Advancing Loss Functions in Recommender Systems: A Comparative Study with a Rényi Divergence-Based Solution

Shengjia Zhang[1,2], Jiawei Chen[1,2,3,*], Changdong Li[2], Sheng Zhou[2], Qihao Shi[2], Yan Feng[1,2], Chun Chen[1,2], Can Wang[1,3]

[1] State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2] College of Computer Science, Zhejiang University, China
[3] Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
{shengjia.zhang, sleepyhunt, lichangdongtw, zhousheng_zju, shiqihao321, fengyan, chenc, wcan}@zju.edu.cn

Abstract

Loss functions play a pivotal role in optimizing recommendation models. Among various loss functions, Softmax Loss (SL) and Cosine Contrastive Loss (CCL) are particularly effective. Their theoretical connections and differences warrant in-depth exploration. This work conducts comprehensive analyses of these losses, yielding significant insights: 1) Common strengths — both can be viewed as augmentations of traditional losses with Distributional Robust Optimization (DRO), enhancing robustness to distributional shifts; 2) Respective limitations — stemming from their use of different distribution distance metrics in DRO optimization, SL exhibits high sensitivity to false negative instances, whereas CCL suffers from low data utilization. To address these limitations, this work proposes a new loss function, DrRL, which generalizes SL and CCL by leveraging Rényi-divergence in DRO optimization. DrRL incorporates the advantageous structures of both SL and CCL, and can be demonstrated to effectively mitigate their limitations. Extensive experiments have been conducted to validate the superiority of DrRL on both recommendation accuracy and robustness.

Code —
https://github.com/cynthia-shengjia/AAAI-2025-DrRL

Introduction

Recommender Systems (RS) (Ricci, Rokach, and Shapira 2021; Gao et al. 2023b; Cui et al. 2024; Liao et al. 2024) are pivotal in delivering personalized suggestions across various online services. Collaborative Filtering (CF) has emerged as an effective method, learning user preferences from historical interactions (He et al. 2017; Shi, Larson, and Hanjalic 2014). Loss functions, which direct the optimization pathways of models, are critically important. Traditionally, RS have primarily utilized point-wise losses, such as Binary Cross-Entropy (BCE) (Johnson et al. 2014) and Mean Squared Error (MSE) (Pan et al. 2008), or pairwise losses like Bayesian Personalized Ranking (BPR) (Rendle et al. 2009). Recent years have witnessed a booming interest in exploring novel loss functions, leading to substantial research advancements in this domain.

*Corresponding author: sleepyhunt@zju.edu.cn

By scrutinizing various existing loss functions, we identify two particularly effective types: 1) Softmax Loss (SL) (Wu et al. 2024c), which employs a softmax function to normalize model predictions and enhance the scores of positive instances relative to negatives; 2) Cosine Contrastive Loss (CCL) (Mao et al. 2021), which augments traditional losses by integrating a truncation mechanism that filters out negative instances with low prediction scores. Figure 1 provides empirical evidence — SL and CCL outperform traditional losses by a significant margin (over 20% on average) on two real-world datasets. These impressive observations pose a compelling question: **What are the inherent common strengths of SL and CCL, and can we develop a new loss function based on these advantages?**

This study conducts comprehensive theoretical analyses to demystify this aspect. Despite their distinct structures, we demonstrate that both SL and CCL can equivalently augment traditional losses with *Distributional Robust Optimization* (DRO) (Lin, Fang, and Gao 2022). DRO is a theoretically robust optimization framework that extends model optimization beyond observed training distributions to a broader family of potential distributions with perturbations. Given the common occurrence of distribution shifts in RS — such as evolving user preferences (Wang et al. 2022b) and inherent biases in data collection (Chen et al. 2020) — the efficacy of DRO-enhanced losses is anticipated.

Beyond their shared advantages, the distinctions and limitations of SL and CCL are also revealed. The primary difference lies in the types of distribution perturbations employed in optimizations — SL uses perturbations constrained by KL-divergence (Wu et al. 2024a), whereas CCL uses perturbations constrained by worst-case regret-divergence (Duchi and Namkoong 2021). These choices lead to the following limitations:

- **SL's sensitivity to noise**: SL employs a KL-divergence-based DRO, this type of DRO has been shown to be highly sensitive to noise (Zhai et al. 2021; Nietert, Goldfeld, and Shafiee 2024) (*e.g.,* false negative instances in RS). Specifically, SL assigns disproportionately large weights to negative instances with higher scores, governed by an exponential function. False negatives, which are common in RS and often result from user unawareness rather than disinterest (Chen et al. 2019; Gao et al.

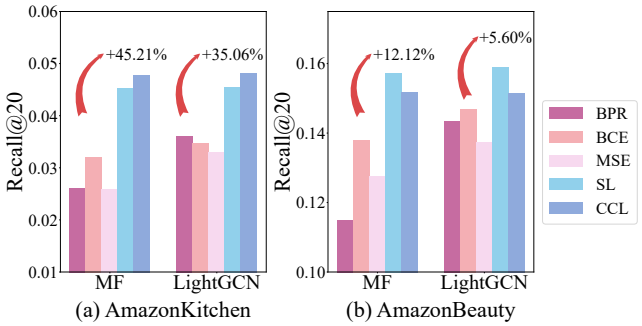

Figure 1: Illustration of how SL and CCL outperforms traditional losses on two recommendation backbones. The detailed experimental settings refer to section .

2022b; Chen et al. 2023), receive excessive emphasis, leading to performance degradation.

- **CCL's low data utilization**: While the truncation mechanism promises enhanced out-of-distribution robustness and model convergence, it naturally reduces data utilization, which is particularly serious in CCL. Specifically, we find that an excessively large proportion (often over 90%) of negative instances are filtered out by CCL. Although these lower-scored negative instances individually contribute less to the gradient than higher-scored ones, their large quantity could still offer valuable training signals. Additionally, the optimal truncation threshold is treated as a hyper-parameter requiring manual tuning, which incurs additional parameter tuning efforts.

To address these limitations, we propose a new loss function, named **Distributional Robust Rényi Loss** (DrRL), which generalizes SL and CCL by employing Rényi-divergence-constrained DRO optimization. Rényi divergence (Van Erven and Harremos 2014), a broad family of divergences that includes common forms like KL-divergence and χ^2 divergence, allows our DrRL to inherit and enhance the strengths of SL and CCL while circumventing their drawbacks: 1) DrRL retains the weighting strategy from SL but provides flexible control over the shape of the weighting distribution, leveraging a polynomial distribution whose order can be flexibly adjusted. This mitigates the excessive impact of false negatives, thereby enhancing robustness to noise. 2) DrRL also inherits the truncation strategy from CCL but demonstrates better data utilization. Additionally, DrRL offers a theoretically sound strategy to learn fine-grained truncation thresholds, avoiding the need for tedious parameter tuning.

In summary, this work makes the following contributions:

- Conducting comprehensive theoretical and empirical analyses to elucidate the theoretical connections between SL and CCL and highlight their limitations.
- Proposing a new recommendation loss function, Distributional Robust Rényi Loss (DrRL), which leverages Rényi-divergence-constrained DRO to inherit the advantages and circumvent the limitations of SL and CCL.
- Performing extensive experiments to demonstrate the effectiveness and robustness of DrRL over existing losses.

Preliminaries

Task Formulation

This work focuses on collaborative filtering (CF) (He et al. 2017; Shi, Larson, and Hanjalic 2014), a conventional recommendation scenario. Let $\mathcal{U}$ and $\mathcal{I}$ denote a user set and an item set. Let $\mathcal{O} \subset \mathcal{U} \times \mathcal{I}$ denote the set of the observed interactions. A user-item pair (u, i) in $\mathcal{O}$ indicates that the user u has interacted with the item i (*e.g.,* click, purchase, etc). For convenience, we define the set of items that user u has interacted with as the positive item set $\mathcal{I}_u^+ = \{i \in \mathcal{I} : (u, i) \in \mathcal{O}\}$, with the remaining items forming the negative item set $\mathcal{I}_u^- = \mathcal{I} \setminus \mathcal{I}_u^+$. The goal of RS is to recommend items to each user that they may be interested in.

Embedding-based methods are widely utilized in RS. These methods first map users and items into d-dimensional embeddings $\mathbf{e}_u$ and $\mathbf{e}_i$, and then generate model predictions $f(u, i)$ with embedding similarity. Recent work has demonstrated cosine similarity is particularly effective (Chen et al. 2023), *i.e.,* $f(u, i) = \frac{\mathbf{e}_u^\top \mathbf{e}_i}{\|\mathbf{e}_u\| \|\mathbf{e}_i\|}$. For convenience, this work adopts cosine similarity for analysis, though our findings can be generalized to other similarity metrics.

Recommendation Loss Functions

Beyond the traditional point-wise and pair-wise loss functions, recent years have witnessed a surge of research on novel recommendation losses. Two representative types are:

Softmax Loss (SL) (Wu et al. 2024c). SL normalizes model predictions using the softmax function, enhancing the scores of positive instances relative to negative ones. The loss for each user u can be written as:

$$\begin{aligned}\mathcal{L}_{\text{SL}}(u) &= -\frac{1}{|\mathcal{I}_u^+|} \sum_{i \in \mathcal{I}_u^+} \log \left(\frac{\exp(f(u,i)/\tau)}{\sum_{j \in \mathcal{I}_u^-} \exp(f(u,j)/\tau)} \right) \\ &= -\frac{1}{|\mathcal{I}_u^+|} \sum_{i \in \mathcal{I}_u^+} f(u,i)/\tau + \log \left(\sum_{j \in \mathcal{I}_u^-} \exp(f(u,j)/\tau) \right),\end{aligned} \tag{1}$$

where τ is a temperature parameter for rescaling the model predictions. Building on SL, recent works have proposed various improvements, such as enhancing popularity debiasing (Zhang et al. 2022) and out-of-distribution (OOD) robustness (Wu et al. 2024b). Since SL forms the foundation for these variants and our analysis can be generalized to them, we focus on the basic SL for our analysis.

Cosine Contrastive Loss (CCL) (Mao et al. 2021). CCL incorporates a *truncation* mechanism in the classical point-wise loss:

$$\mathcal{L}_{\text{CCL}}(u) = -\frac{1}{|\mathcal{I}_u^+|} \sum_{i \in \mathcal{I}_u^+} f(u,i) + \frac{\alpha}{|\mathcal{I}_u^-|} \sum_{j \in \mathcal{I}_u^-} (f(u,j) - \beta)_+, \tag{2}$$

where the symbol $(\cdot)_+$ denotes $\max\{\cdot, 0\}$, β is a margin parameter controlling the truncation threshold, and α is a parameter that rescales the contribution of the negative part.

The truncation mechanism in CCL filters out negative instances with scores lower than β during the model training.

As can be observed, SL and CCL have quite different loss structures, particularly in their treatment of the negative part. In the following sections, we will explore their connections and common strengths.

Distributionally Robust Optimization

The success of machine learning models typically relies on the assumption of independent and identically distributed (IID), *i.e.,* test data is sampled from the same distribution as the observed training data. However, this assumption often fails in real-world scenarios, leading to performance degradation. *Distributionally Robust Optimization* (DRO) has been demonstrated to be effective in mitigating this issue (Lin, Fang, and Gao 2022). DRO extends model optimization beyond the observed training distribution to a broader family of potential distributions with perturbations. Specifically, DRO aims to minimize the worst-case expected loss over a set of potential distributions Q, which surround the observed training distribution Q_0 and are constrained by a distance metric $D(Q, Q_0)$ within a radius η. The objective is formulated as follows:

$$\mathcal{L}_{DRO} = \max_{Q} \mathbb{E}_{x\sim Q}[\mathcal{L}(x;\theta)] \quad \text{s.t. } D(Q, Q_0) \leq \eta \tag{3}$$

where models are optimized under the potential distributions Q, which can be understood as an "adversary", empowering model robustness with adversarial distributional perturbations.

Analyses on CCL and SL

In this section, we first conduct theoretical analyses to reveal the common strengths of SL and CCL, followed by a discussion of the inherent limitations in these loss functions.

Connections between CCL and SL

Given the effectiveness of both CCL and SL, uncovering their shared strengths is valuable, as it not only deepens our understanding of loss function mechanisms, but also inspires the development of new loss functions. Recent work (Wu et al. 2024a) has shown that SL can be equivalent to performing DRO on the negative item distribution. In this work, we further demonstrate that CCL also exhibits this advantageous property, despite its loss structure being quite different from SL.

Specifically, let P_u^- represent the observed negative item distribution of user u, *i.e.,* the uniform distribution over the negative item set $\mathcal{I}_u^-$. We present the following lemma:

Lemma 1 *Optimizing recommendation models with SL and CCL can both be equivalent to solving the following DRO objective:*

$$\mathcal{L}_{DRO}(u) = -\frac{1}{|\mathcal{I}_u^+|}\sum_{i\in\mathcal{I}_u^+} f(u,i) + \max_{Q_u} \mathbb{E}_{j\sim Q_u}\left[f(u,j)\right]. \tag{4}$$

SL is constrained by KL-divergence within robust radius η:

$$D_{KL}(Q_u, P_u^-) := \sum_{j\in\mathcal{I}_u^-} Q_u(j)\log\frac{Q_u(j)}{P_u^-(j)} \leq \eta, \tag{5}$$

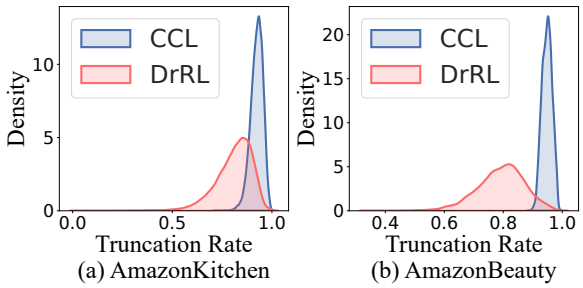

Figure 2: The truncation ratios of DrRL and CCL over users in two real-world datasets.

while CCL is constrained by the worst-case regret-divergence within robust radius $\log(\alpha)$*:*

$$D_{WR}(Q_u, P_u^-) := \sup_{j\in\mathcal{I}_u^-} \log\frac{Q_u(j)}{P_u^-(j)} \leq \log(\alpha), \tag{6}$$

where α is the rescale parameter in Eq.(2).

This lemma clearly demonstrates the common strengths of SL and CCL. Both can be understood as DRO-enhanced versions of the classical point-wise objective:

$$\mathcal{L}_{basic}(u) = -\frac{1}{|\mathcal{I}_u^+|}\sum_{i\in\mathcal{I}_u^+} f(u,i) + \mathbb{E}_{j\sim P_u^-}\left[f(u,j)\right]. \tag{7}$$

SL and CCL improve $\mathcal{L}_{\text{basic}}(u)$ by leveraging DRO on the negative side, where the model is optimized on a set of potential distributions Q_u with distributional perturbations. Due to this adversarial optimization mechanism, DRO endows SL and CCL with robustness against distribution shifts. It is particularly noteworthy that distribution shifts are ubiquitous in RS, *e.g.,* user preferences typically evolve over time (Wang et al. 2022b), and the training data is often polluted by various biases (Chen et al. 2020, 2021; Gao et al. 2023a). This property makes SL and CCL particularly effective.

Limitations of CCL and SL

The above analyses reveal that both CCL and SL can be considered as DRO-enhanced losses, but with different divergence constraints. These different choices of divergences lead to different limitations:

SL is highly sensitive to noise. Recent work (Nietert, Goldfeld, and Shafiee 2024) has demonstrated that DRO with KL-divergence is highly sensitive to noise. To understand this effect, we can derive the worst-case distribution, *i.e.,* $Q_u^* = \arg\max_{Q_u} \mathbb{E}_{j\sim Q_u}\left[f(u,j)\right]$, under which the model is finally optimized:

$$\begin{aligned} Q_u^*(j) =& w_{uj} P_u^-(j) \\ w_{uj} =& \frac{\exp(f(u,j)/\tau)}{\mathbb{E}_{j\sim P_u^-}\left[\exp(f(u,j)/\tau)\right]} \propto \exp(f(u,j)/\tau) \end{aligned} \tag{8}$$

The proof is given by Wu et al. (see Appendix A.6 in (Wu et al. 2024a)). This implies that SL assigns a weight w_{uj} to

each negative instance, with the weights w_{uj} proportional to the exponential of the prediction scores. Given the explosive nature of the exponential function and the fact that τ is usually set to a relatively small value (*e.g.,* $\tau = 0.2$), the weight distribution becomes highly skewed. This skew leads to negative instances with higher scores exerting a disproportionate influence on model training. Such a characteristic renders SL particularly vulnerable to noise, such as false negative instances, which are prevalent in RS. Given the vast number of items and the limited attention of users, some negative instances typically result merely from users' lack of awareness of these items, rather than an active dislike (Chen et al. 2019; Gao et al. 2022b; Wang et al. 2024b). These false negative items, potentially sharing similar features with positive items, can easily receive large prediction scores $f(u, i)$. Consequently, they attain excessively large weights in model training, potentially dominating the optimization directions and severely degrading performance. Table 1 presents the empirical evidence, demonstrating that noisy data receive significantly large weights in SL, typically over 20 times.

Interestingly, recent work (Wu et al. 2024b) claims that SL exhibits robustness to noisy data through DRO. However, we argue that this is not the case. While DRO can indeed enhance model robustness to distribution shifts, it can also increase noise sensitivity rather than decrease, as demonstrated by various studies on DRO (Zhai et al. 2021; Nietert, Goldfeld, and Shafiee 2024). Our analysis also shows that noisy data in SL contribute more, rather than less, to the optimization process. This is further evidenced by empirical results from experiments involving false negative instances (*cf.* Figure 8 in (Wu et al. 2024b)), where the improvements of SL over other baselines do not significantly increase and sometimes even decline as the noise ratio increases.

One might suggest adjusting τ to improve noise resistance. However, τ is a dual parameter of the robust radius η. As discussed in (Wu et al. 2024a), increasing τ would naturally decrease the robust radius η, thereby reducing the model's robustness to distribution shifts. Adjusting τ would significantly decline the merit of SL.

CCL suffers from low data utilization. The truncation mechanism employed by CCL promises to enhance model OOD robustness and accelerate convergence. However, it also naturally reduces data utilization, filtering out a large portion of instances during training, which is particularly severe in CCL. Figure 2 illustrates the ratio of filtered items across users with the optimal β. The ratio is quite extreme, with over 90% of negative instances often being filtered out. Although these lower-scored negative instances individually contribute less to the gradient than higher-scored ones, their large quantity provides valuable training signals.

Another limitation of CCL is that it treats β as a hyperparameter, necessitating manual adjustments and incurring substantial tuning efforts. The reason for this is that the authors of CCL designed it heuristically, without considering its equivalence to DRO. Moreover, we argue that assigning the same β to all users may not be optimal. Figure 3 illustrates that users have diverse preference distributions $f(u, i)$ over items. This is a common phenomenon in practice, since

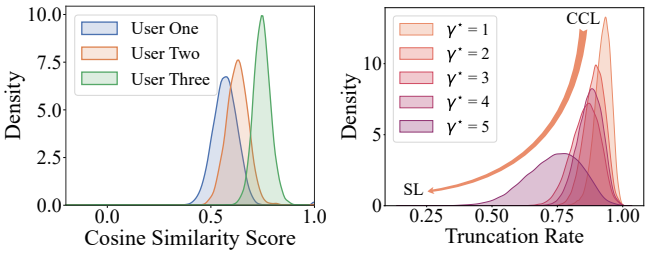

Figure 3: Left: The distribution of $f(u, i)$ for three randomly sampled users; Right: the truncation rate for users with varying γ in the dataset AmazonKitchen.

some users have broader preferences, while others may be more critical. Such diversity naturally motivates us to pursue a personalized β.

Distributionally Robust Rényi Loss

To further capitalize on the strengths of SL and CCL while circumventing their limitations, a straightforward approach is to maintain their DRO-enhanced objective (*cf.* Eq.(4)) while utilizing a more appropriate divergence measure. This work proposes leveraging Rényi divergence (Van Erven and Harremos 2014), a natural generalization of both KL-divergence and worst-case regret-divergence used by SL and CCL. The proposed Distributionally Robust Rényi Loss (DrRL) is formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{DrRL}}(u) = -\frac{1}{|\mathcal{I}_u^+|}\sum_{i\in\mathcal{I}_u^+} f(u,i) + \max_{Q_u} \mathbb{E}_{j\sim Q_u}\left[f(u,j)\right] \\ \text{s.t. } D_\gamma(Q_u, P_u^-) := \sum_{j\in\mathcal{I}_u^-} P_u^-(j)\phi_\gamma\left(\frac{Q_u(j)}{P_u^-(j)}\right) \le \eta \\ \phi_\gamma(t) = \frac{1}{\gamma(\gamma-1)}(t^\gamma - \gamma t + \gamma - 1)\end{aligned} \tag{9}$$

Rényi divergence $D_\gamma(Q_u, P_u^-)$ offers enhanced flexibility by using a parameter γ to adjust the polynomial relationships of the probability distance measure with the probability ratio. Here, we adopt the Cressie-Read family of Rényi divergence (Duchi and Namkoong 2021), which offers analytical benefits. Notably, while Eq.(9) is complex and challenging to optimize directly, it can be substantially simplified with the following lemma:

Lemma 2 *Suppose $\gamma \in (1, +\infty)$, optimizing the objective in Eq.(9) is equivalent to optimizing:*

$$\begin{aligned}\mathcal{L}_{DrRL}(u) = -\frac{\sum_{i\in\mathcal{I}_u^+} f(u,i)}{|\mathcal{I}_u^+|} + c_\gamma(\eta)\left[\frac{1}{|\mathcal{I}_u^-|}\sum_{j\in\mathcal{I}_u^-}(f(u,j)-\beta_u)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}} \\ s.t.\ \beta_u = \arg\min_\beta \left\{\beta + c_\gamma(\eta)\left[\frac{1}{|\mathcal{I}_u^-|}\sum_{j\in\mathcal{I}_u^-}(f(u,j)-\beta)_+^{\gamma^*}\right]^{\frac{1}{\gamma^*}}\right\}\end{aligned} \tag{10}$$

where $\gamma^* = \frac{\gamma}{\gamma-1}$, and $c_\gamma(\eta) = (1+\gamma(\gamma-1)\eta)^{\frac{1}{\gamma}}$.

Loss	Gowalla		AmazonKitchen		AmazonBeauty	
	k_1	k_2	k_1	k_2	k_1	k_2
SL	187.85	110.95	27.96	26.26	24.84	22.48
DrRL	58.16	51.20	20.60	18.68	21.02	19.41

Table 1: The column 'k_1' denotes the average ratios of the largest weight over the average weights; 'k_2' denotes ratios of the weights of false negative instances over the average weights.

The proof is presented in Appendix A.1. This lemma yields a simple closed-form expression of DrRL. It is straightforward and intuitive employing a truncation mechanism while penalizing the prediction scores with a polynomial function. We highlight its significant advantages:

Subsumes SL and CCL. DrRL effectively combines the merits of both CCL and SL, incorporating the truncation mechanism from CCL and weighting strategies from SL. In fact, DrRL can degenerate to SL and CCL under certain conditions, DrRL $\rightarrow$ SL with $\gamma \rightarrow 1$ and DrRL $\rightarrow$ CCL with $\gamma \rightarrow +\infty$ (see Appendix A.4). This demonstrates that our DrRL builds upon the foundations of SL and CCL, ensuring performance at least not worse than these existing methods.

Robustness to False Negative Instances. Unlike SL's exponential weighting, which can lead to disproportionate influence from false negatives due to its explosive nature, DrRL offers a flexible polynomial weight distribution controlled by γ. The following lemma substantiates this:

Lemma 3 *DrRL optimizes models under the following worst-case negative distribution:*

$$Q_u^*(j) = w_{uj} P_u^-(j),$$
$$w_{uj} = c_\gamma(\eta) \frac{(f(u,j) - \beta_u)_+^{\frac{1}{\gamma-1}}}{\mathbb{E}_{j\sim P_u^-}\left[(f(u,j) - \beta_u)_+^{\gamma^*}\right]^{\frac{1}{\gamma}}} \tag{11}$$

where $\mathbb{E}_{j\sim P_u^-}\left[Q_u^*(j)\right] = 1$, $w_{uj} \propto (f(u,j) - \beta_u)_+^{\frac{1}{\gamma-1}}$.

The poof is given in Appendix A.2. The weight w_{uj} is proportional to a polynomial function of prediction scores. The polynomial function is relatively milder and significantly mitigates the impact of false negatives, enhancing model robustness. Table 1 provides the empirical evidence by comparing the weights of DrRL with SL. We can find the weights of noisy data in DrRL are significantly smaller than SL.

Better Data Utilization. Although DrRL also employs a truncation mechanism, it boasts much better data utilization than CCL. Figure 2 compares the ratio of filtered instances between DrRL and CCL across users. DrRL filters significantly fewer instances — on average, 10% less than CCL.

This improvement can be attributed to the inclusion of the weighting strategy in DrRL. The OOD robustness of DrRL can partly originate from the weighting strategy, reducing the need for a large truncation ratio to maintain OOD robustness. Figure 3 also provides empirical evidence that as γ increases, the truncation ratio decreases.

Learnable Personalized Margin Parameter β. Eq.(10) also provides a theoretical framework to learn the personalized margin parameter β. This approach avoids additional hyper-parameter tuning for β and allows for the learning of fine-grained β, being potentially more effective given the diversity of user preference scores. Additionally, the objective with respect to β is convex (*cf.* Appendix A.3), facilitating feasible and efficient learning. In practice, the model and β can be updated iteratively and alternately via gradient descent, incurring minimal computational complexity.

Easy Implementation and Minimal Hyper-parameter Tuning. Appendix C details the implementation of DrRL. It can be straightforward to integrate into various recommendation models with minimal code modifications. Moreover, given β is learnable, DrRL requires minimal hyper-parameter tuning. In our experiments, in many cases, we find that simply setting $c_\gamma(\eta) = 1$ yields satisfactory performance, leaving only γ as a parameter that needs tuning. Nevertheless, fine-tuning $c_\gamma(\eta)$ would be better.

Experiments

Experimental Setups

Datasets. Following CCL and SL (Mao et al. 2021; Wu et al. 2024b), we adopt four widely-used datasets including AmazonKitchen, AmazonElectronics, AmazonBeauty (McAuley et al. 2015) and Gowalla (He et al. 2020). The data information and preprocessing details are presented in Appendix B.

Baseline Methods. The following baselines are included: 1) **MSE** (Pan et al. 2008), **BCE** (Johnson et al. 2014), and **BPR** (Rendle et al. 2009), three classical losses in RS; 2) **CCL** (Mao et al. 2021): a representative loss with leveraging truncation mechanism; 3) **SL** (Wu et al. 2024c): a representative loss with softmax function; 4) **BSL** (Wu et al. 2024b), **AdvInfoNCE** (Zhang et al. 2024): two SOTA losses, improving SL with leveraging DRO on positive side or integrating hardness-aware ranking mechanism; 5) **LLPAUC** (Shi et al. 2024), the SOTA loss approximately optimizing the lower-left partial AUC. Following (Zhang et al. 2024), all losses are integrated with three representative recommendation backbones: 1) the basic **MF** (Rendle et al. 2009); 2) the graph-based model (Kipf and Welling 2016; Wang et al. 2019; Dong et al. 2021; Wu et al. 2022; Gao et al. 2022a; Chen et al. 2024): **LightGCN** (He et al. 2020) and the SOTA **XSimGCL** (Yu et al. 2023).

Hyper-parameter Settings. To maintain fairness across all methods, we tune each method with a very fine granularity to ensure their optimal performance (see Appendix B).

Experimental Results

Overall Performance Comparisons. Table 2 presents the performance of DrRL compared with baselines. Overall, our DrRL consistently outperforms all compared methods across all datasets and backbones. Especially in Gowalla, DrRL achieves impressive improvements — average 6.9% and 5.2% in terms of Recall@20 and NDCG@20 respectively on three backbones. These results demonstrate DrRL can indeed mitigate weaknesses of SL and CCL, making our performance even surpassing other state-of-the-art loss functions by a significant margin. Besides, we find that DRO-enhanced losses (SL, BSL, CCL, AdvInfoNCE, DrRL) gen-

Backbone	Loss	Gowalla		AmazonKitchen		AmazonElectronics		AmazonBeauty	
		Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
MF	MSE	0.1326	0.1087	0.0259	0.0141	0.0744	0.0460	0.1276	0.0728
	BCE	0.1369	0.1065	0.0320	0.0162	0.0715	0.0444	0.1378	0.0781
	BPR	0.1441	0.1202	0.0260	0.0141	0.0663	0.0415	0.1150	0.0668
	CCL	0.1652	0.1282	0.0477	0.0257	0.0870	0.0554	0.1518	0.0897
	LLPAUC	0.1444	0.1219	0.0362	0.0190	0.0793	0.0497	0.1506	0.0877
	SL	0.1623	0.1252	0.0453	0.0262	0.0833	0.0513	0.1573	0.0941
	BSL	0.1613	0.1243	0.0454	0.0262	0.0835	0.0516	0.1587	0.0962
	AdvInfoNCE	0.1628	0.1252	0.0443	0.0259	0.0815	0.0508	0.1596	0.0952
	DrRL	**0.1785**	**0.1435**	**0.0501**	**0.0270**	**0.0895**	**0.0566**	**0.1645**	**0.0978**
	Imp.%	+8.06%*	+11.98%*	+5.04%*	+3.05%*	+2.92%*	+2.10%*	+3.13%*	+1.61%*
LightGCN	MSE	0.1471	0.1146	0.0329	0.0176	0.0781	0.0488	0.1375	0.0773
	BCE	0.1543	0.1303	0.0346	0.0189	0.0795	0.0504	0.1469	0.0836
	BPR	0.1550	0.1280	0.0359	0.0199	0.0797	0.0510	0.1435	0.0823
	CCL	0.1633	0.1281	0.0481	0.0265	0.0890	0.0564	0.1514	0.0881
	LLPAUC	0.1690	0.1433	0.0415	0.0228	0.0876	0.0554	0.1582	0.0926
	SL	0.1609	0.1245	0.0454	0.0258	0.0812	0.0502	0.1589	0.0949
	BSL	0.1611	0.1244	0.0457	0.0258	0.0822	0.0509	0.1556	0.0922
	AdvInfoNCE	0.1628	0.1251	0.0448	0.0254	0.0814	0.0504	0.1587	0.0944
	DrRL	**0.1788**	**0.1447**	**0.0502**	**0.0275**	**0.0902**	**0.0572**	**0.1665**	**0.0974**
	Imp.%	+5.79%*	+1.02%*	+4.40%*	+3.80%*	+1.36%*	+1.49%*	+4.76%*	+2.58%*
XSimGCL	MSE	0.1375	0.1093	0.0358	0.0199	0.0760	0.0494	0.1446	0.0827
	BCE	0.1530	0.1299	0.0356	0.0201	0.0782	0.0509	0.1496	0.0872
	BPR	0.1655	0.1379	0.0382	0.0210	0.0839	0.0539	0.1471	0.0859
	CCL	0.1614	0.1252	0.0472	0.0258	0.0876	0.0561	0.1522	0.0891
	LLPAUC	0.1582	0.1340	0.0435	0.0252	0.0877	0.0561	0.1490	0.0892
	SL	0.1508	0.1146	0.0413	0.0236	0.0748	0.0460	0.1511	0.0895
	BSL	0.1509	0.1144	0.0429	0.0243	0.0770	0.0465	0.1520	0.0898
	AdvInfoNCE	0.1531	0.1148	0.0427	0.0241	0.0771	0.0470	0.1526	0.0907
	DrRL	**0.1774**	**0.1417**	**0.0490**	**0.0266**	**0.0907**	**0.0578**	**0.1608**	**0.0953**
	Imp.%	+7.14%*	+2.81%*	+3.74%*	+3.04%*	+3.45%*	+2.96%*	+5.40%*	+5.14%*

Table 2: Overall performance comparison of DrRL with other losses. The best result is bolded and the runner-up is underlined. Imp.% indicates the relatively improvements of DrRL over the best baselines. The mark '*' suggests the improvement is statistically significant with $p < 0.05$.

Loss	AmazonKitchen		AmazonBeauty	
	Recall@20	NDCG@20	Recall@20	NDCG@20
DrRL-w/o-LP	0.0482	0.0269	0.1582	0.0960
DrRL-w/o-P	0.0477	0.0265	0.1579	0.0922
DrRL	0.0501	0.0270	0.1645	0.0978

Table 3: Ablation Study, we examine two variations of DrRL, where the personalized strategy (DrRL-w/o-P) and the learnable strategy (DrRL-w/o-LP) are omitted.

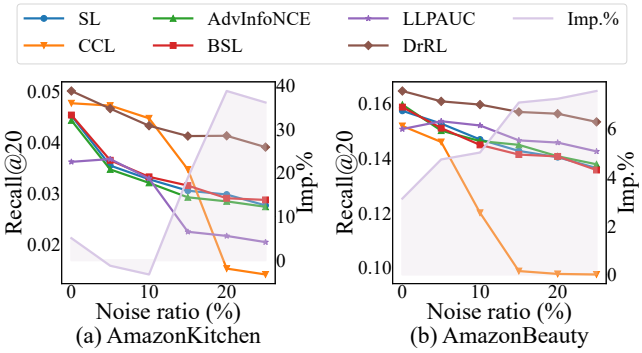

Figure 4: Performance comparisons with the varying ratios of false negative instances. We also present the relative improvements achieved by DrRL over the best baselines.

erally exhibit superior effectiveness. This outcome underscores the significance of OOD robustness in RS.

Performance on Noisy Data. Figure 4 illustrates the noise robustness of compared losses. Here, we closely follow (Wu et al. 2024b) and injects a certain proportion of positive items as false negative during negative sampling. As shown, DrRL consistently outperforms the compared loss functions across most noise ratios. More notably, as the noise ratio increases, DrRL demonstrates even greater improvements over the best baselines.

Another interesting observation is that while CCL exhibits a certain degree of noise robustness at lower noise ratios, its performance deteriorates rapidly as the noise level increases. This phenomenon can be explained as follows: CCL assigns uniform weights to the preserved negative instances rather than giving greater weight to those with higher prediction scores, which initially provides good robustness. However, as the noise ratio increases, the proportion of false

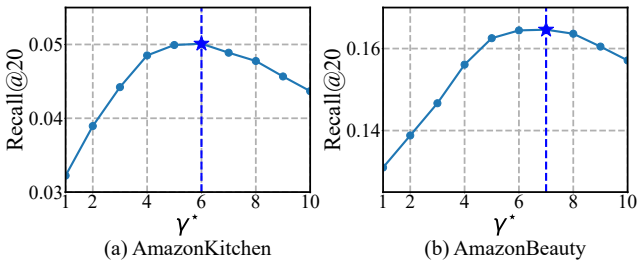

Figure 5: The impact of varying γ.

Loss	AmazonKitchen		AmazonElectronics	
	Recall@20	NDCG@20	Recall@20	NDCG@20
BPR	0.01429	0.00749	0.04390	0.02718
CCL	0.02053	0.01065	0.05013	0.03065
LLPAUC	0.01661	0.00860	0.04637	0.03135
SL	0.01718	0.00914	0.04311	0.02508
BSL	0.01665	0.00907	0.04017	0.02377
AdvInfoNCE	0.01672	0.00892	0.04503	0.02554
DrRL	**0.02077**	**0.01114**	**0.05199**	**0.03160**

Table 4: Performance comparisons under temporal shifts.

negatives among the preserved instances significantly rises. Ultimately, this leads to a scenario where almost exclusively noisy data contributes to the training, resulting in the collapse of CCL's performance under high noise conditions.

Performance under Distribution Shifts. To estimate the robustness of losses against distribution shifts, we follow (Wang et al. 2024a) to construct OOD testing scenario — we divide the training and testing dataset according to the interaction time, where the temporal bias is contained (*cf.* Appendix B). Table 4 presents the results of various loss functions with temporal shifts. DrRL achieves the best performance, demonstrating its superior OOD robustness. The effectiveness can be attributed to the incorporation of Rényi divergence. As a broader family of measures, including the divergences used in SL and CCL, Rényi divergence provides greater flexibility for adapting to complex OOD scenarios.

Ablation Study. In Table 3, we evaluate two variants of DrRL, where the personalized strategy (P) and the learnable strategy (LP) are omitted (w/o). DrRL-w/o-LP outperforms DrRL-w/o-P by a large margin, demonstrating the necessity of leveraging personalized β. Additionally, we observe that DrRL-w/o-P performs slightly worse but remains close to DrRL-w/o-LP. This result indicates that our learning algorithm can indeed find an appropriate β, avoiding the need for exhaustive hyper-parameter tuning.

The Impact of γ. Figure 5 illustrates the model performance with $\gamma^*(\gamma^* = \frac{\gamma}{\gamma-1})$. As shown, model performance initially increases with rising γ^* but declines when γ^* is further increased. This outcome corresponds with our theoretical expectations. Specifically, when γ^* is set close to 1, DrRL degenerates into the basic CCL; when γ^* is set to a sufficiently large value, DrRL approximates SL. Optimal performance is observed at an intermediate value of γ^*, where DrRL effectively integrates the strengths of both SL and CCL, while substantially alleviating their respective limitations. A larger γ^* enhances data utilization as Figure 5 demonstrates, while smaller γ^* flattens the weight distribution and enhances the noise robustness. Reflecting these dual effects, we observe a concave performance curve with respect to γ^*.

Related Work

As this work focuses on loss functions, here we mainly introduce related work on this topic (see Appendix D for recommendation models and recommendation robustness).

Loss functions have drawn increasing attention within the recommendation community. Initially, RS primarily utilized point-wise losses, treating recommendations as classification tasks, including MSE (Pan et al. 2008) and BCE (Johnson et al. 2014). Subsequently, Rendle et al. introduced pairwise loss through Bayesian personalized ranking (Rendle et al. 2009). In more recent developments, there has been a surge in publications on loss functions. Notably, Softmax Loss (SL) (Wu et al. 2024c) and Cosine Contrastive Loss (CCL) (Mao et al. 2021) are particularly effective. There are also some improvements over SL including BSL (Wu et al. 2024b) that applies DRO on positive distributions, AdvInfoNCE (Zhang et al. 2024) that introduces a hardness-aware ranking mechanism, and BC (Zhang et al. 2023) and PopDCL (Liu et al. 2023) that integrate bias-aware terms.

Beyond these developments, explorations from other perspectives have also been conducted. For instance, automated machine learning has been employed to search for optimal loss functions among candidates (Li et al. 2022); Some researchers (Rashed, Grabocka, and Schmidt-Thieme 2021; Pu et al. 2024; Shi et al. 2024) study surrogate objectives for NDCG or partial AUC metrics; while others (Wang et al. 2022a; Park et al. 2023) developed loss functions to enhance embedding alignment and uniformity.

This work focuses on analyzing CCL and SL, exploring their properties, and proposing an improved variant. The most closely related work is (Wu et al. 2024b), but we have significant differences: 1) their analyses are merely on SL, while our analyses include both SL and CCL, elucidating their connections and limitations; 2) we introduce a novel loss that generalizes SL and CCL by incorporating Rényi divergence, distinctly different from their approach of integrating KL-divergence DRO in positive instances.

Conclusions

This work studies loss functions in recommender systems. Our comprehensive analyses have identified that both CCL and SL can be considered as enhancements of traditional losses through the application of DRO. However, their utilization of distinct distribution divergence metrics contributes to SL's high sensitivity to noise and CCL's low data utilization. To address these issues, this work introduces DrRL, which employs Rényi divergence within a DRO framework, effectively inheriting and enhancing the beneficial features of both SL and CCL while alleviating their drawbacks.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62372399, 62476244), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS001), OPPO Research Fund, and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

References

Chen, H.; Bei, Y.; Shen, Q.; Xu, Y.; Zhou, S.; Huang, W.; Huang, F.; Wang, S.; and Huang, X. 2024. Macro graph neural networks for online billion-scale recommender systems. In *Proceedings of the ACM on Web Conference 2024*, 3598–3608.

Chen, J.; Dong, H.; Qiu, Y.; He, X.; Xin, X.; Chen, L.; Lin, G.; and Yang, K. 2021. AutoDebias: Learning to debias for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21–30.

Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; and He, X. 2020. Bias and debias in recommender system: a survey and future directions (2020). *arXiv preprint arXiv:2010.03240*.

Chen, J.; Wang, C.; Zhou, S.; Shi, Q.; Feng, Y.; and Chen, C. 2019. Samwalker: Social recommendation with informative sampling strategy. In *The World Wide Web Conference*, 228–239.

Chen, J.; Wu, J.; Wu, J.; Cao, X.; Zhou, S.; and He, X. 2023. Adap-τ: Adaptively modulating embedding magnitude for recommendation. In *Proceedings of the ACM Web Conference 2023*, 1085–1096.

Cui, Y.; Liu, F.; Wang, P.; Wang, B.; Tang, H.; Wan, Y.; Wang, J.; and Chen, J. 2024. Distillation matters: empowering sequential recommenders to match the performance of large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 507–517.

Dong, H.; Chen, J.; Feng, F.; He, X.; Bi, S.; Ding, Z.; and Cui, P. 2021. On the equivalence of decoupled graph convolution network and label propagation. In *Proceedings of the Web Conference 2021*, 3651–3662.

Duchi, J. C.; and Namkoong, H. 2021. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3): 1378–1406.

Gao, C.; Huang, K.; Chen, J.; Zhang, Y.; Li, B.; Jiang, P.; Wang, S.; Zhang, Z.; and He, X. 2023a. Alleviating matthew effect of offline reinforcement learning in interactive recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 238–248.

Gao, C.; Wang, S.; Li, S.; Chen, J.; He, X.; Lei, W.; Li, B.; Zhang, Y.; and Jiang, P. 2023b. CIRS: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems*, 42(1): 1–27.

Gao, C.; Wang, X.; He, X.; and Li, Y. 2022a. Graph neural networks for recommender system. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 1623–1625.

Gao, Y.; Du, Y.; Hu, Y.; Chen, L.; Zhu, X.; Fang, Z.; and Zheng, B. 2022b. Self-guided learning to denoise for robust recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1412–1422.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. 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*, 639–648.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, 173–182.

Johnson, C. C.; et al. 2014. Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, 27(78): 1–9.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Li, Z.; Ji, J.; Ge, Y.; and Zhang, Y. 2022. Autolossgen: Automatic loss function generation for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1304–1315.

Liao, J.; Li, S.; Yang, Z.; Wu, J.; Yuan, Y.; Wang, X.; and He, X. 2024. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1785–1795.

Lin, F.; Fang, X.; and Gao, Z. 2022. Distributionally robust optimization: A review on theory and applications. *Numerical Algebra, Control and Optimization*, 12(1): 159–212.

Liu, Z.; Li, H.; Chen, G.; Ouyang, Y.; Rong, W.; and Xiong, Z. 2023. PopDCL: Popularity-aware Debiased Contrastive Loss for Collaborative Filtering. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1482–1492.

Mao, K.; Zhu, J.; Wang, J.; Dai, Q.; Dong, Z.; Xiao, X.; and He, X. 2021. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1243–1252.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Nietert, S.; Goldfeld, Z.; and Shafiee, S. 2024. Outlier-robust Wasserstein DRO. *Advances in Neural Information Processing Systems*, 36.

Pan, R.; Zhou, Y.; Cao, B.; Liu, N. N.; Lukose, R.; Scholz, M.; and Yang, Q. 2008. One-class collaborative filtering. In *2008 Eighth IEEE international conference on data mining*, 502–511. IEEE.

Park, S.; Yoon, M.; Lee, J.-w.; Park, H.; and Lee, J. 2023. Toward a Better Understanding of Loss Functions for Collaborative Filtering. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2034–2043.

Pu, Y.; Chen, X.; Huang, X.; Chen, J.; Lian, D.; and Chen, E. 2024. Learning-Efficient Yet Generalizable Collaborative Filtering for Item Recommendation. In *Forty-first International Conference on Machine Learning*.

Rashed, A.; Grabocka, J.; and Schmidt-Thieme, L. 2021. 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*, 605–613.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 452–461.

Ricci, F.; Rokach, L.; and Shapira, B. 2021. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, 1–35.

Shi, W.; Wang, C.; Feng, F.; Zhang, Y.; Wang, W.; Wu, J.; and He, X. 2024. Lower-Left Partial AUC: An Effective and Efficient Optimization Metric for Recommendation. In *Proceedings of the ACM on Web Conference 2024*, 3253–3264.

Shi, Y.; Larson, M.; and Hanjalic, A. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 47(1): 1–45.

Van Erven, T.; and Harremos, P. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7): 3797–3820.

Wang, B.; Chen, J.; Li, C.; Zhou, S.; Shi, Q.; Gao, Y.; Feng, Y.; Chen, C.; and Wang, C. 2024a. Distributionally Robust Graph-based Recommendation System. In *Proceedings of the ACM on Web Conference 2024*, 3777–3788.

Wang, B.; Liu, F.; Chen, J.; Wu, Y.; Lou, X.; Wang, J.; Feng, Y.; Chen, C.; and Wang, C. 2024b. Llm4dsr: Leveraing large language model for denoising sequential recommendation. *arXiv preprint arXiv:2408.08208*.

Wang, C.; Yu, Y.; Ma, W.; Zhang, M.; Chen, C.; Liu, Y.; and Ma, S. 2022a. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 1816–1825.

Wang, W.; Lin, X.; Feng, F.; He, X.; Lin, M.; and Chua, T.-S. 2022b. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, 3562–3571.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 165–174.

Wu, J.; Chen, J.; Wu, J.; Shi, W.; Wang, X.; and He, X. 2024a. Understanding contrastive learning via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 36.

Wu, J.; Chen, J.; Wu, J.; Shi, W.; Zhang, J.; and Wang, X. 2024b. BSL: Understanding and improving softmax loss for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, 816–830. IEEE.

Wu, J.; Wang, X.; Gao, X.; Chen, J.; Fu, H.; and Qiu, T. 2024c. On the effectiveness of sampled softmax loss for item recommendation. *ACM Transactions on Information Systems*, 42(4): 1–26.

Wu, S.; Sun, F.; Zhang, W.; Xie, X.; and Cui, B. 2022. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5): 1–37.

Yu, J.; Xia, X.; Chen, T.; Cui, L.; Hung, N. Q. V.; and Yin, H. 2023. XSimGCL: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 36(2): 913–926.

Zhai, R.; Dan, C.; Kolter, Z.; and Ravikumar, P. 2021. DORO: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, 12345–12355. PMLR.

Zhang, A.; Ma, W.; Wang, X.; and Chua, T.-S. 2022. Incorporating bias-aware margins into contrastive loss for collaborative filtering. *Advances in Neural Information Processing Systems*, 35: 7866–7878.

Zhang, A.; Sheng, L.; Cai, Z.; Wang, X.; and Chua, T.-S. 2024. Empowering collaborative filtering with principled adversarial contrastive loss. *Advances in Neural Information Processing Systems*, 36.

Zhang, A.; Zheng, J.; Wang, X.; Yuan, Y.; and Chua, T.-S. 2023. Invariant collaborative filtering to popularity distribution shift. In *Proceedings of the ACM Web Conference 2023*, 1240–1251.