A Generic Learning Framework for Sequential Recommendation with Distribution Shifts

Zhengyi Yang
University of Science and Technology of China
yangzhy@mail.ustc.edu.cn

Xiangnan He*
University of Science and Technology of China
xiangnanhe@gmail.com

Jizhi Zhang
University of Science and Technology of China
cdzhangjizhi@mail.ustc.edu.cn

Jiancan Wu
University of Science and Technology of China
wujcan@gmail.com

Xin Xin
Shandong University
xinxin@sdu.edu.cn

Jiawei Chen
Zhejiang University
sleepyhunt@zju.edu.cn

Xiang Wang*
University of Science and Technology of China
xiangwang1223@gmail.com

ABSTRACT

Leading sequential recommendation (SeqRec) models adopt empirical risk minimization (ERM) as the learning framework, which inherently assumes that the training data (historical interaction sequences) and the testing data (future interactions) are drawn from the same distribution. However, such i.i.d. assumption hardly holds in practice, due to the online serving and dynamic nature of recommender system. For example, with the streaming of new data, the item popularity distribution would change, and the user preference would evolve after consuming some items. Such distribution shifts could undermine the ERM framework, hurting the model's generalization ability for future online serving.

In this work, we aim to develop a generic learning framework to enhance the generalization of recommenders in the dynamic environment. Specifically, on top of ERM, we devise a Distributionally Robust Optimization mechanism for SeqRec (**DROS**). At its core is our carefully-designed distribution adaption paradigm, which considers the dynamics of data distribution and explores possible distribution shifts between training and testing. Through this way, we can endow the backbone recommenders with better generalization ability. It is worth mentioning that DROS is an effective model-agnostic learning framework, which is applicable to general recommendation scenarios. Theoretical analyses show that DROS enables the backbone recommenders to achieve robust performance in future testing data. Empirical studies verify the effectiveness against dynamic distribution shifts of DROS. Codes are anonymously open-sourced at https://github.com/YangZhengyi98/DROS.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Retrieval models and ranking**.

KEYWORDS

Sequential Recommendation, Distributionally Robust Optimization, Robust Learning

ACM Reference Format:
Zhengyi Yang, Xiangnan He, Jizhi Zhang, Jiancan Wu, Xin Xin, Jiawei Chen, and Xiang Wang. 2023. A Generic Learning Framework for Sequential Recommendation with Distribution Shifts. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591624

1 INTRODUCTION

Targeting at making recommendation based on users' previous behavior sequences, sequential recommendation (SeqRec) is becoming increasingly important in online platforms, such as e-commerce, streaming media and social networks [2, 6, 44, 53]. Scrutinizing the current studies on SeqRec [14, 16, 34], we can summarize a common pipeline: treating historical interaction sequences of users as the training data, employing the recommender models upon them to capture the sequential patterns, and then predicting users' future interactions in the testing data.

To parameterize this pipeline and optimize the recommenders, empirical risk minimization (ERM) [3, 14, 26, 27] has become the dominant framework, which minimizes the loss over the empirical training distribution. It inherently assumes that the training and testing data are drawn from the same distribution. However, this assumption hardly holds in real-world scenarios, since it ignores the online serving and dynamic nature of recommender system. In particular, the properties of streaming data (*e.g.,* popularity distribution of items) are usually changing over time [37, 49], thereby making

*Corresponding authors, and they are also affiliated with Institute of Artificial Intelligence, Institute of Dataspace, Hefei Comprehensive National Science Center.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591624

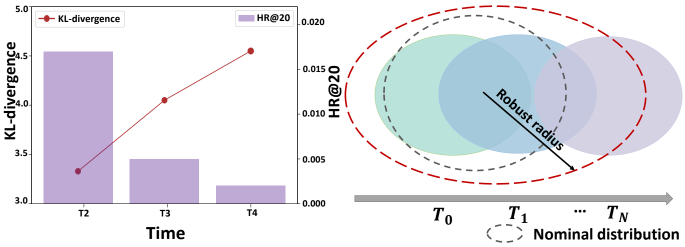


Figure 1: The left figure shows that the recommendation accuracy of GRU4Rec drops with the increase of item distribution discrepancy along time. The right figure illustrates our proposal: after estimating the nominal distribution from historical interactions, we optimize the model within a proper robust radius.

users' sequential patterns dynamically vary across historical and future interactions. This yields the distribution shifts between the training and testing data.

Distribution shifts undermine the ERM framework, and consequently, the generalization of recommenders deteriorates in serving future data. Figure 1 demonstrates this issue with empirical evidence. We first divide the YooChoose [1] data into four disjoint shards in chronological order, and calculate the KL-divergence *w.r.t.* item distribution between the first and other shards. Then a GRU4Rec [14] model is trained upon the first shard, and tested on the rest. Clearly, there is a steady increase in the divergence of item distribution, whereas the dynamic distribution shifts result in severe drop of recommendation accuracy. Therefore, it is crucial to enhance the dynamic adaption of SeqRec models.

Recently, there is increasing interest in using debiasing or data augmentation to enhance the generalization of recommenders [26, 31, 40, 44, 48, 49]. However, they suffer from some inherent limitations when facing the unique characteristics of SeqRec. For debasing strategies [31, 40, 47, 49, 52], they only consider one case of fixed bias per time, thus can hardly deal with the dynamics of SeqRec. For data augmentation strategies [26, 42, 44], they rely heavily on human-designed augmentations and usually have no theoretical guarantee. Some studies [9, 45] leverage prior knowledge of downstream tasks to guide the training, however, the prior knowledge is not always available in real SeqRec scenarios. By far, none of them consider the dynamic characteristics of SeqRec. Therefore, it is desirable to develop effective solutions to strengthen the dynamic adaptation ability of recommenders, better with theoretical guarantees.

Another promising way is Distributionally Robust Optimization (DRO) [1, 19], which hedges against the discrepancy between training and testing distributions. The basic idea is training the model over the distributional family which is determined by a *nominal distribution* with a *robust radius*, so as to handle the distributional uncertainty. However, it is challenging to directly apply DRO to SeqRec, due to the following reasons. First of all, the testing information is usually required to estimate the nominal distribution, which is supposed to lay around the testing data [19, 24, 30]. However, accessing to the user behaviors in future (*i.e.,* testing data) during training is impractical in SeqRec. Secondly, existing DRO methods exert the distributional family generation [35, 50] on the continuous data (*e.g.,* images), and it remains unknown how to do it on the discrete data. Thus, generating the sequences of discrete item ID obstacles the DRO deployment in SeqRec. Worse still, the dynamic characteristics of SeqRec could deteriorate the generation of discrete sequences. In a nutshell, a paradigm of incorporating DRO into SeqRec is until-now lacking to mitigate the distribution shifts, to the best of our knowledge.

In this work, we propose a Distributionally Robust Optimization mechanism for SeqRec (**DROS**). Specifically, DRO requires to estimate the nominal distribution, which is supposed to lay around the testing data [19, 24, 30]. Since it is impractical to access testing data in advance, we make an intuitive assumption that the noimal distribution can be estimated from historical interactions. This assumption is based on our observation in Figure 1 that the dynamics in SeqRec are continuous *w.r.t.* time. It is manifested as the gradual increase of item distribution divergence and the gradual decline of performance. Therefore, we estimate the nominal distribution from historical interactions, which stays close to the testing data based on this observation. To approach the testing data, we further minimize the risks beyond the nominal distribution within a proper robust radius as shown in Figure 1. It makes the sequential recommender promise to take the testing distribution into account and improve the dynamic adaptation ability consequently. Our theoretical analysis admits that if the distance between training and testing data is bounded, the robustness of DROS can be guaranteed. We also empirically demonstrate that our DROS can better adapt to future inference phase in the dynamic SeqRec process.

Our contributions are summarized as follows:

- We reveal the dynamic distribution shifts between the training and testing data in sequential recommendation, which makes ERM suffer from poor generalization.
- We propose a simple yet effective framework equipped with SeqRec-oriented DRO mechanism, such that the sequential recommenders can effectively adapt to dynamic testing stage.
- Theoretical analysis and extensive experiments on three real-world datasets demonstrate the superiority of our DROS.

2 RELATED WORK

This section reviews the work on sequential recommendation , and then discusses the work on DRO that is related to our work from the technique perspective.

2.1 Sequential Recommendation

Sequential recommendation (SeqRec) aims to infer users' preference from their previous interaction sequences. In the early stage, SeqRec is mainly based on Markov Chain [5, 13] or factorization machine [28]. Recently, deep learning models have been applied in SeqRec to model users' behavior sequence, such as recurrent neural network (RNN) [14], convolutional neural network (CNN) [34, 46], and Transformer encoder [7, 10, 12, 16, 25, 32].

Besides, several studies [23, 45] have leveraged domain generalization [17, 20, 29], mainly focusing on pretraining a recommender in one domain and applying it in downstream domains. Another research line is based on data augmentation, which can improve the generalization ability of SeqRec recommenders [4, 26, 33, 44].

[1] https://recsys.acm.org/recsys15/challenge/

Moreover, work of debiasing [8, 31, 36, 38, 40, 49, 52] can remove the side effect of bias issues, from the training data.

Despite effectiveness, these work does not consider the natural dynamicsin SeqRec and applying ERM for optimization. Our work, instead, investigates these dynamics and further promotes the performance of SeqRec recommenders.

2.2 Distributional Robust Optimization

Distributional Robust Optimization (DRO) [1, 15, 18, 19] is an optimization framework for tasks with uncertainty involved – DRO allows distributions of training and testing data to be different within a pre-defined *uncertainty set*:

$$\mathcal{L}_{DRO} = \max_{D(\mu||\mu_0)\leq\rho} \mathbb{E}_{(x,y)\sim\mu}\ell(x,y,\theta), \tag{1}$$

where $\{\mu : D(\mu||\mu_0) \leq \rho\}$ is the *uncertainty set*, μ_0 is called as *nominal distribution* that should be acquired from prior knowledge, and ρ is called as *robust radius* [19, 24, 30].

Recently, DRO has been applied to improve the robustness of vision models [35, 39, 50, 51]. However, directly applying these existing frameworks on SeqRec is challenging with practical issues. Work [35, 50] proposes to generate new image data with DRO, but generating sequence-item piars is more challenging since the ID space in SeqRec is discrete. In SeqRec, [41] proposes to use group DRO [30] to improve user fairness of different groups. Leveraging DRO to address the challenge of dynamic adaptation in SeqRec remains largely unexplored.

3 METHOD

In this section, we first introduce basic notations and commonly-used ERM frameworks in SeqRec, and then elaborate our proposed DROS in detail. Finally, we present theoretical analysis to guarantee the effectiveness of our proposal.

3.1 Problem Formulation

Let $\mathbf{s}_{1:t} = \{v_1, v_2, \ldots, v_t\}$ denote the historical interaction sequence of a user. The task of SeqRec is to recommend the potential next item $\hat{v}_{t+1}$ that best matches her/his current preference. Generally, we can represent a sequential recommender as f, which takes $\mathbf{s}_{1:t}$ as input and outputs the sequence embedding $\mathbf{e}_t$: $\mathbf{e}_t = f(\mathbf{s}_{1:t})$. Subsequently, we can feed $\mathbf{e}_t$ into a prediction layer g to generate the prediction logits $\mathbf{y}$ over all candidate items: $\mathbf{y} = g(\mathbf{e}_t)$. Typically g can be defined as a fully-connected layer or an inner-product layer upon embeddings of candidate items [14, 16, 34].

After acquiring the prediction logits, we can train the sequential recommender under ERM by minimizing loss functions such as Mean Square Error (MSE) [3], Binary Cross Entropy (BCE) [14, 16, 34], or Bayesian Personalized Ranking (BPR) [27]:

$$\begin{aligned}
\mathcal{L}_{MSE} &= \sum_{(\mathbf{s},v)\in O^+} (y_{v|\mathbf{s}} - 1)^2 + \sum_{(\mathbf{s},w)\in O^-} (y_{w|\mathbf{s}} - 0)^2 \\
\mathcal{L}_{BCE} &= -[\sum_{(\mathbf{s},v)\in O^+} \log\sigma(y_{v|\mathbf{s}}) + \sum_{(\mathbf{s},w)\in O^-} \log(1-\sigma(y_{w|\mathbf{s}}))] \\
\mathcal{L}_{BPR} &= -\sum_{(\mathbf{s},v)\in O^+(\mathbf{s},w)\in O^-} \log\sigma(y_{v|\mathbf{s}} - y_{w|\mathbf{s}}),
\end{aligned} \tag{2}$$

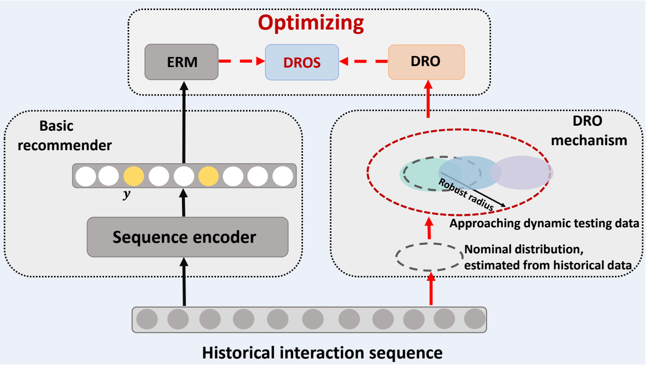

Optimizing
ERM
DROS
DRO
Basic recommender
y
Sequence encoder
DRO mechanism
Robust radius
Approaching dynamic testing data
Nominal distribution, estimated from historical data
Historical interaction sequence

Figure 2: Framework of DROS. We first estimate the nominal distribution from historical data, and optimize distribution within the robust radius to approach the testing data.

where O^+ and O^- denote the dataset of positive and negative samples respectively, $y_{v|\mathbf{s}}$ denotes the predicted score of item v being the next item of sequence $\mathbf{s}$, and σ denotes the Sigmoid function.

3.2 SeqRec-oriented DRO Framework

Note that a potential assumption of ERM is that the training and testing data are drawn from the same distribution, which is not practical in SeqRec due to its dynamic propriety as we discussed in Section 1. Therefore it is insufficient to simply use ERM as the optimization framework, which drives us to develop SeqRec-oriented DRO framework to address this dynamic adaptation challenge.

We next consider how to construct the DRO framework in SeqRec by two important components: 1) the inner loss function ℓ, and 2) the nominal distribution μ_0 defined in Equation (1).

3.2.1 *Inner loss function ℓ.*

Generally, ℓ represents the goodness of the predicted score of given samples, which is often measured by the distance between the predicted score and the score of ground-truth [1, 19]. In SeqRec, it is often the case that training samples are divided into positive samples and negative ones, so there is no absolute score for ground-truth [16, 26, 27, 34]. Inspired by previous work [3] which demonstrates that MSE can achieve plausible performance in recommendation with simple predefined ground-truth scores, we directly use L2-norm to measure the goodness of a given predicted score with a predefined score of ground-truth:

$$\ell(\mathbf{s}, v) = (y_{v|\mathbf{s}} - I(\mathbf{s}, v))^2, \tag{3}$$

where

$$I(\mathbf{s}, v) = \begin{cases} 1, & \text{if } (\mathbf{s}, v) \text{ is a positive sample} \\ 0, & \text{if } (\mathbf{s}, v) \text{ is a negative sample} \end{cases}. \tag{4}$$

3.2.2 *Nominal distribution μ_0.*

Typically in DRO, μ_0 should cover the testing distribution with robust radius ρ, which is the reason why μ_0 should lay around the distribution of the testing data [1, 19]. In SeqRec, for a given sequence $\mathbf{s}$, we desire to acquire the distribution of next item v as the nominal distribution. This poses practical challenges, since we have no access to the testing data beforehand. Towards this end, we further investigate the dynamics in SeqRec, and find that the dynamics is continuous — the gradual increase of divergence of item distribution shown in Figure 1 also reveals the existence of overlap between the distributions of previous interactions and current testing stage. Under these observations, we can safely assume that the distributions of training and testing data should not be too far from each other in SeqRec. Consequently, the nominal distribution can be estimated by item frequency distribution from training data:

$$\mu_0(v_i|\mathbf{s}) = p(v_i) = \left(\frac{D_{v_i}}{\sum_j D_{v_j}}\right)^{\gamma}, \tag{5}$$

where $p(v)$ is the frequency of item v estimated from the training data, D_{v_i}denotes the number of observed interactions for item v_i, and γ is a smoothing factor used in experiment inspired from [49].

Thereafter, we can define the DRO part in DROS following the definition in Equation (1):

$$\mathcal{L}_{DRO} = \max_{D(\mu||\mu_0)\leq\rho} \mathbb{E}_{(\mathbf{s},v)\sim\mu}(y_{v|\mathbf{s}} - I(\mathbf{s},v))^2. \tag{6}$$

3.3 Analysis under KL-divergence

Generally, minimizing our proposed objective Equation (6) can in principle enhance the dynamic adaptation of SeqRec recommenders, but in practice the learning process would be unstable due to the maximization term in $\mathcal{L}_{DRO}$. If we can further acquire a closed form of $\mathcal{L}_{DRO}$, this unstable issue can be addressed. To this end, we further investigate to choose KL-divergence as the measurement of distribution distance D in Equation (6), then the maximization problem of Equation (6) can be derived into a closed form as is shown in Theorem 3.1 and Theorem 3.2 (for a derivation, see Appendix A.2).

Theorem 3.1. *Set $D(\mu||\mu_0) = D_{KL}(\mu||\mu_0)$, the maximization problem in Equation (6) is equivalent to the following form:*

$$\mathcal{L}'_{DRO} = \inf_{\beta\geq 0}\left[\beta\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta}\right) + \beta\rho\right], \tag{7}$$

where β is a Lagrange multiplier.

Note that by applying Theorem 3.1, minimizing $\mathcal{L}_{DRO}$ is equals to minimizing $\mathcal{L}'_{DRO}$ under D set as KL-divergence. But this is still a two-layer optimization problem and directly optimizing is still unstable. Fortunately, we can fix β in Equation (7), then minimizing $\mathcal{L}'_{DRO}$ equals minimizing its tight upper bound (Theorem 3.2).

Theorem 3.2. *Set $\beta = \beta_0$ to be a constant, optimizing*

$$\mathcal{L}''_{DRO} = \beta_0\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right), \tag{8}$$

equals to optimizing all the upper bounds of $\mathcal{L}'_{DRO}$ for $\forall\rho > 0$. And for $\forall\beta_0 > 0$ there always exists a ρ_{β_0} to guarantee the upper bound to be tight.

Note that the robust radius ρ does not appear in Equation (8), which is quite unusual since robust radius is crucial in distance-based robust optimization: if the robust radius is small, the uncertainty set will almost only contain the nominal distribution, and thus DRO is reduced to ERM; if the robust radius is larger, the uncertainty set are likely to contain pathological distributions [19]. Towards this end, we further theoretically demonstrate that the robust radius ρ is highly related to β_0: robust radius decreases monotonically *w.r.t.* β_0 (Theorem 3.3). In practice, if we tune β_0 to be larger, the robust radius would be smaller, and the reverse holds true as well, which is meaningful in guiding our implementation.

Theorem 3.3. *[relation between β and ρ]. ρ_{β_0} decreases monotonically w.r.t β_0, and the following equality holds:*

$$\lim_{\beta_0\to\infty}\rho_{\beta_0} = 0. \tag{9}$$

3.4 Theoretical Guarantee

We present the generalization bound of our proposed methods in Theorem 3.4. This bound depends on the size of training samples and robust radius. Details of the derivation are illustrated in Appendix A.2.4.

Theorem 3.4. *[Generalization Bound]. Suppose $\ell(\mathbf{s},v) \in [0,M]$. $\mathcal{L}''_{DRO,N}$ stands for DRO loss with N samples. Then for any distribution μ satisfied $D_{KL}(\mu||\mu_0) \leq \rho_{\beta_0}$ with loss denoted as $\mathcal{L}_\mu$, we have that with probability at least $1-\eta$:*

$$\mathcal{L}_\mu \leq \mathcal{L}''_{DRO,N} + \mathcal{B}(\eta,N,\beta_0), \tag{10}$$

where:

$$\mathcal{B}(\eta,N,\beta_0) = \frac{M}{N-1+\exp\left(\frac{M}{\beta_0}\right)}\sqrt{\frac{N\exp\left(\frac{2M}{\beta_0}\right)\log\left(\frac{1}{\eta}\right)}{2}}.$$

Note that in SeqRec scenario, μ denotes the distribution of the unseen testing data. Thus the ideal loss on the testing set is upper bounded by our proposed method considering that $\mathcal{B}(\eta,N,\beta_0)\to 0$ as $N\to\infty$. In addition, this bound is instructive and meaningful in practice: if we increase the robust radius, β_0 would decrease according to Theorem 3.3, and $\mathcal{B}(\eta,N,\beta_0)$ will increase consequently. Under this condition, we need more samples (N to be larger) so that the bound can be tighter.

3.5 Discussion

To conclude, optimizing Equation (8) can make the recommender more robust in the dynamic environment of SeqRec according to our throughout analysis. However, pursuing robustness may restrict the model accuracy. To this end, we propose DROS: Distributionally Robust Optimization mechanism for SeqRec by compromising DRO and ERM, which is quite simple yet very effective in not only considering the dynamics in SeqRec but also preserving the recommendation accuracy.

Finally, we can formally define the objective of DROS as:

$$\mathcal{L}_{DROS} = \alpha\mathcal{L}''_{DRO} + \mathcal{L}_{ERM}, \tag{11}$$

where α is the compromise coefficient, and $\mathcal{L}_{ERM}$ represents the ERM objective function such as MSE, BCE, and BPR (Equation (2)).

It is worth mentioning that although our analysis is conducted in SeqRec, our proposed DROS can be applied to other recommendation seniors such as collaborative filtering and click through rate prediction with minor modifications, since they also suffer from the distribution dynamics issue. We leave this for our future work.

4 EXPERIMENTS

In this section, we conduct experiments to exhibit the superiority of our proposed DROS and reveal the reasons of its effectiveness.

Table 1: Performance comparison of different ERMs (MSE, BEC, BPR) and DROS. ’Avg.Imp.’ denotes the average improvement.

		YooChoose		KuaiRec		RetailRocket		
		HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20	Avg.Imp.
GRU4Rec	MSE	0.0302±0.0009	0.0112±0.0005	0.0224±0.0008	0.0081±0.0002	0.0588±0.0013	0.0230±0.0008	–
	DROS	0.0413±0.0006	0.0170±0.0002	0.0349±0.0015	0.0152±0.0004	0.0880±0.0021	0.0362±0.0010	56.5%
	BCE	0.0389±0.0011	0.0162±0.0002	0.0332±0.0011	0.0123±0.0008	0.0686±0.0014	0.0286±0.0009	–
	DROS	0.0506±0.0004	0.0251±0.0003	0.0437±0.0019	0.0203±0.0006	0.1308±0.0037	0.0613±0.0016	62.6%
	BPR	0.0381±0.0005	0.0173±0.0002	0.0375±0.0007	0.0186±0.0004	0.0556±0.0008	0.0229±0.0005	–
	DROS	0.0465±0.0010	0.0238±0.0006	0.0422±0.0007	0.0201±0.0003	0.0863±0.0014	0.0407±0.0006	35.5%
Caser	MSE	0.0290±0.0004	0.0107±0.0003	0.0252±0.0007	0.0092±0.0003	0.0350±0.0005	0.0134±0.0006	–
	DROS	0.0393±0.0008	0.0163±0.0005	0.0334±0.0006	0.0130±0.0006	0.0657±0.0007	0.0260±0.0008	57.2%
	BCE	0.0406±0.0012	0.0188±0.0009	0.0288±0.0019	0.0107±0.0007	0.0532±0.0009	0.0215±0.0004	–
	DROS	0.0471±0.0005	0.0237±0.0003	0.0384±0.0011	0.0174±0.0007	0.0957±0.0012	0.0445±0.0008	54.1%
	BPR	0.0400±0.0007	0.0182±0.0005	0.0333±0.0006	0.0129±0.0005	0.0443±0.0010	0.0178±0.0007	–
	DROS	0.0466±0.0004	0.0234±0.004	0.0351±0.0009	0.0155±0.0007	0.0831±0.0015	0.0397±0.0009	46.9%
SASRec	MSE	0.0308±0.0008	0.0114±0.0003	0.0269±0.0005	0.0091±0.0002	0.0355±0.0006	0.0132±0.0004	–
	DROS	0.0405±0.0010	0.0166±0.0005	0.0376±0.0004	0.0160±0.0007	0.0759±0.0012	0.0308±0.0007	73.3%
	BCE	0.0368±0.0008	0.0163±0.0002	0.0392±0.0018	0.0153±0.0011	0.0807±0.0022	0.0319±0.0010	–
	DROS	0.0488±0.0007	0.0244±0.0003	0.0448±0.0003	0.0219±0.0007	0.1304±0.0017	0.0597±0.0008	48.1%
	BPR	0.0358±0.0009	0.0163±0.0007	0.0397±0.0014	0.0176±0.0011	0.0572±0.0009	0.0228±0.0011	–
	DROS	0.0479±0.0002	0.0245±0.0001	0.0437±0.0009	0.0209±0.0006	0.1280±0.0019	0.0606±0.0012	67.1%

Table 2: Statistics of datasets.

Dataset	YooChoose	KuaiRec	RetailRocket
#sequences	128,468	92,090	95,865
#items	9,514	7,261	94,130
#interactions	539,436	737,163	539,005

4.1 Experiment Settings

4.1.1 Datasets.

- **YooChoose** dataset comes from RecSys Challenge 2015 [2]. We preserve the purchase sequences for a moderate size of data. Items interacted by less than 5 times are removed to avoid cold-start issue. Sequences whose length is less than 3 are also removed.
- **KuaiRec** [11] dataset is collected from the recommendation logs of a video-sharing mobile app. We also remove items interacted by less than 5 times and sequences with length less than 3.
- **RetailRocket** dataset is collected from a real-world e-commerce website [3]. We leverage the sequences of viewing and keep the data processing the same as YooChoose.

For all datasets, we first sort all sequences in chronological order, and then split the data into training, validation and testing data at the ratio of 8:1:1. Table 2 summarizes the statics of datasets.

4.1.2 Evaluation Protocols. Following previous work [9, 16, 33], we leverage the next-item recommendation scheme and adopt two widely used metrics to evaluate the top-K recommendation quality: hit ratio (HR) and normalized discounted cumulative gain (NDCG).

4.1.3 Implementation Details. We implement all methods with Python 3.7 and PyTorch 1.12.1 in Nvidia GeForce RTX 3090. We would pad the sequence with an additional padding item if the sequence length is less than 10 and preserve the last 10 interacted items as the historical sequence. We use Adam optimizer, the learning rate is tuned as 0.001 and the batch size is set as 256. The embedding dimension is fixed as 64 across all models. We adopt L2 regularization for all models and the coefficient is searched in [1e-3, 1e-4, 1e-5, 1e-6, 1e-7]. For DROS, we search β_0 in the range of [0.5, 1.5] at the step size 0.1, and α in the range of [0.1, 0.5] at the stop size 0.1. The experiments are conducted 5 times and the average and standard deviation are reported.

4.1.4 Backbone models. Since DROS is model-agnostic, we test the performance with the representative sequential recommenders:

- **GRU4Rec** [14] a RNN-based sequential recommender, which leverages GRU to encode users’ interaction sequence.
- **Caser** [34] is a CNN-based sequential recommender. We apply one vertical filter and 16 horizontal filters with heights {2, 3, 4}.
- **SASRec** [16] is an attention-based sequential recommender.

4.2 Improvement over ERM

We compare our proposal with representative ERM frameworks widely used in SeqRec: MSE, BCE, BPR, as mentioned in Equation (2). The results are reported in Table 1.

From Table 1, we can observe that our proposed DROS outperforms all ERM frameworks substantially and consistently, with regard to all sequential recommenders and datasets. Since the datasets are divided chronologically, this notable performance improvement demonstrates that ERM only achieves sub-optimal result in the dynamic environment of SeqRec. Taking this distribution dynamics into account, our proposal successfully enhances the dynamic adaptation ability of sequential recommenders and offers more robust recommendations in inference stage.

4.3 Improvement over Baselines

Some work may have the effect of dynamic adaptation. We consider them as potential baselines using BCE as the ERM framework.

[2] https://recsys.acm.org/recsys15/challenge/

[3] https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset

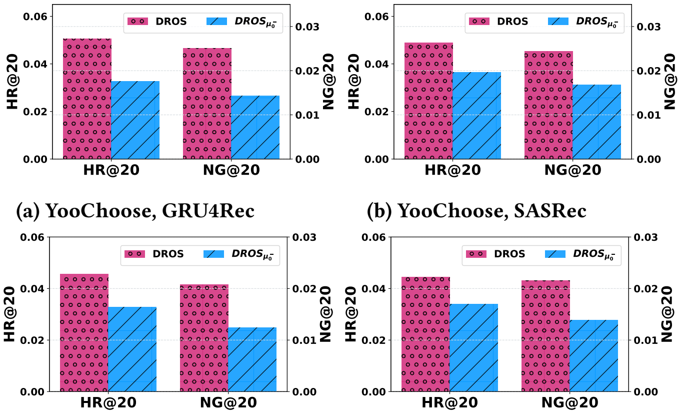
Table 3: Comparison with baselines . Boldface denotes the best performance while underline indicates the second best. 'Imp.' denotes the relative improvement over the second best.

		YooChoose		KuaiRec		RetailRocket	
		HR@20	NDCG@20	HR@20	NDCG@20	HR@20	NDCG@20
GRU4Rec	IPS	0.0390±0.0009	0.0172±0.0002	0.0312±0.0006	0.0112±0.0003	0.0777±0.0009	0.0365±0.0010
	PD	0.0408±0.0007	0.0181±0.0003	0.0336±0.0012	0.0127±0.0002	0.0677±0.0012	0.0285±0.0008
	AdaRanker	0.0350±0.0006	0.0152±0.0005	0.0288±0.0006	0.0093±0.0005	0.0727±0.0007	0.0296±0.0006
	SSM	0.0421±0.0006	0.0180±0.0002	0.0357±0.0007	0.0164±0.0006	0.0978±0.0013	0.0467±0.0003
	CL4SRec	0.0445±0.0008	0.0186±0.0002	0.0370±0.0005	0.0159±0.0004	0.0926±0.0011	0.0397±0.0005
	S-DRO	0.0405±0.0004	0.0180±0.0003	0.0321±0.0005	0.0139±0.0015	0.0852±0.0002	0.0375±0.0005
	GADA	0.0407±0.0005	0.0181±0.0009	0.0303±0.0003	0.0108±0.0005	0.0830±0.0004	0.0356±0.0007
	DROS	**0.0506**±0.0004	**0.0251**±0.0003	**0.0437**±0.0019	**0.0203**±0.0006	**0.1308**±0.0037	**0.0613**±0.0016
	Imp.	13.7%	34.9%	18.1%	23.7%	33.7%	31.2%
Caser	IPS	0.0418±0.0007	0.0194±0.0004	0.0299±0.0005	0.0115±0.0004	0.0642±0.0011	0.0265±0.0006
	PD	0.0427±0.0008	0.0192±0.0005	0.0335±0.0009	0.0124±0.0007	0.0651±0.0008	0.0279±0.0006
	AdaRanker	0.0398±0.0011	0.0181±0.0002	0.0281±0.0006	0.0107±0.0002	0.0577±0.0010	0.0250±0.0008
	SSM	0.0465±0.0007	0.0219±0.0005	0.0345±0.0008	0.0157±0.0003	0.0643±0.0007	0.0336±0.0010
	CL4SRec	0.0431±0.0005	0.0195±0.0009	0.0371±0.0008	0.0162±0.0004	0.0752±0.0007	0.0316±0.0009
	S-DRO	0.0417±0.0009	0.0195±0.0006	0.0307±0.0009	0.0117±0.0007	0.0773±0.0007	0.0332±0.0011
	GADA	0.0427±0.0013	0.0189±0.0009	0.0344±0.0007	0.0127±0.0004	0.0755±0.0011	0.0326±0.0014
	DROS	**0.0477**±0.0005	**0.0237**±0.0003	**0.0384**±0.0011	**0.0174**±0.0007	**0.0957**±0.0012	**0.0445**±0.0008
	Imp.	2.6%	8.2%	3.5%	7.4%	23.8%	34.0%
SASRec	IPS	0.0381±0.0005	0.0173±0.0003	0.0373±0.0003	0.0140±0.0005	0.0920±0.0010	0.0373±0.0009
	PD	0.0411±0.0008	0.0184±0.0005	0.0424±0.0006	0.0187±0.0002	0.0899±0.0014	0.0361±0.0005
	AdaRanker	0.0374±0.0006	0.0167±0.0004	0.0414±0.0009	0.0189±0.0005	0.0824±0.0009	0.0388±0.006
	SSM	0.0445±0.0008	0.0188±0.0002	0.0416±0.0005	0.0177±0.0006	0.0943±0.0008	0.0474±0.0006
	CL4SRec	0.0452±0.0004	0.0193±0.0002	0.0425±0.0010	0.0201±0.0009	0.1155±0.0013	0.0483±0.0007
	S-DRO	0.0401±0.0004	0.0177±0.0004	0.0333±0.0006	0.0135±0.0012	0.0942±0.0025	0.0394±0.0018
	GADA	0.0386±0.0006	0.0172±0.0005	0.0349±0.0004	0.0138±0.0004	0.0964±0.0013	0.0399±0.0006
	DROS	**0.0488**±0.0007	**0.0244**±0.0003	**0.0448**±0.0003	**0.0219**±0.0007	**0.1304**±0.0017	**0.0597**±0.0008
	Imp.	7.9%	26.4%	5.4%	8.9%	12.9%	23.6%

- **IPS** [31, 40]. This is a representative debiasing method for addressing selection bias. It was first developed by [31] and was introduced to sequential recommendation recently by [40].
- **PD** [49]. It considers to decouple the popularity effects and user-item matching scores, and adjusts the influence of popularity bias in the training stage by causal intervention.
- **SSM** [43]. It revises InfoNCE, a classic contrastive learning method in recommendation.
- **CL4SRec** [44]. It uses augmentation methods (item crop, item mask and item reorder) in SeqRec and applies contrastive learning techniques to derive self-supervision signals.
- **AdaRanker** [9]. It develops to adjust model parameters according to given candidate items. Since the datasets in our experiment have no category information, we adopt distribution-mixer sampling to be popularity-based and uniform in the original paper.
- **S − DRO** [41]. It applies group-DRO in SeqRec to improve fairness of different user groups.
- **GADA** [35]. It adopts adversarial data augmentation in domain generalization of computer vision tasks.

The results are shown in Table 3, from which we can observe:

- In general, our proposal performs best on all datasets in terms of all metrics. This performance improvement can be attributed to the superiority of DROS, which includes the dynamics of data in the training process, instead of statically considering the previous interactions in the training data.
- As debiasing strategies, IPS and PD target at alleviating the side effect of exposure bias and popularity bias respectively. Both of them can improve the performance in certain cases. But they perform worse than DROS, so considering particular bias is inadequate to resolve the dynamic challenge in SeqRec.
- AdaRanker assumes that distribution of the future testing data is similar with distribution of the candidate items. Different from AdaRanker, DROS also promotes the dynamic adaptation of sequential recommenders by approaching the testing distribution with a proper robust radius, which is the reason why DROS performs better than AdaRanker.
- SSM revises InfoNCE in recommendation, and CL4SRec leverages three carefully-designed data augmentation techniques in SeqRec. They also boost the performance of backbone recommenders, but fail to outperform DROS (with an exception in KuaiRec dataset with Caser as the backbone recommender). This result indicates that ignoring the data distribution dynamics can only achieve sub-optimal performance.
- S-DRO is designed for improving fairness of SeqRec, so the overall performance is not as good as DROS. In terms of GADA, it demands distance of samples, which faces obstacles in SeqRec



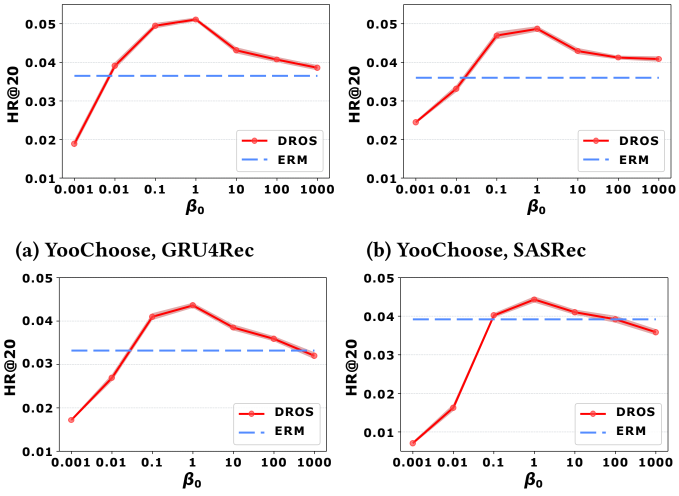
(a) YooChoose, GRU4Rec

(b) YooChoose, SASRec

(c) KuaiRec, GRU4Rec

(d) KuaiRec, SASRec

Figure 3: Ablation study of nominal distribution.



(a) YooChoose, GRU4Rec

(b) YooChoose, SASRec

(c) KuaiRec, GRU4Rec

(d) KuaiRec, SASRec

Figure 4: Ablation study of robust radius. Note that robust radius decreases monotonically *w.r.t.* β in DROS.

with samples as sequence-item pairs instead of images. We modify it by adding two embedding distances, but the unsatisfactory performance implies more effort before it can benefit SeqRec.

4.4 Ablation Study of DROS

There exist several crucial designs in our proposal: 1) nominal distribution μ_0; and 2) robust radius ρ. To verify the impact of each design, we conduct ablation experiments by replacing it with some variant or changing its values.

4.4.1 Nominal Distribution. Generally, nominal distribution μ_0 should lay around the distribution of the testing data [19]. Since we have no access to testing data in practice, we estimate the nominal distribution with item frequency in previous interactions under the observation that dynamics in SeqRec is continuous. To demonstrate the rationality of this estimation, we in contrast use the inverse frequency score to estimate the nominal distribution, *i.e.,* items with higher frequency in previous interactions have lower probability in the nominal distribution. We denote this variant as $DROS_{u_0^-}$.

We illustrate the comparison in Figure 3. It is apparent that taking the inverse frequency score as the estimation of nominal distribution results in worse performance. The dramatic decline of performance indicates that this ill-designed nominal distribution lays too far away from the distribution of the testing data, such that DROS fails to generalize to testing data from this nominal distribution. Therefore it is necessary to estimate a reasonable nominal distribution.

4.4.2 Robust Radius. Robust radius ρ is crucial in DRO. As discussed in Lemma 3.3, ρ decreases monotonically *w.r.t.* β in DROS, thus we conduct experiments to illustrate the impact of robust radius by varying β. The result is shown in Figure 4.

We can observe that the performance change *w.r.t.* β well justifies our analysis of robust radius: 1) As the value of β increases, robust radius ρ would decrease according to Lemma 3.3. Under this circumstance, the performance reduces to be similar to basic sequential recommender since DRO approaches ERM; 2) As the value of β decreases, robust radius ρ would increase. In this case, the performance collapses as shown in Figure 4, which is because the uncertainty set contains too much redundant distributions beyond distribution of testing set if robust radius is too large, and DROS is prone to reaching a pathological distribution.

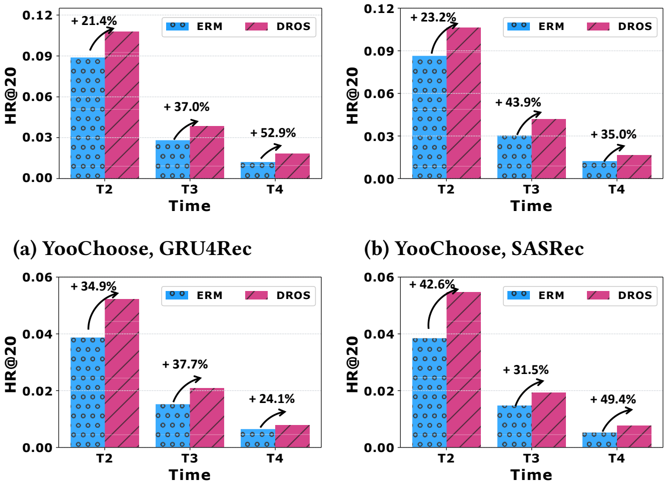
4.5 Benefits of DROS

In this section, we explore whether our proposed DROS can enhance the dynamic adaptation ability of sequential recommenders. Similarly to Figure 1, we first divide dataset into four shards in chronological order, such that distributions of latter shards of data differ more from the first shard of data. Then we train the recommender on the first shard of data and evaluate on the rest. The results are illustrated in Figure 5.

We can observe that as time passes by, the performance of recommenders declines sharply due to the data distribution is shifting over time. Relying on ERM framework, traditional sequential recommenders tend to overfit historical data, which results in poor robustness in future stages. Note that DROS can promote the performance of base sequential recommenders (GRU4Rec and SASRec) consistently in latter stages, which indicates the effectiveness of DROS in enhancing the dynamic adaptation of recommenders. Since the rationality of each components has been verified, we can safely attribute the performance improvement to that DROS can consider date distribution dynamics with the carefully-designed distribution adaptation paradigm.

5 CONCLUSION

In this work, we study how to enhance the dynamic adaptation ability of sequential recommenders. We develop a generic framework equipped with SeqRec-oriented DRO mechanism. We first estimate the nominal distribution from historical interactions, and promote recommenders to approach future distributions with a proper robust radius. Theoretical analysis guarantees the robustness of DROS against distribution shifts in the dynamic recommendation environment. Experimental results further confirm that DROS outperforms baselines effectively and consistently.



(a) YooChoose, GRU4Rec **(b) YooChoose, SASRec**

(c) KuaiRec, GRU4Rec **(d) KuaiRec, SASRec**

Figure 5: Benefits of DROS in enhancing dynamic adaptation ability of sequential recommenders.

This work reveals the limitation of ERM in SeqRec, in spite of its dominant role in both research and industry of recommendation. We propose a simple yet effective framework, DROS, to deal with the natural distribution shifts in SeqRec. We believe it is crucial to put more effort in the dynamic nature of online serving in recommendation, and DRO provides a power foundation in the process. We will work on achieving the full promise of DRO in SeqRec based on our study of DROS. Moreover, we are also interested in verifying the effectiveness of DROS in related tasks such as collaborative filtering and click through rate prediction.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2021ZD0111802), the National Natural Science Foundation of China (9227010114, U21B2026), the University Synergy Innovation Program of Anhui Province (GXXT-2022-040), and the CCCD Key Lab of Ministry of Culture and Tourism.

A APPENDIX

A.1 Lemma

LEMMA A.1. *(Holder's inequality [21]) For random variable X and functions $g(\cdot)$ and $f(\cdot)$, we have that for any $p > 0$ and $q = \frac{p}{1-p}$.:*

$$\mathbb{E}_X[|g(X)f(X)|] \leq \mathbb{E}_X[|g(X)|^p]^{\frac{1}{p}}\mathbb{E}_X[|f(X)|^q]^{\frac{1}{q}}. \tag{12}$$

LEMMA A.2. *(McDiarmid's inequality [22]) Let $X_1, ..., X_N \in \mathcal{X}^N$ be a set of $N \geq 1$ independent random variables and assume that there exists $c_1, ..., c_N > 0$ such that $f : \mathcal{X}^N \rightarrow \mathbb{R}$ satisfies:*

$$|f(x_1, ..., x_i, ..., x_N) - f(x_1, ..., x_i', ..., x_N)| \leq c_i. \tag{13}$$

For all $i \in 1, 2, ...N$ and any points $x_1, ...x_N, x_i' \in \mathcal{X}$. Let $f(S)$ denote $f(X_1, ..., X_N)$, then for all $\epsilon > 0$, the following inequalities hold:

$$\mathbb{P}[f(S) - \mathbb{E}\{f(S)\} \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^N c_i^2}\right). \tag{14}$$

A.2 Proof of Theorems

A.2.1 Proof of Theorem 3.1.

PROOF. We follow the deviation from [15]. Set D as KL-divergence, the inner maximization problem can be formulated as:

$$\begin{aligned} &\max_{\mu} \ \mathbb{E}_{(\mathbf{s},v)\sim\mu} \ [\ell(\mathbf{s},v)] \\ &\text{subject to } \ D_{KL}(\mu||\mu_0) \leq \rho. \end{aligned} \tag{15}$$

Let $\eta(\mathbf{s},v) = \mu(\mathbf{s},v)/\mu_0(\mathbf{s},v)$, we have:

$$\begin{aligned} D_{KL}(\mu||\mu_0) &= \iint \mu(\mathbf{s},v)\log\frac{\mu(\mathbf{s},v)}{\mu_0(\mathbf{s},v)}dsdv \\ &= \mathbb{E}_{(\mathbf{s},v)\sim\mu_0}[\eta(\mathbf{s},v)\log\eta(\mathbf{s},v)]. \end{aligned} \tag{16}$$

$$\mathbb{E}_{(\mathbf{s},v)\sim\mu} \ [\ell(\mathbf{s},v)] = \mathbb{E}_{(\mathbf{s},v)\sim\mu_0}[\ell(\mathbf{s},v)\eta(\mathbf{s},v)]. \tag{17}$$

Therefore Equation (15) can be reformulated as:

$$\begin{aligned} &\max_{\eta} \ \mathbb{E}_{(\mathbf{s},v)\sim\mu_0} \ [\ell(\mathbf{s},v)\eta] \\ &\text{subject to } \ \mathbb{E}_{(\mathbf{s},v)\sim\mu_0}[\eta\log\eta] \leq \rho, \mathbb{E}_{(\mathbf{s},v)\sim\mu_0}[\eta] = 1. \end{aligned} \tag{18}$$

Its Lagrangian function is:

$$l_0(\beta,\gamma,\eta,\rho) = \mathbb{E}_{\mu_0}[\ell(\mathbf{s},v)\eta] - \beta(\mathbb{E}_{(\mu_0}[\eta\log\eta] - \rho) - \gamma(\mathbb{E}_{\mu_0}[\eta] - 1). \tag{19}$$

We can obtain the corresponding Lagrangian dual of Problem (18):

$$\min_{\beta\geq 0, \ \gamma\geq 0} \ \max_{\eta} \ l_0^*(\beta,\gamma,\eta,\rho). \tag{20}$$

l_0^* is the close form of Problem (18) by solving Problem (20). Omitting the term $\beta\rho$ and γ, then define the functional:

$$\mathcal{F}(\eta(\mathbf{s},v)) := \mathbb{E}_{(\mathbf{s},v)\sim\mu_0}[\ell(\mathbf{s},v)\eta - \beta\eta\log\eta - \gamma\eta]. \tag{21}$$

Note $\mathcal{F}(\eta(\mathbf{s},v))$ is convex in η. For any feasible direction $u = u(\mathbf{s},v)$, we can calculate the derivative of the functional:

$$\begin{aligned} &\mathrm{D}\mathcal{F}(\eta(\mathbf{s},v))u = \lim_{t\rightarrow 0}(\mathcal{F}(\eta(\mathbf{s},v) + tu(\mathbf{s},v)) - \mathcal{F}(\eta(\mathbf{s},v)))/t \\ &= \mathbb{E}_{\mu_0}[\ell u] - \gamma\mathbb{E}_{\mu_0}[u] - \beta\lim_{t\rightarrow 0}\mathbb{E}_{\mu_0}((\eta + tu)\log(\eta + tu) - \eta\log\eta)/t. \end{aligned} \tag{22}$$

By the Monotone Convergence Theorem, we can interchange the order of *limitation* and *expectation* in Equation (22) and get:

$$\mathrm{D}\mathcal{F}(\eta(\mathbf{s},v))u = \mathbb{E}_{\mu_0}[(\ell - \beta(\log\eta + 1) - \gamma)u]. \tag{23}$$

To acquire $\eta^*(\mathbf{s},v)$, let $\mathrm{D}\mathcal{F}(\eta(\mathbf{s},v))u = 0$ for all feasible direction $u(\mathbf{s},v)$, we can have:

$$\ell(\mathbf{s},v) - \beta(\log\eta + 1) - \gamma = 0. \tag{24}$$

Solving the equation, we get:

$$\eta^*(\mathbf{s},v) = \exp\left(\ell(\mathbf{s},v)/\beta - (\gamma + \beta)/\beta\right). \tag{25}$$

From Equation (18) we have $\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}[\eta^*] = 1$, then:

$$\gamma^* = \beta\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta\right) - \beta. \tag{26}$$

Then $\eta^*(\mathbf{s},v)$ can be further reformulated as:

$$\eta^*(\mathbf{s},v) = \exp\left(\ell(\mathbf{s},v)/\beta\right)/\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta\right). \tag{27}$$

Put η^* in Lagrangian function Equation (19), we can obtain:

$$l_0^*(\beta,\gamma,\eta,\rho) = \inf_{\beta\geq 0}\left[\beta\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta\right) + \beta\rho\right]. \tag{28}$$

Thus, we get the desired form of the theorem:

$$\mathcal{L}_{DRO}' = \inf_{\beta\geq 0}\left[\beta\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta}\right) + \beta\rho\right]. \tag{29}$$

□

A.2.2 *Proof of Theorem 3.2.*

Proof. Let θ denote the learn-able parameters in the neural network. Since $\beta_0\rho$ has no gradient to θ, thus for $\forall\rho$ we can get:

$$\begin{aligned}\nabla_\theta\mathcal{L}''_{DRO} &= \nabla_\theta\left[\beta_0\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta_0\right)\right]\\ &= \nabla_\theta\underbrace{\left[\beta_0\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta_0\right)+\beta_0\rho\right]}_{\mathcal{L}^U_{DRO}},\end{aligned}\tag{30}$$

this shows using $\mathcal{L}''_{DRO}$ is equal to use $\mathcal{L}^U_{DRO}$ while training. Interestingly, $\mathcal{L}^U_{DRO}$ is a upper bound for all possible robust radius ρ. For $\forall\rho$, we have:

$$\begin{aligned}\mathcal{L}^U_{DRO} &= \beta_0\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta_0\right)+\beta_0\rho\\ &\geq \inf_{\beta>0}\left[\beta\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta\right)+\beta\rho\right]\\ &= \mathcal{L}'_{DRO},\end{aligned}\tag{31}$$

Then we will show there exists a ρ_{β_0} to let the bound to be tight. Since we choose $\beta=\beta_0$ as a constant, the ρ_{β_0} must let the Equation (32) exists:

$$\beta_0 = \arg\min_\beta[\beta\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\ell(\mathbf{s},v)/\beta\right)+\beta\rho_{\beta_0}],\tag{32}$$

Thus, we have:

$$\begin{aligned}&\partial_\beta\mathcal{L}'_{DRO|\beta=\beta_0}\\ &= \log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right)-\frac{\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\ell(\mathbf{s},v)\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right)}{\beta_0\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right)}+\rho_{\beta_0}=0.\end{aligned}\tag{33}$$

By solving Equation (33), we get:

$$\rho_{\beta_0} = -\log\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right)+\frac{\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\ell(\mathbf{s},v)\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right)}{\beta_0\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}\exp\left(\frac{\ell(\mathbf{s},v)}{\beta_0}\right)}.\tag{34}$$

□

A.2.3 *Proof of Theorem3.3.*

Proof. In this deviation, we use $\mathbb{E}$ to denote $\mathbb{E}_{(\mathbf{s},v)\sim\mu_0}$ and ℓ to denote $\ell(\mathbf{s},v)$. we first show ρ_{β_0} decreases monotonically with β_0:

$$\partial_{\beta_0}\rho_{\beta_0} = \frac{\left\{\left\{\mathbb{E}\left[\ell\exp\left(\frac{\ell}{\beta_0}\right)\right]\right\}^2-\mathbb{E}\left[\ell^2\exp\left(\frac{\ell}{\beta_0}\right)\right]\mathbb{E}\left[\exp\left(\frac{\ell}{\beta_0}\right)\right]\right\}}{\beta_0^3\left\{\mathbb{E}\left[\exp\left(\frac{\ell}{\beta_0}\right)\right]\right\}^2}.\tag{35}$$

By using Lemma A.1, we get:

$$\mathbb{E}\left[\ell\exp\left(\frac{\ell}{\beta_0}\right)\right] \leq \left\{\mathbb{E}\left[\left(\ell\exp\left(\frac{\ell}{2\beta_0}\right)\right)^2\right]\mathbb{E}\left[\left(\exp\left(\frac{\ell}{2\beta_0}\right)\right)^2\right]\right\}^{\frac{1}{2}}.\tag{36}$$

Since $\beta_0^3\left\{\mathbb{E}\left[\exp\left(\frac{\ell}{\beta_0}\right)\right]\right\}^2 \geq 0$, we get $\partial_{\beta_0}\rho_{\beta_0} \leq 0$ for all β_0. This means that ρ_{β_0} decreases monotonically with β_0. To be specifically, when $\beta_0 \to \infty$, we have:

$$\lim_{\beta_0\to\infty}\left|\rho_{\beta_0}\right| \leq \lim_{\beta_0\to\infty}\left|-\log\mathbb{E}\left[\exp\left(\frac{\ell}{\beta_0}\right)\right]\right|+\lim_{\beta_0\to\infty}\left|\frac{\mathbb{E}\left[\ell\exp\left(\frac{\ell}{\beta_0}\right)\right]}{\beta_0\mathbb{E}\left[\exp\left(\frac{\ell}{\beta_0}\right)\right]}\right|.\tag{37}$$

The limit is constructed by two parts, the first part:

$$\lim_{\beta_0\to\infty}\left|-\log\mathbb{E}\left[\exp\frac{\ell}{\beta_0}\right]\right| \leq \log\mathbb{E}\left[\lim_{\beta_0\to\infty}\left|\exp\left(\frac{\ell}{\beta_0}\right)\right|\right] = \log\mathbb{E}1 = 0.\tag{38}$$

Before get the limit of the second part, we first get:

$$\left|\frac{\mathbb{E}\left[\ell\exp\left(\ell/\beta_0\right)\right]}{\mathbb{E}[\exp\left(\ell/\beta_0\right)]}\right| = \left|\mathbb{E}\left[\ell\frac{\exp\left(\ell/\beta_0\right)}{\mathbb{E}[\exp\left(\ell/\beta_0\right)]}\right]\right| \leq \max_{(\mathbf{s},v)\sim\mu_0}\ell.\tag{39}$$

Thus, the second part can be derived as:

$$\lim_{\beta_0\to\infty}\left|\frac{\mathbb{E}\left[\ell\exp\left(\ell/\beta_0\right)\right]}{\beta_0\mathbb{E}\left[\exp\left(\ell/\beta_0\right)\right]}\right| \leq \lim_{\beta_0\to\infty}\frac{\max_{(\mathbf{s},v)\sim\mu_0}\ell}{\beta_0} = 0.\tag{40}$$

Finally, we get $\lim_{\beta_0\to\infty}|\rho_{\beta_0}|=0$, this means $\lim_{\beta_0\to\infty}\rho_{\beta_0}=0$. □

A.2.4 *Proof of Theorem 3.4.*

Proof. For any μ satisfying $D_{KL}(\mu||\mu_0) \leq \rho_{\beta_0}$, we have:

$$\mathcal{L}_\mu = \mathbb{E}_{(\mathbf{s},v)\sim\mu}\ell(\mathbf{s},v) \leq \max_{D_{KL}(\mu||\mu_0)\leq\rho_{\beta_0}}\mathbb{E}_{(\mathbf{s},v)\sim\mu}\ell(\mathbf{s},v) = \mathcal{L}'_{DRO}.\tag{41}$$

Using Theorem 3.2, we analyze $\mathcal{L}''_{DRO}$ instead of $\mathcal{L}'_{DRO}$ since they are equal when robust radius is ρ_{β_0}.

Let $w_{(\mathbf{s},v)} = \frac{\exp(\ell(\mathbf{s},v)/\beta_0)}{\sum_v\ell(\exp(\ell(\mathbf{s},v)/\beta_0))}$ denote weight for DRO re-weighting. We have:

$$|w_{(\mathbf{s},v)}\ell(\mathbf{s},v)-w'_{(\mathbf{s},v)}\ell'(\mathbf{s},v)| \leq \sup_{(\mathbf{s},v)\sim\mu_0}|w_{(\mathbf{s},v)}\ell(\mathbf{s},v)| \leq \frac{\exp\left(\frac{M}{\beta_0}\right)M}{N-1+\exp\left(\frac{M}{\beta_0}\right)},\tag{42}$$

for all $\mathbf{s},v$ exist. By using McDiarmid's inequality in Lemma A.2, for any $\epsilon > 0$, there exists:

$$\mathbb{P}[\mathcal{L}''_{DRO}-\mathcal{L}''_{DRO,N} \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2\left(N-1+\exp\left(\frac{M}{\beta_0}\right)\right)^2}{N\exp\left(\frac{2M}{\beta_0}\right)M^2}\right),\tag{43}$$

let

$$\eta = \exp\left(\frac{-2\epsilon^2\left(N-1+\exp\left(\frac{M}{\beta_0}\right)\right)^2}{N\exp\left(\frac{2M}{\beta_0}\right)M^2}\right),\tag{44}$$

we get:

$$\epsilon = \frac{M}{N-1+\exp\left(\frac{M}{\beta_0}\right)}\sqrt{\frac{N\exp\left(\frac{2M}{\beta_0}\right)\log\left(\frac{1}{\eta}\right)}{2}}.\tag{45}$$

Thus, for $\forall\eta \in (0,1)$, we get that with probability at least $1-\eta$:

$$\mathcal{L}_\mu \leq \mathcal{L}''_{DRO} \leq \mathcal{L}''_{DRO,N}+\frac{M}{N-1+\exp\left(\frac{M}{\beta_0}\right)}\sqrt{\frac{N\exp\left(\frac{2M}{\beta_0}\right)\log\left(\frac{1}{\eta}\right)}{2}}.\tag{46}$$

□

REFERENCES

[1] Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust Solutions of Optimization Problems Affected by Uncertain Probabilities. *Manag. Sci.* 59 (2013), 341–357.

[2] Pedro Dalla Vecchia Chaves, Bruno L. Pereira, and Rodrygo L. T. Santos. 2022. Efficient Online Learning to Rank for Sequential Music Recommendation. In *WWW*. 2442–2450.

[3] Chong Chen, Min Zhang, Yongfeng Zhang, Yiqun Liu, and Shaoping Ma. 2020. Efficient Neural Matrix Factorization without Sampling for Recommendation. *TOIS* 38 (2020), 14:1–14:28.

[4] Yongjun Chen, Zhiwei Liu, Jia Li, Julian J. McAuley, and Caiming Xiong. 2022. Intent Contrastive Learning for Sequential Recommendation. In *WWW*. 2172–2182.

[5] Chen Cheng, Haiqin Yang, Michael R Lyu, and Irwin King. 2013. Where you like to go next: Successive point-of-interest recommendation. In *IJCAI*. 2605–2611.

[6] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *RecSys*. 191–198.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[8] Sihao Ding, Fuli Feng, Xiangnan He, Jinqiu Jin, Wenjie Wang, Yong Liao, and Yongdong Zhang. 2022. Interpolative Distillation for Unifying Biased and Debiased Recommendation. In *SIGIR*. 40–49.

[9] Xinyan Fan, Jianxun Lian, Wayne Xin Zhao, Zheng Liu, Chaozhuo Li, and Xing Xie. 2022. Ada-Ranker: A Data Distribution Adaptive Ranking Paradigm for Sequential Recommendation. In *SIGIR*. 1599–1610.

[10] Xinyan Fan, Zheng Liu, Jianxun Lian, Wayne Xin Zhao, Xing Xie, and Ji-Rong Wen. 2021. Lighter and Better: Low-Rank Decomposed Self-Attention Networks for Next-Item Recommendation. In *SIGIR*. 1733–1737.

[11] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In *CIKM*.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. 2022. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*. 15979–15988.

[13] Ruining He and Julian McAuley. 2016. Fusing similarity models with markov chains for sparse sequential recommendation. In *ICDM*. 191–200.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based Recommendations with Recurrent Neural Networks. In *ICLR*.

[15] Zhaolin Hu and L Jeff Hong. 2013. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online* (2013), 1695–1724.

[16] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *ICDM*. 197–206.

[17] David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron C. Courville. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *ICML*, Vol. 139. 5815–5826.

[18] Henry Lam. 2016. Robust Sensitivity Analysis for Stochastic Systems. *Math. Oper. Res.* 41 (2016), 1248–1275.

[19] Fengming Lin, Xiaolei Fang, and Zheming Gao. 2022. Distributionally Robust Optimization: A review on theory and applications. *Numerical Algebra, Control & Optimization* 12 (2022), 159.

[20] Divyat Mahajan, Shruti Tople, and Amit Sharma. 2021. Domain generalization using causal matching. In *ICML*. 7313–7324.

[21] Lech Maligranda. 1998. Why Hölder's inequality should be called Rogers' inequality. *Mathematical Inequalities & Applications* 1 (1998), 69–83.

[22] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press.

[23] Yabo Ni, Dan Ou, Shichen Liu, Xiang Li, Wenwu Ou, Anxiang Zeng, and Luo Si. 2018. Perceive Your Users in Depth: Learning Universal User Representations from Multiple E-commerce Tasks. In *SIGKDD*. 596–605.

[24] Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. 2019. Distributionally Robust Language Modeling. In *EMNLP*. 4226–4236.

[25] Aleksandr Petrov and Craig Macdonald. 2022. Effective and Efficient Training for Sequential Recommendation using Recency Sampling. In *RecSys*. ACM, 81–91.

[26] Ruihong Qiu, Zi Huang, Hongzhi Yin, and Zijian Wang. 2022. Contrastive Learning for Representation Degeneration Problem in Sequential Recommendation. In *WSDM*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). 813–823.

[27] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. 452–461.

[28] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *WWW*. 811–820.

[29] Alexander Robey, George J Pappas, and Hamed Hassani. 2021. Model-based domain generalization. *NeurIPS* 34 (2021), 20210–20229.

[30] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *ICLR*.

[31] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In *ICML*. JMLR.org, 1670–1679.

[32] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *CIKM*. 1441–1450.

[33] Qiaoyu Tan, Jianwei Zhang, Jiangchao Yao, Ninghao Liu, Jingren Zhou, Hongxia Yang, and Xia Hu. 2021. Sparse-Interest Network for Sequential Recommendation. In *WSDM*. 598–606.

[34] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *WSDM*. 565–573.

[35] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. 2018. Generalizing to Unseen Domains via Adversarial Data Augmentation. In *NeurIPS*. 5339–5349.

[36] Qi Wan, Xiangnan He, Xiang Wang, Jiancan Wu, Wei Guo, and Ruiming Tang. 2022. Cross Pairwise Ranking for Unbiased Item Recommendation. In *WWW*. 2370–2378.

[37] Xuejian Wang, Lantao Yu, Kan Ren, Guanyu Tao, Weinan Zhang, Yong Yu, and Jun Wang. 2017. Dynamic Attention Deep Model for Article Recommendation by Learning Human Editors' Demonstration. In *SIGKDD*. 2051–2059.

[38] Zimu Wang, Yue He, Jiashuo Liu, Wenchao Zou, Philip S. Yu, and Peng Cui. 2022. Invariant Preference Learning for General Debiasing in Recommendation. In *SIGKDD*. 1969–1978.

[39] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. 2021. Learning to Diversify for Single Domain Generalization. In *ICCV*. 814–823.

[40] Zhenlei Wang, Shiqi Shen, Zhipeng Wang, Bo Chen, Xu Chen, and Ji-Rong Wen. 2022. Unbiased Sequential Recommendation with Latent Confounders. In *WWW*. 2195–2204.

[41] Hongyi Wen, Xinyang Yi, Tiansheng Yao, Jiaxi Tang, Lichan Hong, and Ed H. Chi. 2022. Distributionally-robust Recommendations for Improving Worst-case User Experience. In *WWW*. 3606–3610.

[42] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. In *SIGIR*. ACM, 726–735.

[43] Jiancan Wu, Xiang Wang, Xingyu Gao, Jiawei Chen, Hongcheng Fu, Tianyu Qiu, and Xiangnan He. 2022. On the Effectiveness of Sampled Softmax Loss for Item Recommendation. *CoRR* abs/2201.02327 (2022).

[44] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive Learning for Sequential Recommendation. In *ICDE*. 1259–1273.

[45] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-Efficient Transfer from Sequential Behaviors for User Modeling and Recommendation. In *SIGIR*. 1469–1478.

[46] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M. Jose, and Xiangnan He. 2019. A Simple Convolutional Generative Network for Next Item Recommendation. In *WSDM*. 582–590.

[47] An Zhang, Wenchang Ma, Xiang Wang, and Tat seng Chua. 2022. Incorporating Bias-aware Margins into Contrastive Loss for Collaborative Filtering. In *NeurIPS*.

[48] An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat seng Chua. 2023. Invariant Collaborative Filtering to Popularity Distribution Shift. In *WWW*.

[49] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *SIGIR*. 11–20.

[50] Long Zhao, Ting Liu, Xi Peng, and Dimitris N. Metaxas. 2020. Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Robustness. In *NeurIPS*.

[51] Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. 2020. Domain Generalization via Entropy Regularization. In *NeurIPS*.

[52] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *WWW*. 2980–2991.

[53] Guorui Zhou, Xiaoqiang Zhu, Chengru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep Interest Network for Click-Through Rate Prediction. In *SIGKDD*. 1059–1068.