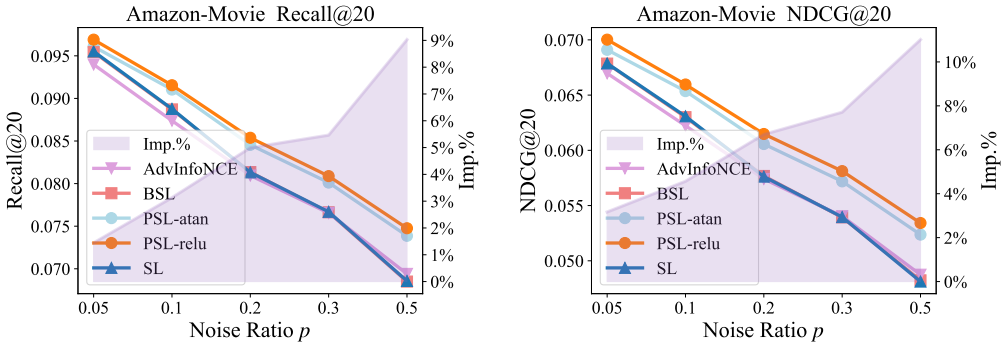
Figure C.1: Noise results on Amazon-Book dataset.



(a) Amazon-Electronic (Recall@20) (b) Amazon-Electronic (NDCG@20)

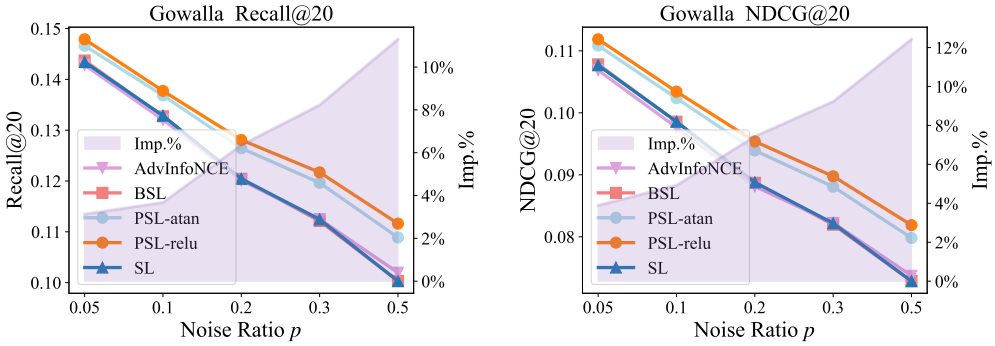
Figure C.2: Noise results on Amazon-Electronic dataset.

https://doi.org/10.52202/079017-3844



(a) Amazon-Movie (Recall@20) (b) Amazon-Movie (NDCG@20)

Figure C.3: Noise results on Amazon-Movie dataset.



(a) Gowalla (Recall@20) (b) Gowalla (NDCG@20)

Figure C.4: Noise results on Gowalla dataset.

https://doi.org/10.52202/079017-3844

C.2 PSL-softplus Results

BPR uses Softplus [69] as $\log \sigma$, i.e., $\sigma(d_{uij}) = \exp(d_{uij}) + 1$, which is looser than SL. That is, this surrogate activation is not a suitable choice for PSL. We call this PSL variant as **PSL-softplus**.

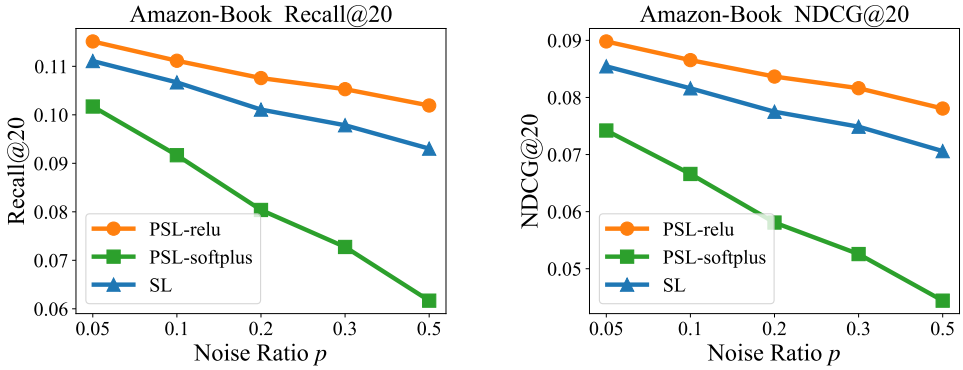
In this section, we conduct experiments to evaluate the performance of PSL-softplus with surrogate activation $\sigma(d_{uij}) = \exp(d_{uij}) + 1$. The IID, OOD, and Noise results of PSL-softplus are shown in Tables C.4 and C.5, Figures C.5 to C.8, respectively. Results demonstrate that PSL-softplus is inferior to SL and three PSLs in all settings. This confirms our claim – the choice of surrogate activation σ is crucial, and an unreasonable or intuitive design will decrease the accuracy.

Table C.4: IID results of PSL-softplus. The results of SL, PSL-tanh, PSL-atan, and PSL-relu have been listed in Table 1. The blue-colored results are better than PSL-softplus.

Model	Loss	Amazon-Book		Amazon-Electronic		Amazon-Movie		Gowalla	
		Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
MF	SL	0.1559	0.1210	0.0821	0.0529	0.1286	0.0929	0.2064	0.1624
	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
	PSL-softplus	0.1536	0.1149	0.0826	0.0522	0.1280	0.0919	0.2053	0.1613
LightGCN	SL	0.1567	0.1220	0.0823	0.0526	0.1304	0.0941	0.2068	0.1628
	PSL-tanh	0.1575	0.1233	0.0825	0.0532	0.1300	0.0947	0.2091	0.1648
	PSL-atan	0.1575	0.1233	0.0825	0.0532	0.1300	0.0948	0.2091	0.1648
	PSL-relu	0.1575	0.1233	0.0830	0.0536	0.1300	0.0953	0.2086	0.1648
	PSL-softplus	0.1536	0.1152	0.0814	0.0514	0.1296	0.0932	0.2053	0.1613
XSimGCL	SL	0.1549	0.1207	0.0772	0.0490	0.1255	0.0905	0.2005	0.1570
	PSL-tanh	0.1567	0.1225	0.0790	0.0501	0.1308	0.0926	0.2034	0.1591
	PSL-atan	0.1565	0.1225	0.0792	0.0502	0.1253	0.0917	0.2035	0.1591
	PSL-relu	0.1571	0.1228	0.0801	0.0507	0.1313	0.0935	0.2037	0.1593
	PSL-softplus	0.1545	0.1161	0.0770	0.0484	0.1242	0.0894	0.1996	0.1557

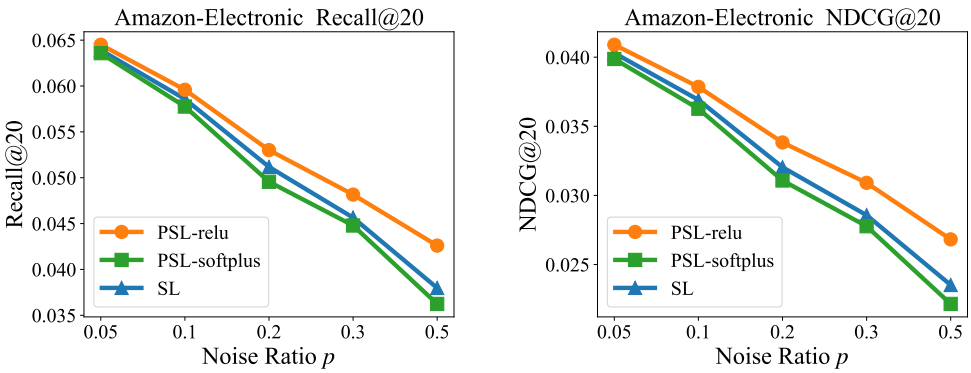
Table C.5: OOD results of PSL-softplus. The results of SL, PSL-tanh, PSL-atan, and PSL-relu have been listed in Table 2. The blue-colored results are better than PSL-softplus.

Loss	Amazon-CD		Amazon-Electronic		Gowalla		Yelp2018	
	Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
SL	0.1184	0.0815	0.0230	0.0142	0.1006	0.0737	0.0349	0.0224
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
PSL-softplus	0.1169	0.0799	0.0232	0.0139	0.0909	0.0665	0.0346	0.0222



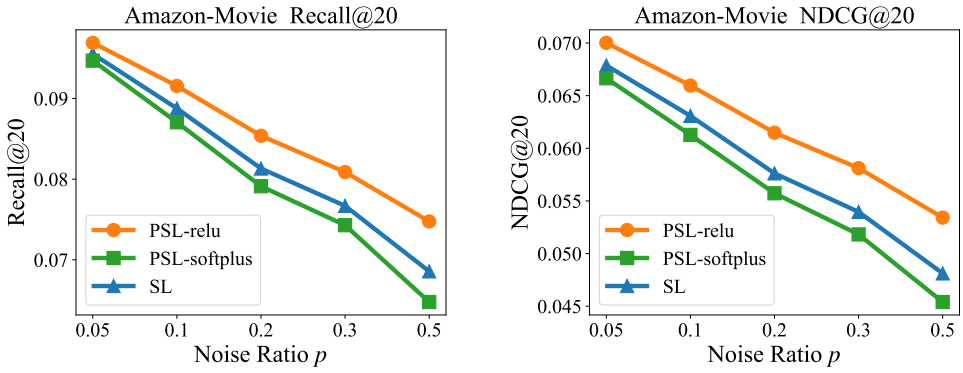
(a) Amazon-Book (Recall@20) (b) Amazon-Book (NDCG@20)

Figure C.5: Noise results of PSL-softplus on Amazon-Book dataset.



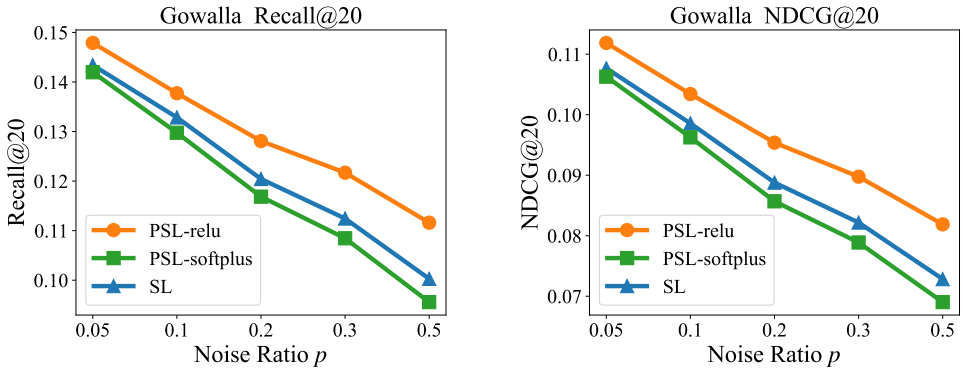
(a) Amazon-Electronic (Recall@20) (b) Amazon-Electronic (NDCG@20)

Figure C.6: Noise results of PSL-softplus on Amazon-Electronic dataset.



(a) Amazon-Movie (Recall@20) (b) Amazon-Movie (NDCG@20)

Figure C.7: Noise results of PSL-softplus on Amazon-Movie dataset.



(a) Gowalla (Recall@20) (b) Gowalla (NDCG@20)

Figure C.8: Noise results of PSL-softplus on Gowalla dataset.

https://doi.org/10.52202/079017-3844

C.3 Comparisons of Two Extension Forms

In this section, we compare the two different extension forms from SL to PSL, i.e., **outside form** $\sigma(d_{uij})^{1/\tau}$ and **inside form** $\sigma(d_{uij}/\tau)$. As discussed in Section 4.2, the outside form scales in the value domain, while the inside form scales in the definition domain. Therefore, the inside form will lead to certain drawbacks: 1) the condition (4.2) must be satisfied over the entire $d_{uij} \in \mathbb{R}$ to ensure a tighter DCG surrogate loss (cf. Lemma 4.1), which is hard to achieve; 2) the value of $\sigma(d_{uij}/\tau)$ and its gradient may be quickly exploded when $\tau \to 0$, as the range of d_{uij}/τ is hard to control, which may cause numerical instability.

To empirically compare the above two extension forms, we conduct experiments on MF backbone and four IID datasets. Specifically, since there exists serious numerical instability, we expand the range of τ to $\{0.005, 0.025, 0.05, 0.1, 0.25, 0.5, 1.0\}$ for the inside form, where the outside form remains the same search space $\tau \in \{0.005, 0.025, 0.05, 0.1, 0.25\}$. The results are shown in Table C.6, demonstrating that the outside form is superior to the inside form in all cases.

Table C.6: Extension forms comparisons on MF under IID setting. The blue-colored results are better than the counterpart.

Form	Loss	Amazon-Book		Amazon-Electronic		Amazon-Movie		Gowalla	
		Recall	NDCG	Recall	NDCG	Recall	NDCG	Recall	NDCG
$\sigma(d_{uij})^{1/\tau}$	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
$\sigma(d_{uij}/\tau)$	PSL-tanh	0.1415	0.1041	0.0767	0.0494	0.0876	0.0590	0.1956	0.1507
	PSL-atan	0.0307	0.0213	0.0453	0.0268	0.0363	0.0247	0.0982	0.0727
	PSL-relu	0.1366	0.1053	0.0723	0.0452	0.1210	0.0855	0.1732	0.1304

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract accurately conclude this paper's contribution, including theoretical findings and experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of this work is discussed in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-3844

Justification: Full theory assumptions and proofs are provided in Section 2 to Section 4, and Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the experimental results in Section 5, Appendices B and C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code are provided in https://github.com/Tiny-Snow/IR-Benchmark.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the experimental details are included in Section 5, Appendices B and C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The p-value of the main performance is reported in Tables 1 and 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-3844

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sufficient information on the computer resources is provided in Appendix B.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work improves the effect of recommendation loss and has no societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no risks for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original paper that produced the codes and datasets are cited in the paper without any omissions.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

 https://doi.org/10.52202/079017-3844

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper are well documented and the details of the code are included.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3844