

Temporally and Distributionally Robust Optimization for Cold-start Recommendation

Xinyu Lin¹, Wenjie Wang^{1*}, Jujia Zhao¹, Yongqi Li², Fuli Feng^{3,4}, Tat-Seng Chua¹

¹National University of Singapore

²The Hong Kong Polytechnic University

³MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition

⁴University of Science and Technology of China

{xylin1028, wenjiawang96, zhao.jujia.0913, liyongqi0, fulifeng93}@gmail.com, dcscts@nus.edu.sg

Abstract

Collaborative Filtering (CF) recommender models highly depend on user-item interactions to learn CF representations, thus falling short of recommending cold-start items. To address this issue, prior studies mainly introduce item features (*e.g.*, thumbnails) for cold-start item recommendation. They learn a feature extractor on warm-start items to align feature representations with interactions, and then leverage the feature extractor to extract the feature representations of cold-start items for interaction prediction. Unfortunately, the features of cold-start items, especially the popular ones, tend to diverge from those of warm-start ones due to temporal feature shifts, preventing the feature extractor from accurately learning feature representations of cold-start items.

To alleviate the impact of temporal feature shifts, we consider using Distributionally Robust Optimization (DRO) to enhance the generation ability of the feature extractor. Nonetheless, existing DRO methods face an inconsistency issue: the worse-case warm-start items emphasized during DRO training might not align well with the cold-start item distribution. To capture the temporal feature shifts and combat this inconsistency issue, we propose a novel temporal DRO with new optimization objectives, namely, 1) to integrate a worst-case factor to improve the worst-case performance, and 2) to devise a shifting factor to capture the shifting trend of item features and enhance the optimization of the potentially popular groups in cold-start items. Substantial experiments on three real-world datasets validate the superiority of our temporal DRO in enhancing the generalization ability of cold-start recommender models.

1 Introduction

Recommender systems are widely deployed to filter the overloaded multimedia information on the web for meeting users' personalized information needs (He et al. 2017). Technically speaking, Collaborative Filtering (CF) is the

most representative method (Koren, Bell, and Volinsky 2009). In essence, CF methods learn the CF representations of users and items from historical interactions and utilize the learned CF representations to predict the users' future interactions. As content production capabilities continue to advance, recommender systems face the challenge of accommodating an increasing influx of new items (*a.k.a.* cold-start items¹). For example, 500 hours of video are uploaded to YouTube every minute². Since the new items lack historical interactions and thereby have no CF representations, traditional CF methods fail to effectively recommend these cold items to users, disrupting the ecological balance of recommender systems on the item side. In light of this, it is essential to improve the cold-start item recommendation.

Prior literature has integrated item features, such as categories and thumbnails of micro-videos, for cold-start item recommendation (Shalaby et al. 2022; Zhao et al. 2022). These methods essentially learn a feature extractor that encodes warm items (*i.e.*, items in the training set) into feature representations and utilizes feature representations to fit the user-item interactions during training. For inference for cold items, given the lack of CF counterparts, only feature representations from the feature extractor are used to estimate user preference. The key of this paradigm lies in devising training strategies to align feature representations and user-item interactions, which mainly fall into two research lines. 1) Robust training-based methods (Volkovs, Yu, and Poutanen 2017; Du et al. 2020) use both feature representations and CF representations to predict interactions while CF representations are randomly corrupted to strengthen the alignment. 2) Auxiliary loss-based methods (Zhu et al. 2020) pay attention to minimizing the distance between the feature representations and CF representations learned from interactions via the auxiliary loss, *e.g.*, contrastive loss (Wei et al. 2021) and GAN loss (Chen et al. 2022).

Despite their success, existing methods suffer from a critical issue: item features temporally shift from warm to cold

*Corresponding author. This work is supported by the National Key Research and Development Program of China (2022YFB3104701), the National Natural Science Foundation of China (62272437), and Huawei International Pte Ltd. Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹For simplicity, cold-start items and warm-start items are referred to as cold and warm items, respectively.

²<https://www.statista.com/>.

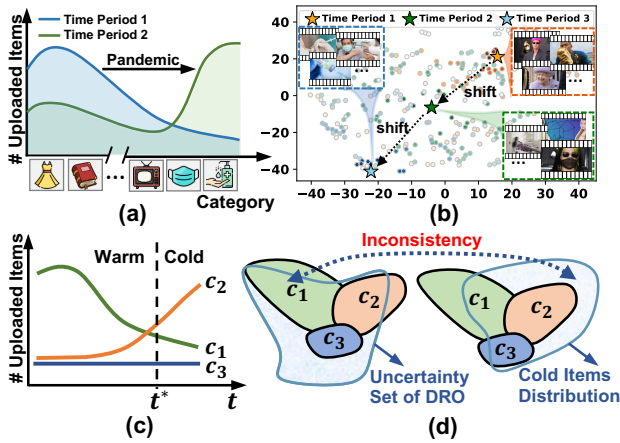


Figure 1: (a) An example of item category feature shifts towards sanitary products. (b) T-SNE visualization of visual features of item thumbnails in three time periods on a Micro-video dataset. The stars represent the average item features in each time period. (c) An example of the shifting trend of three item groups over time. (d) Illustration of the inconsistency issue of DRO.

items. As illustrated in Figure 1(a), the category features of newly-uploaded items are shifting over time due to various environmental factors, such as a pandemic outbreak. Empirical evidence from a real-world Micro-video dataset further substantiates this phenomenon. In Figure 1(b), we divide the micro-videos into three time periods according to the upload time and visualize the micro-video features, where a star represents the average item features in each time period. The moving stars across time periods validate that item features are gradually shifting over time. Since the feature extractor is typically trained on warm items using Empirical Risk Minimization (ERM) (Vapnik 1991), it easily overfits the majority group of warm items. Unfortunately, the majority group of cold items could deviate from that of warm items as depicted in Figure 1(a) and (b). Such temporal feature shifts hinder the feature extractor’s ability to accurately extract feature representations for cold items, thus degrading the performance of cold-start item recommendation. To tackle this issue, we consider learning a feature extractor with robust generalization ability to enhance the interaction prediction on temporally shifted cold items.

To strengthen the generalization ability, Distributionally Robust Optimization (DRO) is a promising approach³. In general, DRO aims to enhance the worst-case performance over the pre-defined uncertainty set, *i.e.*, potential shifted distributions (Duchi and Namkoong 2018). However, directly applying DRO in cold-start recommendation suffers from the inconsistency issue. DRO will overemphasize the minority groups⁴ in warm items at the expense of other groups’ performance (Oren et al. 2019). Due to the fact

³Other potential solutions are discussed in Section 5.

⁴Minority group usually yields worse performance in recommendation (Wen et al. 2022). In DRO, the training distribution is

that minority groups in warm items may not guarantee their popularity in subsequent cold items, the overemphasis on the minority group of warm items might compromise the performance of the popular groups in cold items. For example, in Figure 1(c), c_1 , c_2 , and c_3 denote three item groups, where c_3 is the minority group in the warm items that traditional DRO pays special attention to. However, c_2 is gradually becoming popular, dominating the cold items. The inconsistency between the excessive emphasis on c_3 and the shifting trend towards c_2 prevents DRO from alleviating the impact of temporal feature shifts (see Figure 1(d)). To address this inconsistency issue and strengthen the generalization ability of the feature extractor under the temporal feature shifts, we put forth two objectives for DRO training: 1) enhancing the worst-case optimization on the minority group of warm items, thereby raising the lower bound of performance; and 2) capturing the shifting trend of item features and emphasizing the optimization of the groups likely to become popular.

To this end, we propose a **Temporal DRO (TDRO)**, which considers the temporal shifting trend of item features for cold-start recommendation. In particular, we consider two factors for the training of TDRO: 1) *a worst-case factor* to guarantee worst-case performance, where we divide the warm items into groups by the similarity of item features, and prioritize the improvements of the item groups with large training loss; and 2) *a shifting factor* to capture the shifting trend of item features, which utilizes a gradient-based strategy to emphasize the optimization towards the gradually popular item groups across time periods. We instantiate the TDRO on two State-Of-The-Art (SOTA) cold-start recommender methods and conduct extensive experiments on three real-world datasets. The empirical results under multiple settings (*e.g.*, cold-start and warm-start recommendation, and recommendation with differing degrees of temporal feature shifts) validate the superiority of TDRO in enhancing the generalization ability of cold-start models. We release our codes and datasets at <https://github.com/Linxyhaha/TDRO/>.

The contributions of this work are summarized as follows.

- We point out a crucial issue of temporal feature shifts in cold-start item recommendation and highlight the importance of strengthening the generalization ability of cold-start models.
- We propose a novel TDRO objective for cold-start recommendation, which extends the conventional DRO to avoid overemphasizing the minority groups and capture the temporal shifting trend of item features.
- We conduct extensive experiments on three datasets, demonstrating the effectiveness of temporal DRO in attaining robust prediction under temporal feature shifts.

2 Related Work

- **Cold-start Recommendation.** Traditional CF methods typically rely on CF representations learned from historical

assumed to be a mixture of subgroups, and the uncertainty set is defined on mixtures of these subgroups (*cf.* Section 3).

interactions (Wang et al. 2022; Li et al. 2019; Sun et al. 2022). However, the influx of cold items hinders traditional CF methods from providing appropriate recommendations due to the lack of historical interactions (Zhao et al. 2022; Rajapakse and Leith 2022; Raziperchikolaei, Liang, and Chung 2021; Pulis and Bajada 2021; Du et al. 2022a; Huan et al. 2022; Zhu et al. 2021; Sun et al. 2021; Wang et al. 2021a; Chu et al. 2023). To remedy this, existing methods align the feature representations with interactions (Meng et al. 2020; Guo et al. 2017), falling into two research lines. 1) Robust training-based methods utilize both feature and CF representations for prediction while the CF representations are randomly corrupted (Volkovs, Yu, and Poutanen 2017). 2) Auxiliary loss-based methods introduce different auxiliary losses for minimizing the distance between the feature and CF representations (Wei et al. 2021; Chen et al. 2022). However, previous methods suffer from temporal feature shifts from warm to cold items, where the feature representations of cold items may not be well captured by the feature extractor learned from warm items.

• **Distributionally Robust Optimization.** DRO aims to achieve uniform performance against distribution shifts (He et al. 2022) by optimizing the worst-case performance over a pre-defined uncertainty set (Rahimian and Mehrotra 2019; Michel, Hashimoto, and Neubig 2022). The most representative line of work is discrepancy-based DRO which defines the uncertainty set as a ball surrounding the training distributions with different discrepancy metrics (Duchi and Namkoong 2018; Staib and Jegelka 2019; Liu et al. 2022). Since discrepancy-based DRO suffers from over-pessimism issue (Oren et al. 2019; Sagawa et al. 2020; Duchi, Hashimoto, and Namkoong 2023), another line of research falls into Group-DRO (Zhou et al. 2021; Goel et al. 2021). It defines the uncertainty set as a set of mixtures of subgroups in the training set, encouraging DRO to focus on meaningful distribution shifts (Oren et al. 2019; Wen et al. 2022). Some prior work (Zhou et al. 2023) explores DRO to alleviate long-tail users and items for warm-start recommendation, *e.g.*, S-DRO (Wen et al. 2022), PDRO (Zhao et al. 2023). However, directly applying DRO to cold-start recommendation may cause inconsistency issue. In this work, we consider leveraging a temporally DRO to focus on the mitigation of temporal item feature shifts for cold-start recommendation.

More detailed discussions on related works are presented in Appendix A.8, including other literature on robustness enhancement, such as invariant learning (Arjovsky et al. 2019; Du et al. 2022b; Koyama and Yamaguchi 2020; Liu et al. 2021) and re-weighting strategy (Zhang et al. 2021; Kim et al. 2021; Yang et al. 2023).

3 Preliminary

Cold-start Recommendation. To address the cold-start item issue, existing methods leverage the item features (*e.g.*, categories and visual features) to predict the user-item interactions. Specifically, given the users \mathcal{U} , warm items \mathcal{I}_w with features $\{\mathbf{s}_i | i \in \mathcal{I}_w\}$, and user-item interactions $\mathcal{D} = \{(u, i, y_{ui}) | u \in \mathcal{U}, i \in \mathcal{I}_w\}$ with $y_{ui} \in \{0, 1\}$ indicating whether the user u likes the item i ($y_{ui} = 1$)

or not ($y_{ui} = 0$), the cold-start recommender model aims to learn a feature extractor, an interaction predictor, and the CF representations of users and items for aligning feature representations with user-item interactions. The learnable parameters of the cold-start recommender model, denoted as θ , are optimized via Empirical Risk Minimization (ERM). Formally, we have

$$\theta_{ERM}^* := \arg \min_{\theta \in \Theta} \mathbb{E}_{(u, i, y_{ui}) \in \mathcal{D}} [\mathcal{L}(\theta; (u, i, y_{ui}, \mathbf{s}_i))], \quad (1)$$

where $\mathcal{L}(\cdot)$ is the loss function of the cold-start recommender model and is particularly tailored to different cold-start methods to regulate the alignment.

Nevertheless, such a learning paradigm merely minimizes the expected loss under the same distribution as the training data (Rahimian and Mehrotra 2019). The feature extractor could under-represent the minority groups (Wen et al. 2022), which however might be popular in cold items, leading to the vulnerability to the shifted cold item features.

Distributionally Robust Optimization. To alleviate temporal feature shifts, DRO⁵ is an effective solution that could achieve consistently high performance across various distribution shifts (Zhou et al. 2021; Duchi and Namkoong 2018; Oren et al. 2019; Sagawa et al. 2020; Hu et al. 2018). In detail, DRO assumes the training distribution to be a mixture of K pre-defined groups $\{P_i | i = 1, \dots, K\}$. Then, it optimizes the worst-case performance over the K subgroups for controlling the performance lower bound. Formally,

$$\theta_{DRO}^* := \arg \min_{\theta \in \Theta} \left\{ \max_{j \in [K]} \mathbb{E}_{(u, i, y_{ui}) \sim P_j} [\mathcal{L}(\theta; (u, i, y_{ui}, \mathbf{s}_i))] \right\}. \quad (2)$$

A practical solution to Eq. (2) is to conduct interleave step-wise optimization (Piratla, Netrapalli, and Sarawagi 2022; Sagawa et al. 2020). Specifically, at each update step t , DRO first selects the group with the worst empirical performance:

$$\begin{aligned} j^* &= \arg \max_{j \in \{1, \dots, K\}} \mathbb{E}_{(u, i, y_{ui}) \sim P_j} [\mathcal{L}(\theta; (u, i, y_{ui}, \mathbf{s}_i))] \\ &\approx \arg \min_{j \in \{1, \dots, K\}} -\bar{\mathcal{L}}_j, \end{aligned} \quad (3)$$

where $\bar{\mathcal{L}}_j = \frac{1}{N_j} \sum_{(u, i, y_{ui}) \sim \tilde{P}_j} \mathcal{L}_j(\theta; (u, i, y_{ui}, \mathbf{s}_i))$, \tilde{P}_j is the empirical distribution of group j in dataset \mathcal{D} , and N_j is the number of samples in group j . Subsequently, the model parameters θ are updated based on the selected group, *i.e.*, $\theta^{t+1} = \theta^t - \eta \nabla_{\theta} \bar{\mathcal{L}}_{j^*}(\theta^t)$, where η is the learning rate.

Despite the success of DRO in various domains (*e.g.*, image classification (Zhai et al. 2021; Sagawa et al. 2020), natural language modeling (Oren et al. 2019; Michel, Hashimoto, and Neubig 2022)), directly applying DRO in cold-start recommendation faces an inconsistency issue. It is likely that DRO will overemphasize the minority group in warm items at the expense of performance of other groups (Wen et al. 2022). Besides, the majority and minority item groups may change due to temporal feature shifts, thereby hurting the cold item performance (*cf.* Section 1).

⁵We adopt Group-DRO to avoid over-pessimism issue (refer to Appendix A.1 for details).

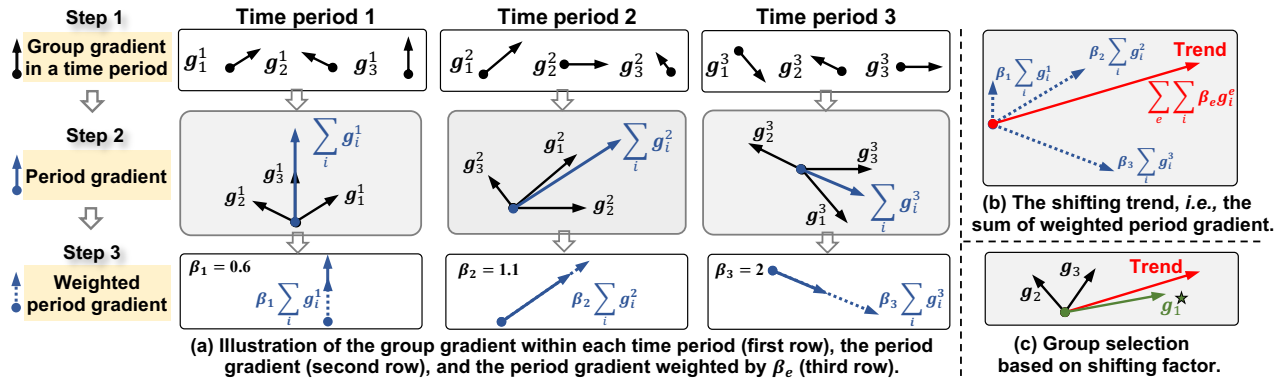


Figure 2: Illustration of the shifting factor with three groups and three time periods (i.e., $i \in \{1, 2, 3\}$ and $e \in \{1, 2, 3\}$). (a) depicts the three steps of obtaining the weighted period gradient in each time period. And then, by summing up the weighted period gradient, we can obtain the shifting trend as shown in (b). Finally, the shifting factor for each group is obtained by calculating the similarity between the group gradient and the shifting trend as presented in (c).

4 Temporally DRO

To alleviate the impact of temporal feature shifts for cold-start recommendation, we propose two new objectives for DRO training: 1) enhancing the worst-case optimization on minority groups to raise the lower bound of performance, and 2) capturing the temporal shifting trend of item features and emphasizing the optimization of groups that are likely to become popular.

4.1 Group Selection

It is noted that the group selection plays a critical role in DRO (Eq. (3)) to strengthen the model’s robustness (Piratla, Netrapalli, and Sarawagi 2022). As such, we propose a novel TDRO, which introduces two factors in group selection: 1) a *worst-case factor* to focus more on minority groups with larger losses and give them priorities for group selection, and 2) a *shifting factor* to emphasize the potentially popular groups in cold items by leveraging the temporal shifting trend. Besides, the shifting factor can alleviate the overemphasis on one particular worst-case group.

Shifting Trend-guided Group Selection. In detail, we first split the warm items into K groups via K -means clustering based on their item features (e.g., visual features of thumbnails). We then split the chronologically sorted interactions into E time periods, $e \in \{1, \dots, E\}$. We denote the average loss of group i in time period e as $\mathcal{L}_i^e(\cdot)$. At each update step t , we consider both the worst-case factor and the shifting factor to select the group j^* for optimization, which is formulated as

$$j^* = \arg \min_{j \in \{1, \dots, K\}} \underbrace{-(1 - \lambda) \bar{\mathcal{L}}_j(\theta^t)}_{\text{(worst-case factor)}} + \lambda \underbrace{\sum_{e=1}^E \sum_{i=1}^K \beta_e \mathcal{L}_i^e(\theta^t - \eta \nabla_{\theta} \bar{\mathcal{L}}_j(\theta^t))}_{\text{(shifting factor)}}, \quad (4)$$

where λ is the hyper-parameter to balance the strength between two factors. The *worst-case factor* calculates the loss value of each group $\bar{\mathcal{L}}_j(\theta^t)$ for group selection. The group with a larger loss will have a smaller $-\bar{\mathcal{L}}_j(\theta^t)$, thus being more likely to be selected. Besides, the *shifting factor* consists of two perspectives:

- To alleviate the overemphasis on one particular worst-case group, the shifting factor selects the optimization group to improve the performance on *all* groups. Specifically, $\theta^t - \eta \nabla_{\theta} \bar{\mathcal{L}}_j(\theta^t)$ is the updated parameters if we choose group j for optimization. Thereafter, the loss of each group i in a time period e after parameter updating will be $\mathcal{L}_i^e(\theta^t - \eta \nabla_{\theta} \bar{\mathcal{L}}_j(\theta^t))$. And the performance improvements for all groups across all periods are measured by $\sum_{e=1}^E \sum_{i=1}^K \mathcal{L}_i^e(\theta^t - \eta \nabla_{\theta} \bar{\mathcal{L}}_j(\theta^t))$.
- To emphasize the potentially popular groups in cold items, the shifting factor upweights the later time periods closer to the test phase. In detail, we use β_e to re-weight the performance improvements over all groups for each time period e . We define $\beta_e = \exp(p \cdot e)$, where a later period e will have a higher weight and $p > 0$ is the hyper-parameter to control the steepness. A smaller p encourages time periods to be uniformly important, while a larger p upweights the time periods closer to the test phase.

However, directly applying Eq. (4) for group selection will incur extensive resource costs as we need to consider all possible cases of the updated parameters. Fortunately, we can approximate Eq. (4) into a gradient-based formulation via First-order Taylor formulation (see Appendix A.2 for detailed derivation):

$$j^* = \arg \max_{j \in \{1, \dots, K\}} \underbrace{(1 - \lambda) \bar{\mathcal{L}}_j(\theta^t)}_{\text{(worst-case factor)}} + \lambda \langle \mathbf{g}_j, \underbrace{\sum_{e=1}^E \sum_{i=1}^K \beta_e \mathbf{g}_i^e}_{\text{(shifting factor)}} \rangle, \quad (5)$$

where $\mathbf{g}_j = \nabla_{\theta} \bar{\mathcal{L}}_j(\theta)$ denotes the gradient of the average loss of group j , and $\mathbf{g}_i^e = \nabla_{\theta} \mathcal{L}_i^e(\theta)$ denotes the gradient of group i ’s average loss in time period e . The $\langle \cdot, \cdot \rangle$ represents

the inner product computation. Since $\sum_{e=1}^E \sum_{i=1}^K \beta_e \mathbf{g}_i^e$ is a constant vector (referred to as *shifting trend*) for any group j , we can avoid this cumbersome computations in Eq. (4) for efficient group selection.

Interpretation of Shifting Factor. For an intuitive understanding of the gradient-based shifting factor, we visualize a toy example in Figure 2, where we set $K = 3$ and $E = 3$.

- **Factor decomposition.** As shown in Figure 2(a), we have three decomposed group gradients, $\mathbf{g}_{i \in \{1,2,3\}}^e$, for each time period e . We can then obtain the period gradient $\sum_{i=1}^K \mathbf{g}_i^e$ of time period e by summing up the decomposed group gradients. Since the gradient indicates the optimization direction, the sum of the gradient within each time period, *i.e.*, period gradient, represents the optimal updating direction in each temporally shifted distribution. Subsequently, by multiplying the period importance β_e to each time period and summing up the weighted period gradient, we can obtain the shifting trend $\sum_{e=1}^E \sum_{i=1}^K \beta_e \mathbf{g}_i^e$ that reflects optimization direction on potentially popular groups (Figure 2(b)).

- **Factor interpretation.** Finally, the shifting factor is obtained by calculating the inner product of the shifting trend and the group gradient \mathbf{g}_j (see Figure 2(c)). Since the shifting trend is a constant vector for all groups, the shifting factor essentially measures the similarity between each group gradient and the shifting trend, *i.e.*, optimization direction emphasizing the potentially popular item groups.

As for model optimization at each step, we first select the optimal group j^* via Eq. (5), and then update the parameters θ by gradient descent $\theta^{t+1} = \theta^t - \eta \nabla_{\theta} \bar{\mathcal{L}}_{j^*}(\theta^t)$.

4.2 Gradient Smoothing

Despite the success of step-wise optimization in many applications (Sagawa et al. 2020), directly employing such strategy in recommender systems suffers from training instability (Wen et al. 2022). As such, we follow the previous work (Piratla, Netrapalli, and Sarawagi 2022; Wen et al. 2022) by incorporating gradient smoothing for optimization from two aspects: group importance smoothing and loss consistency enhancement.

- **Group importance smoothing.** We consider assigning weight vector \mathbf{w} for groups and regulate the weight dynamic by η_w . Formally,

$$\mathbf{w}^{t+1} = \arg \max_{\mathbf{w}_i \in [K]} \sum_i w_i [(1 - \lambda) \bar{\mathcal{L}}_i(\theta) + \lambda \langle \mathbf{g}_i, \sum_{e=1}^E \sum_{j=1}^K \beta_e \mathbf{g}_j^e \rangle] - \frac{1}{\eta_w} \text{KL}(\mathbf{w}, \mathbf{w}^t), \quad (6)$$

where w_i is the i -th entry of \mathbf{w} , η is the learning rate, and $\text{KL}(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log \frac{p_i}{q_i}$ is the KL-divergence between \mathbf{p} and \mathbf{q} . By applying KKT conditions (refer to Appendix A.2), we obtain the closed-form solution of Eq. (6):

$$w_i^{t+1} = \frac{w_i^t \exp(\eta_w [(1 - \lambda) \bar{\mathcal{L}}_i(\theta) + \lambda \langle \mathbf{g}_i, \sum_{e=1}^E \sum_{j=1}^K \beta_e \mathbf{g}_j^e \rangle])}{\sum_s w_s^t \exp(\eta_w [(1 - \lambda) \bar{\mathcal{L}}_s(\theta) + \lambda \langle \mathbf{g}_s, \sum_{e=1}^E \sum_{j=1}^K \beta_e \mathbf{g}_j^e \rangle])}. \quad (7)$$

Thereafter, the model parameters θ are updated through

$$\theta^{t+1} = \theta^t - \eta \sum_i w_i^{t+1} \nabla \bar{\mathcal{L}}_i(\theta^t). \quad (8)$$

Algorithm 1: Training Procedure of TDRO

Input: Number of groups K , number of time periods E , initial model parameters θ^0 , initial group weight $\mathbf{w} = (\frac{1}{K}, \frac{1}{K}, \dots, \frac{1}{K})$, initial group loss $\bar{\mathcal{L}}_{i \in [K]}^0$, item features $\{\mathbf{s}_i | i \in \mathcal{I}_w\}$, interactions \mathcal{D} , shifting factor strength λ , period importance $\beta_{e \in [E]}$, weight step size η_w , streaming step size μ , and learning rate η .

```

1: while not converge do
2:   for all  $i \in \{1, \dots, K\}$  do
3:     Calculate  $\bar{\mathcal{L}}_i^t(\theta^t)$  via cold-start loss function.
4:      $\bar{\mathcal{L}}_i^t(\theta^t) \leftarrow (1 - \mu) \bar{\mathcal{L}}_i^{t-1}(\theta^{t-1}) + \mu \bar{\mathcal{L}}_i^t(\theta^t)$ 
5:   for all  $i \in \{1, \dots, K\}$  do
6:      $w_i^{t+1} \leftarrow w_i^t \exp(\eta_w [(1 - \lambda) \bar{\mathcal{L}}_i^t(\theta^t) + \lambda (\nabla \bar{\mathcal{L}}_i^t(\theta^t) \sum_{e=1}^E \sum_{j=1}^K \beta_e \nabla \bar{\mathcal{L}}_j^{e,t}(\theta^t))])$ 
7:    $w_i^{t+1} \leftarrow w_i^{t+1} / \|\mathbf{w}^{t+1}\|_1, \forall i \in \{1, \dots, K\}$   $\triangleright$  Normalize
8:    $\theta^{t+1} \leftarrow \theta^t - \eta \sum_{i \in [K]} w_i^{t+1} \nabla \bar{\mathcal{L}}_i^t(\theta^t)$   $\triangleright$  Update

```

Output: Optimized model parameters θ .

- **Loss consistency enhancement.** To alleviate the training instability caused by aggravated data sparsity after group and time period division, we follow (Wen et al. 2022) to keep the streaming estimations of empirical loss:

$$\bar{\mathcal{L}}_j^t \leftarrow (1 - \mu) \bar{\mathcal{L}}_j^{t-1} + \mu \bar{\mathcal{L}}_j^t,$$

where μ is the hyper-parameter to control the streaming step size. A smaller μ leads to more conservative training (see Appendix A.2 for details).

- **Instantiation.** To instantiate TDRO on cold-start recommender models, we first calculate the group weight \mathbf{w} via Eq. (7), where $\mathcal{L}(\theta)$ can be substituted by any form of the loss function from the backend cold-start models. The model parameters will then be optimized based on weighted gradient descent via Eq. (8). Training details of TDRO are presented in Algorithm 1.

5 Experiments

We conduct extensive experiments on three real-world datasets to answer the following research questions:

- **RQ1:** How does our proposed TDRO perform compared to the baselines under temporal feature shifts?
- **RQ2:** How do the different components of TDRO (*i.e.*, two factors for group selection) affect the performance?
- **RQ3:** How does TDRO perform over different strengths of temporal feature shifts and how does TDRO mitigate the impact of shifts?

5.1 Experimental Settings

Datasets. We conducted experiments on three real-world datasets across different domains: 1) **Amazon** (He and McAuley 2016) is a representative clothing dataset with rich visual features of clothing images. 2) **Micro-video** is a real-world industry dataset collected from a popular micro-video platform, with rich visual and textual features from thumbnails and textual descriptions. 3) **Kwai**⁶ is a benchmark rec-

⁶<https://www.kwai.com/>.

Metric	Models	Amazon			Micro-video			Kwai		
		All	Warm	Cold	All	Warm	Cold	All	Warm	Cold
Recall@20	DUIF	0.0042	0.0048	0.0129	0.0318	0.0537	0.0771	0.0208	0.0248	0.0158
	DropoutNet	0.0050	0.0110	0.0050	0.0187	0.0494	0.0222	0.0099	0.0118	0.0066
	M2TRec	0.0065	0.0058	0.0068	0.0131	0.0056	0.0298	0.0317	0.0320	0.0009
	MTPR	0.0057	0.0116	0.0082	0.0303	0.0723	0.0542	0.0464	0.0550	0.0049
	Heater	0.0065	0.0136	0.0040	0.0469	0.1153	0.0868	0.0452	0.0536	0.0087
	CB2CF	0.0078	0.0170	0.0074	0.0496	0.0961	0.0928	0.0624	0.0737	0.0064
	CCFCRec	0.0071	0.0175	0.0117	0.0435	0.0750	0.0699	0.0098	0.0141	0.0129
	InvRL	0.0120	0.0183	0.0150	0.0578	0.0899	0.0754	0.0588	0.0701	0.0191
	CLCRec	0.0106	0.0200	0.0135	0.0583	0.1135	0.0623	0.0743	0.0884	0.0160
	+S-DRO	0.0121	0.0237	0.0144	0.0656	0.1173	0.0719	0.0661	0.0787	0.0172
	+TDRO	0.0130*	0.0237*	0.0166*	0.0703*	0.1180*	0.0761*	0.0841*	0.1016*	0.0186*
	GAR	0.0079	0.0200	0.0124	0.0644	0.0962	0.0840	0.0588	0.0706	0.0051
	+S-DRO	0.0078	0.0189	0.0132	0.0626	0.0894	0.0874	0.0579	0.0690	0.0050
	+TDRO	0.0087*	0.0236*	0.0150*	0.0711*	0.1104*	0.0947*	0.0598*	0.0719*	0.0052
NDCG@20	DUIF	0.0020	0.0023	0.0058	0.0204	0.0295	0.0511	0.0158	0.0181	0.0070
	DropoutNet	0.0021	0.0043	0.0021	0.0117	0.0286	0.0121	0.0054	0.0061	0.0030
	M2TRec	0.0032	0.0029	0.0030	0.0075	0.0036	0.0211	0.0247	0.0248	0.0004
	MTPR	0.0029	0.0056	0.0030	0.0175	0.0389	0.0362	0.0324	0.0369	0.0021
	Heater	0.0037	0.0075	0.0015	0.0290	0.0653	0.0484	0.0276	0.0312	0.0030
	CB2CF	0.0037	0.0076	0.0031	0.0254	0.0490	0.0636	0.0446	0.0504	0.0026
	CCFCRec	0.0032	0.0074	0.0050	0.0321	0.0410	0.0464	0.0068	0.0092	0.0058
	InvRL	0.0056	0.0079	0.0072	0.0355	0.0493	0.0503	0.0390	0.0444	0.0088
	CLCRec	0.0054	0.0093	0.0061	0.0417	0.0728	0.0444	0.0536	0.0610	0.0071
	+S-DRO	0.0060	0.0107	0.0071	0.0451	0.0747	0.0480	0.0472	0.0536	0.0076
	+TDRO	0.0066*	0.0112*	0.0077*	0.0507*	0.0794*	0.0511*	0.0597*	0.0719*	0.0081*
	GAR	0.0041	0.0088	0.0060	0.0375	0.0496	0.0625	0.0421	0.0485	0.0021
	+S-DRO	0.0033	0.0089	0.0052	0.0385	0.0474	0.0532	0.0423	0.0481	0.0021
	+TDRO	0.0041	0.0110*	0.0066*	0.0419*	0.0571*	0.0638*	0.0431*	0.0495*	0.0024*

Table 1: Overall performance comparison between the baselines and two SOTA models enhanced by TDRO on three datasets. The bold results highlight the better performance in the comparison between the backbone models with and without TDRO. * implies that the improvements over the backbone models are statistically significant (p -value < 0.01) under one-sample t-tests.

ommendation dataset provided with rich visual features. For Amazon and Micro-video datasets, we split the interactions into training, validation, and testing sets chronologically at the ratio of 8:1:1 according to the timestamps. For the Kwai dataset, due to the lack of global timestamps, we instead follow previous work (Wei et al. 2021) that randomly split the interactions. In addition, we divide the items in the validation and testing sets into warm and cold sets, where items that do not appear in the training set are regarded as cold items, and the rest as warm items. The statistics of datasets are summarized in Appendix Table 5.

Evaluation. We adopt the full-ranking protocol (Wei et al. 2021) for evaluation. We consider three different settings: full-ranking over 1) all items, 2) warm items only, and 3) cold items only, denoted respectively as “all”, “warm”, and “cold” settings. The widely-used Recall@20 and NDCG@20 are employed as evaluation metrics.

Baselines. We compare TDRO with competitive cold-start recommender models, including 1) *robust training-based methods*: DUIF (Geng et al. 2015), DropoutNet (Volkovs, Yu, and Poutanen 2017), M2TRec (Shalaby et al. 2022), and MTPR (Du et al. 2020)), and 2) *auxiliary loss-based methods*: Heater (Zhu et al. 2020), CB2CF (Barkan et al. 2019), CCFCRec (Zhou, Zhang, and Yang 2023), CLCRec (Wei et al. 2021), and GAR (Chen et al. 2022). Additionally, we also consider 3) *potential methods* to overcome temporal feature shifts: S-DRO (Wen et al. 2022) and invariant

learning framework (Du et al. 2022b). Details of baselines and the hyper-parameter tuning of baselines and TDRO are summarized in Appendix A.4 and A.5.

5.2 Overall Performance (RQ1)

The overall performance of the baselines and the two SOTA cold-start methods equipped with S-DRO and TDRO is reported in Table 1, from which we can observe the following:

- Auxiliary loss-based methods typically outperform the robust training-based ones. The reason is that robust training-based methods directly utilize feature representations to fit interactions, which inevitably introduces noises and hurt the learning of the recommender model. Meanwhile, auxiliary loss-based methods decouple the CF and feature representations space, which protects the CF representations from feature noises and improves cold performance effectively via different auxiliary losses.
- CLCRec consistently yields impressive performance across the three datasets. This is attributed to the integration of contrastive loss for aligning feature and CF representations, where mutual information between feature and CF space is maximized for robust prediction. Besides, by introducing adversarial constraints for similar distributions of CF and feature representations, GAR exhibits competitive performance despite its instability.
- In most cases, S-DRO improves the performance of cold items compared to the backbone model. The stable improvements are attributed to the tail performance guarantee over

Methods	Amazon			Micro-video			Kwai		
	All Recall@20	Warm Recall@20	Cold Recall@20	All Recall@20	Warm Recall@20	Cold Recall@20	All Recall@20	Warm Recall@20	Cold Recall@20
CLCRec	0.0106	0.0200	0.0135	0.0583	0.1135	0.0623	0.0743	0.0884	0.0160
w/o Worst-case Factor	0.0121	0.0219	0.0157	0.0648	0.1138	0.0687	0.0790	0.0997	0.0145
w/o Shifting Factor	0.0126	0.0228	0.0160	0.0643	0.1145	0.0622	0.0797	0.0986	0.0165
TDRO	0.0130	0.0237	0.0166	0.0703	0.1180	0.0761	0.0814	0.1016	0.0186

Table 2: Ablation study of worst-case factor and shifting factor. The best results are highlighted in bold.

	Amazon			Micro-video		
	Group1	Group2	Group3	Group1	Group2	Group3
Distance	48	62	123	13	19	39
CLCRec	0.0218	0.0075	0.0024	0.1131	0.0503	0.0116
TDRO	0.0254	0.0110	0.0027	0.1321	0.0598	0.0139

Table 3: Recall@20 over user groups with different strengths of temporal feature shifts under “all” setting.

potential shifted distributions, which may partially cover the shifted cold item distribution. In addition, our proposed TDRO consistently outperforms S-DRO and the backbone model on all and cold performance by a large margin. This justifies the effectiveness of TDRO in enhancing the generalization ability of the feature extractor. Moreover, capturing the shifting patterns is also helpful for achieving steady improvements for warm items, reflecting the superiority of TDRO in alleviating the temporal feature shifts issue. Possible reasons for inferior performance on InvRL and M2TRec are discussed in Appendix A.6.

5.3 In-depth Analysis

In this subsection⁷, we study two factors of TDRO, investigate the effectiveness of TDRO under different strengths of temporal feature shifts, and explore how EQUAL mitigates the impact of temporal feature shifts.

Ablation Study (RQ2). To study the effectiveness of the worst-case and shifting factor, we implement TDRO without (w/o) each factor, separately. From Table 2, we can find that: 1) The performance declines if either the worst-case factor or the shifting factor is removed. This verifies the effectiveness of incorporating the optimization over worst-case group and the performance improvements for all groups based on the shifting trend. 2) Removing each factor still outperforms CLCRec (“all” setting). This indicates that either performance lower bound guarantee or leveraging shifting trends improves generalization ability.

User Group Evaluation (RQ3). We further inspect how TDRO performs under different strengths of temporal feature shifts by evaluating TDRO on different user groups. Specifically, we calculate the Euclidean distance of the average item features between the training set and testing set for each user. Next, we rank the users according to the distance, and then split the users into three groups (denoted

⁷We focus on the analysis of TDRO instantiated on CLCRec due to its better performance compared to GAR to save space.

	All		Cold	
	Worst-case	Popular	Worst-case	Popular
CLCRec	0.0166	0.0168	0.0088	0.0088
TDRO	0.0173	0.0195	0.0123	0.0125

Table 4: Recall@20 of the item group with the worst performance and the item group of top 25% popular items.

as Group 1, Group 2, and Group 3) based on the ranking. The results *w.r.t.* Recall@20 is given in Table 3. Despite that the performance of both CLCRec and TDRO declines gradually as the shifts become more significant, TDRO consistently outperforms CLCRec in each group, validating the effectiveness of TDRO in enhancing the generalization ability to temporal feature shifts.

Item Group Analysis (RQ3). To explore how TDRO alleviates the impact of temporal feature shifts, we analyze the generalization ability enhancement of TDRO on Amazon *w.r.t.* item groups. In detail, we calculate the item popularity (*i.e.*, interaction proportion) in the testing set and divide the items into four subgroups based on the popularity scores. We then conduct evaluation on each item subgroup to see whether TDRO: 1) guarantees the worst-case group performance, and 2) enhances the performance over the group with the top 25% popular items. As shown in Table 4, the boosted performance on worst-case group and popular items partially explains the superior performance of TDRO.

6 Conclusion and Future Work

In this work, we revealed the critical issue of temporal item feature shifts in the cold-start recommendation. To overcome this issue, we proposed a novel temporal DRO learning framework called TDRO, which 1) considers the worst-case performance for the performance lower bound guarantee, and 2) leverages the shifting trend of item features to enhance the performance of popular groups in subsequent cold items. Empirical results on three real-world datasets validated the effectiveness of TDRO in achieving robust prediction under temporal item feature shifts.

This work highlights temporal feature shifts in cold-start recommendation, leaving many promising directions to be explored in the future. One is to consider adaptive environment importance for more fine-grained modeling of the shifting trend. Moreover, it is worthwhile to explore more effective group division strategies beyond the pre-defined ones, to fulfill the potential of TDRO in enhancing the model’s generalization ability enhancement.

References

- Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv:1907.02893*.
- Barkan, O.; Koenigstein, N.; Yogev, E.; and Katz, O. 2019. CB2CF: a neural multiview content-to-collaborative filtering model for completely cold item recommendations. In *RecSys*, 228–236. ACM.
- Chen, H.; Wang, Z.; Huang, F.; Huang, X.; Xu, Y.; Lin, Y.; He, P.; and Li, Z. 2022. Generative adversarial framework for cold-start item recommendation. In *SIGIR*, 2565–2571. ACM.
- Chu, Z.; Wang, H.; Xiao, Y.; Long, B.; and Wu, L. 2023. Meta policy learning for cold-start conversational recommendation. In *WSDM*, 222–230. ACM.
- Du, J.; Ye, Z.; Yao, L.; Guo, B.; and Yu, Z. 2022a. Socially-aware dual contrastive learning for cold-start recommendation. In *SIGIR*, 1927–1932. ACM.
- Du, X.; Wang, X.; He, X.; Li, Z.; Tang, J.; and Chua, T.-S. 2020. How to learn item representation for cold-start multimedia recommendation? In *MM*, 3469–3477. ACM.
- Du, X.; Wu, Z.; Feng, F.; He, X.; and Tang, J. 2022b. Invariant representation learning for multimedia recommendation. In *MM*, 619–628. ACM.
- Duchi, J.; Hashimoto, T.; and Namkoong, H. 2023. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 71(2): 649–664.
- Duchi, J.; and Namkoong, H. 2018. Learning models with uniform performance via distributionally robust optimization. *arXiv:1810.08750*.
- Geng, X.; Zhang, H.; Bian, J.; and Chua, T.-S. 2015. Learning image and user features for recommendation in social networks. In *ICCV*, 4274–4282. IEEE.
- Goel, K.; Gu, A.; Li, Y.; and Ré, C. 2021. Model patching: closing the subgroup performance gap with data augmentation. In *ICLR*.
- Guo, C.; Lu, H.; Shi, S.; Hao, B.; Liu, B.; Zhang, M.; Liu, Y.; and Ma, S. 2017. How integration helps on cold-start recommendations. In *RecSys Challenge*, 1–6. ACM.
- Hao, B.; Zhang, J.; Yin, H.; Li, C.; and Chen, H. 2021. Pre-training graph neural networks for cold-start users and items representation. In *WSDM*, 265–273. ACM.
- He, R.; and McAuley, J. 2016. Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, 507–517. ACM.
- He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *WWW*, 173–182. ACM.
- He, Y.; Wang, Z.; Cui, P.; Zou, H.; Zhang, Y.; Cui, Q.; and Jiang, Y. 2022. CausPref: Causal Preference Learning for Out-of-Distribution Recommendation. In *WWW*, 410–421. ACM.
- Houlsby, N.; Hernández-Lobato, J. M.; and Ghahramani, Z. 2014. Cold-start active learning with robust ordinal matrix factorization. In *ICML*, 766–774. PMLR.
- Hu, W.; Niu, G.; Sato, I.; and Sugiyama, M. 2018. Does distributionally robust supervised learning give robust classifiers? In *ICML*, 2029–2037. PMLR.
- Huan, Z.; Zhang, G.; Zhang, X.; Zhou, J.; Wu, Q.; Gu, L.; Gu, J.; He, Y.; Zhu, Y.; and Mo, L. 2022. An industrial framework for cold-start recommendation in zero-shot scenarios. In *SIGIR*, 3403–3407. ACM.
- Kim, M.; Song, H.; Kim, D.; Shin, K.; and Lee, J.-G. 2021. Pre-mere: meta-reweighting via self-ensembling for point-of-interest recommendation. In *AAAI*, 4164–4171. AAAI press.
- Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.
- Koyama, M.; and Yamaguchi, S. 2020. When is invariance useful in an out-of-distribution generalization problem? *arXiv:2008.01883*.
- Li, Y.; Liu, M.; Yin, J.; Cui, C.; Xu, X.-S.; and Nie, L. 2019. Routing micro-videos via a temporal graph-guided recommendation system. In *MM*, 1464–1472.
- Liu, J.; Hu, Z.; Cui, P.; Li, B.; and Shen, Z. 2021. Heterogeneous risk minimization. In *ICML*, 6804–6814. PMLR.
- Liu, J.; Wu, J.; Li, B.; and Cui, P. 2022. Distributionally robust optimization with data geometry. In *NeurIPS*, 33689–33701. Curran Associates, Inc.
- Meng, Y.; Yan, X.; Liu, W.; Wu, H.; and Cheng, J. 2020. Wasserstein collaborative filtering for item cold-start recommendation. In *UMAP*, 318–322. ACM.
- Michel, P.; Hashimoto, T.; and Neubig, G. 2021. Modeling the second player in distributionally robust optimization. In *ICLR*.
- Michel, P.; Hashimoto, T.; and Neubig, G. 2022. Distributionally robust models with parametric likelihood ratios. In *ICLR*.
- Neupane, K. P.; Zheng, E.; Kong, Y.; and Yu, Q. 2022. A dynamic meta-learning model for time-sensitive cold-start recommendations. In *AAAI*, 7868–7876. AAAI press.
- Oren, Y.; Sagawa, S.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust language modeling. *arXiv:1909.02060*.
- Pan, F.; Li, S.; Ao, X.; Tang, P.; and He, Q. 2019. Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings. In *SIGIR*, 695–704. ACM.
- Piratla, V.; Netrapalli, P.; and Sarawagi, S. 2022. Focus on the common good: group distributional robustness follows. In *ICLR*.
- Pulis, M.; and Bajada, J. 2021. Siamese neural networks for content-based cold-start music recommendation. In *RecSys*, 719–723. ACM.
- Rahimian, H.; and Mehrotra, S. 2019. Distributionally robust optimization: A review. *arXiv:1908.05659*.
- Rajapakse, D. C.; and Leith, D. 2022. Fast and accurate user cold-start learning using monte carlo tree search. In *RecSys*, 350–359. ACM.
- Raziperchikolaei, R.; Liang, G.; and Chung, Y.-j. 2021. Shared neural item representations for completely cold start problem. In *RecSys*, 422–431. ACM.
- Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. In *ICLR*.
- Shalaby, W.; Oh, S.; Afsharinejad, A.; Kumar, S.; and Cui, X. 2022. M2TRec: metadata-aware multi-task transformer for large-scale and cold-start free session-based recommendations. In *RecSys*, 573–578. ACM.
- Shi, S.; Zhang, M.; Liu, Y.; and Ma, S. 2018. Attention-based adaptive model to unify warm and cold starts recommendation. In *CIKM*, 127–136. ACM.
- Staib, M.; and Jegelka, S. 2019. Distributionally robust optimization and generalization in kernel methods. In *NeurIPS*, 9131–9141. Curran Associates, Inc.
- Sun, T.; Wang, W.; Jing, L.; Cui, Y.; Song, X.; and Nie, L. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *MM*, 15–23.
- Sun, X.; Shi, T.; Gao, X.; Kang, Y.; and Chen, G. 2021. FORM: follow the online regularized meta-leader for cold-start recommendation. In *SIGIR*, 1177–1186. ACM.

Togashi, R.; Otani, M.; and Satoh, S. 2021. Alleviating cold-start problems in recommendation through pseudo-labelling over knowledge graph. In *WSDM*, 931–939. ACM.

Vapnik, V. 1991. Principles of risk minimization for learning theory. In *NeurIPS*, 831–838. Curran Associates, Inc.

Vartak, M.; Thiagarajan, A.; Miranda, C.; Bratman, J.; and Larochelle, H. 2017. A meta-learning perspective on cold-start recommendations for items. In *NeurIPS*, 6904–6914. Curran Associates, Inc.

Volkovs, M.; Yu, G.; and Poutanen, T. 2017. Dropoutnet: addressing cold start in recommender systems. In *NeurIPS*, 4957–4966. Curran Associates, Inc.

Wang, S.; Zhang, K.; Wu, L.; Ma, H.; Hong, R.; and Wang, M. 2021a. Privileged graph distillation for cold start recommendation. In *SIGIR*, 1187–1196. ACM.

Wang, W.; Feng, F.; He, X.; Nie, L.; and Chua, T.-S. 2021b. Denoising implicit feedback for recommendation. In *WSDM*, 373–381. ACM.

Wang, W.; Lin, X.; Feng, F.; He, X.; Lin, M.; and Chua, T.-S. 2022. Causal Representation Learning for Out-of-Distribution Recommendation. In *WWW*, 3562–3571. ACM.

Wei, T.; Wu, Z.; Li, R.; Hu, Z.; Feng, F.; He, X.; Sun, Y.; and Wang, W. 2020. Fast adaptation for cold-start collaborative filtering with meta-learning. In *ICDM*, 661–670. IEEE.

Wei, Y.; Wang, X.; Li, Q.; Nie, L.; Li, Y.; Li, X.; and Chua, T.-S. 2021. Contrastive learning for cold-start recommendation. In *MM*, 5382–5390. ACM.

Wen, H.; Yi, X.; Yao, T.; Tang, J.; Hong, L.; and Chi, E. H. 2022. Distributionally-robust recommendations for improving worst-case user experience. In *WWW*, 3606–3610. ACM.

Yang, L.; Wang, S.; Tao, Y.; Sun, J.; Liu, X.; Yu, P. S.; and Wang, T. 2023. DGRec: graph neural network for recommendation with diversified embedding generation. In *WSDM*, 661–669. ACM.

Zhai, R.; Dan, C.; Kolter, Z.; and Ravikumar, P. 2021. DORO: distributional and outlier robust optimization. In *ICML*, 12345–12355. PMLR.

Zhang, Y.; Feng, F.; He, X.; Wei, T.; Song, C.; Ling, G.; and Zhang, Y. 2021. Causal intervention for leveraging popularity bias in recommendation. In *SIGIR*, 11–20. ACM.

Zhao, J.; Wang, W.; Lin, X.; Qu, L.; Zhang, J.; and Chua, T.-S. 2023. Popularity-aware Distributionally Robust Optimization for Recommendation System. In *CIKM*, 4967–4973.

Zhao, X.; Ren, Y.; Du, Y.; Zhang, S.; and Wang, N. 2022. Improving Item Cold-start Recommendation via Model-agnostic Conditional Variational Autoencoder. In *SIGIR*, 2595–2600. ACM.

Zhou, C.; Ma, X.; Michel, P.; and Neubig, G. 2021. Examining and combating spurious features under distribution shift. In *ICML*, 12857–12867. PMLR.

Zhou, R.; Wu, X.; Qiu, Z.; Zheng, Y.; and Chen, X. 2023. Distributionally Robust Sequential Recommendation. In *SIGIR*, 279–288.

Zhou, Z.; Zhang, L.; and Yang, N. 2023. Contrastive collaborative filtering for cold-start item recommendation. In *WWW*, 928–937. ACM.

Zhu, Z.; Kim, J.; Nguyen, T.; Fenton, A.; and Caverlee, J. 2021. Fairness among new items in cold start recommender systems. In *SIGIR*, 767–776. ACM.

Zhu, Z.; Sefati, S.; Saadatpanah, P.; and Caverlee, J. 2020. Recommendation for new users and new items via randomized training and mixture-of-experts transformation. In *SIGIR*, 1121–1130. ACM.

A Appendix

A.1 DRO

Overpessimism issue. DRO is an effective solution that could achieve consistently high performance across various distribution shifts (Zhou et al. 2021; Duchi and Namkoong 2018). In formal,

$$\theta_{\text{DRO}}^* := \arg \min_{\theta \in \Theta} \left\{ \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(u,i,y_{ui}) \in Q} [\mathcal{L}(\theta; (u, i, y_{ui}, \mathbf{s}_i))] \right\}, \quad (9)$$

where the uncertainty set \mathcal{Q} encodes the possible shifted distributions that we want our model to perform well on.

The uncertainty set \mathcal{Q} is typically defined as a ball surrounding the training distributions endowed with a certain discrepancy metric, such as f-divergence (Duchi and Namkoong 2018) and Wasserstein Distance (Liu et al. 2022). By choosing the discrepancy metric and the radius for determining the boundary of the possible shifted distributions, we could confer the robustness of the model to a wide range of distribution shifts. Nevertheless, such a definition of uncertainty set could lead to the over-pessimism issue, where implausible shifted distributions are being overwhelmingly considered (Liu et al. 2022; Oren et al. 2019). Therefore, to overcome the over-pessimism issue, Group-DRO is proposed to find the potential shifted distributions on a group level (Oren et al. 2019; Sagawa et al. 2020; Hu et al. 2018).

Objective of Group-DRO. In Group-DRO, the training distribution \mathcal{P} is assumed to be a mixture of K pre-defined groups $\{P_i | i = 1, \dots, K\}$, i.e., $\mathcal{P} := \sum_{i=1}^K w_i P_i$, where w_i is the mixture ratio of the group i . Then, the uncertainty set of Group-DRO is defined as $\mathcal{Q} := \{\sum_{i=1}^K w'_i P_i : w'_i \in \Delta_K\}$, where Δ_K is the $K - 1$ -dimensional probability simplex. Since the maximum of $\sum_{i=1}^K w'_i P_i$ is attained at a vertex due to linearity, the objective of Group-DRO is reformulated as

$$\theta_{\text{DRO}}^* := \arg \min_{\theta \in \Theta} \left\{ \max_{j \in [K]} \mathbb{E}_{(u,i,y_{ui}) \sim P_j} [\mathcal{L}(\theta; (u, i, y_{ui}, \mathbf{s}_i))] \right\}. \quad (10)$$

A.2 Formula Derivation

Interpretation of shifting factor. Via First-order Taylor, we can rewrite Eq. (4) into,

$$\begin{aligned} j^* &\approx \arg \min_{j \in \{1, \dots, K\}} (\lambda - 1) \bar{\mathcal{L}}_j(\theta^t) + \\ &\quad \lambda \sum_{e=1}^E \sum_{i=1}^K \beta_e [\mathcal{L}_i^e(\theta^t) + \eta \nabla_{\theta} \mathcal{L}_i^e(\theta^t) (\theta^{t+1} - \theta^t)] \\ &= \arg \min_{j \in \{1, \dots, K\}} (\lambda - 1) \bar{\mathcal{L}}_j(\theta^t) + \lambda \sum_{e=1}^E \sum_{i=1}^K \beta_e [\mathcal{L}_i^e(\theta^t) - \eta \langle \mathbf{g}_i^e, \mathbf{g}_j \rangle], \\ &= \arg \min_{j \in \{1, \dots, K\}} (\lambda - 1) \bar{\mathcal{L}}_j(\theta^t) - \lambda \sum_{e=1}^E \sum_{i=1}^K \tilde{\beta}_e \langle \mathbf{g}_i^e, \mathbf{g}_j \rangle, \end{aligned} \quad (11)$$

where $\tilde{\beta}_e = \beta_e \cdot \eta$. We use β_e to denote $\tilde{\beta}_e$ in Eq. (5) for notation brevity.

Group weight smoothing. We consider the optimization problem,

$$\begin{aligned} \min_{\mathbf{w}} & - \sum_i w_i c_i + \frac{1}{\eta_w} \sum_i w_i \log \frac{w_i}{w_i^t} + \delta_{\geq 0}(\mathbf{w}), \\ \text{s.t.} & \sum_i w_i = 1. \end{aligned} \quad (12)$$

where $c_i = (1 - \lambda) \bar{\mathcal{L}}_i(\theta^t) + \lambda \langle \mathbf{g}_i, \sum_{e=1}^E \sum_{j=1}^K \beta_e \mathbf{g}_j^e \rangle$, and $\delta_{\geq 0}(\mathbf{w})$ is the indicator function of set $\{x | x \geq 0\}$. For $y \in \mathbb{R}$, we have the Lagrangian,

$$L(\mathbf{w}, y) = \sum_i \left[w_i c_i + \frac{1}{\eta_w} \sum_i w_i \log \frac{w_i}{w_i^t} + \delta_{\geq 0}(\mathbf{w}) + w_i y \right] - y. \quad (13)$$

Then the Lagrange dual function is $\theta_y = \max L(\mathbf{w}, y)$. For every i , we solve,

$$\begin{aligned} \max & -w_i c_i + \frac{1}{\eta_w} \log \frac{w_i}{w_i^t} + w_i y, \\ \text{s.t.} & \sum_i w_i = 1. \end{aligned} \quad (14)$$

By setting $\nabla_{w_i} = -c_i + \frac{1}{\eta_w} (\log \frac{w_i}{w_i^t} + 1) + y = 0$, we obtain the solution as,

$$w_i = \frac{w_i^t \exp(\eta_w c_i)}{\sum_j w_j^t \exp(\eta_w c_j)}. \quad (15)$$

Streaming estimations of empirical loss. In DRO training, the loss variance for the selected group can be high (Wen et al. 2022), thus leading to instability. To reduce the loss variance, previous work (Wen et al. 2022) proposes a streaming algorithm. The key idea is to keep streaming estimations of the empirical loss $\bar{\mathcal{L}}_j^t$ by updating the loss value in a small step μ , in a similar way to stochastic gradient descent:

$$\bar{\mathcal{L}}_j^t \leftarrow (1 - \mu) \bar{\mathcal{L}}_j^{t-1} + \mu \bar{\mathcal{L}}_j^t.$$

We can find that a smaller μ preserves more loss value at the last time step, i.e., $\bar{\mathcal{L}}_j^{t-1}$. As such, the optimization will be less affected by batches where sparse subgroups do not exist.

A.3 Micro-video Dataset

Micro-video is an industrial dataset collected from a worldwide micro-video APP, which contains user-item interactions on extensive micro-videos from October 15 to November 14, 2021. This dataset consists of rich side information, especially the content feature of items. The content features of items (i.e., micro-videos) include visual features, such as thumbnails of different resolutions. Besides, each micro-video has diverse textual features, such as titles, descriptions, and tags in multiple languages.

Though Amazon and Kwai are two popular benchmark datasets for recommendation in different domains, they only have visual features for extracting feature representations of cold items. Therefore, we conduct experiments on Micro-video datasets, to validate the effectiveness of TDRO on different modalities of item features.

A.4 Baselines

- **DUIF** (Geng et al. 2015) discards CF representations and solely utilizes feature representations to align with interactions.
- **DropoutNet** (Volkovs, Yu, and Poutanen 2017) leverages feature and CF representations for training interaction predictor, where item CF representations are randomly removed to strengthen the alignment between feature representations and interactions.
- **M2TRec** (Shalaby et al. 2022) is a sequential-based model which integrates a Transformer encoder to learn feature representations, enabling the ability to capture the shifting trend. Besides, it removes CF representations in training and inference.
- **MTPR** (Du et al. 2020) replaces some warm item CF representations with zero vectors for robust learning in a multi-task manner.
- **Heater** (Zhu et al. 2020) utilizes a mixture of experts to extract personalized feature representations, and achieves alignment with interactions by minimizing the distance between feature representations and pre-trained CF representations.
- **CB2CF** (Barkan et al. 2019) introduces a general feature extractor to align feature representations with CF representations learned from interactions via MSE loss.
- **CCFCRec** (Zhou, Zhang, and Yang 2023) encourages the feature extractor to learn robust feature representations by minimizing the distance between feature representations of co-occurring warm items.
- **CLCRec** (Wei et al. 2021) aligns feature and CF representations effectively by maximizing the mutual information between feature and CF representations via contrastive loss.
- **GAR** (Chen et al. 2022) uses adversarial training to bridge the distribution gap between feature and CF representations.
- **InvRL** (Du et al. 2022b) leverages the invariant learning framework, which pays attention to learning invariant feature representations that can achieve robust interaction prediction across different distributions.
- **S-DRO** (Wen et al. 2022) is a method that employs DRO to mitigate the performance gap between user groups in recommender systems. We implement S-DRO on both CLCRec and GAR.
- **DUIF**. The best learning rate is $1e^{-3}$ on the three datasets. The best weight decay is $1e^{-3}$ on Micro-video and Kwai, and $1e^{-4}$ on Amazon.
- **DroupNet**. The dropout ratio is tuned in $\{0.2, 0.5, 0.8\}$. The best dropout ratio is 0.8 on Amazon, and 0.5 on Micro-video and Kwai.
- **M2TRec**. We tune the number of Transformer layers and attention heads in $\{1, 2, 4\}$ and $\{1, 2, 4, 8\}$, separately. Both of the best numbers of Transformer layers and attention heads are 2 on the three datasets.
- **CCFCRec, CLCRec, CB2CF, and Heater**. The coefficient of the auxiliary loss is searched in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. As for CCFCRec, the numbers of positive and negative samples for auxiliary loss are tuned in $\{1, 2, 4\}$ and $\{10, 20, 40\}$, respectively. Besides, we choose the ratio for CLCRec to randomly replaces CF representations with feature representations in $\{0.1, 0.5, 0.9\}$. As for CCFCRec, the best coefficients of the auxiliary loss are 0.5, 0.3, and 0.1 on Amazon, Micro-video, and Kwai, respectively. The best numbers of positive and negative samples are 1 and 10 on the three datasets. As for CLCRec, the best coefficient of auxiliary loss is 0.1 on the three datasets, and the ratio for random replacing is 0.5, 0.9, and 0.3 on Amazon, Micro-video, and Kwai, respectively.
- **GAR**. We separately search the coefficients of adversarial loss and interaction prediction loss in $\{0.1, 0.5, 0.9\}$. The best coefficients of adversarial loss are 0.5, 0.9, and 0.9 on Amazon, Micro-video, and Kwai, respectively. The best coefficients of interaction prediction are 0.6, 0.9, and 0.9 on Amazon, Micro-video, and Kwai, respectively.
- **InvRL**. The number of environments, the coefficient of environment constraint, and the strength of mask regularization are chosen from $\{5, 10, 20, 30\}$, $\{0.1, 0.5, 1\}$, and $\{0.1, 0.5, 1\}$, respectively. The best number of environments, the coefficient of environment constraint, and the strength of mask regularization are 10, 1, and, 0.1, respectively, on the three datasets.
- **S-DRO**. We select the group number K from $\{1, 3, 5, 7, 10\}$. The streaming step size μ and the regularization strength for group importance smoothing η_w are searched in $\{0.1, 0.2, 0.3, 0.5\}$ and $\{0.01, 0.1, 0.2, 0.3, 0.5\}$, respectively. As for the CLCRec backbone model, the best group numbers are 3, 3, and 5 on Amazon, Micro-video, and Kwai, respectively. The best streaming step size and the regularization strength are 0.2 and 0.1, respectively on the three datasets. As for the GAR backbone model, the best group numbers are 5, 3, and 5 on Amazon, Micro-video, and Kwai, respectively. The best streaming step size and the regularization strength are 0.2 and 0.1, respectively on the three datasets.
- **TDRO**. We choose the time period number E , the shifting factor strength λ , and the steepness control factor p for period importance are searched in $\{1, 3, 5, 7, 10\}$, $\{0.1, 0.3, 0.5, 0.7\}$ and $\{0.05, 0.2, 0.5, 1\}$, respectively. The searching scopes for other shared hyper-parameters are consistent with S-DRO. As for the CLCRec backbone

A.5 Hyper-parameters Settings

To ensure a fair comparison, we set the same dimension of CF and feature representations to 128 and adopt a two-layer MLP with a hidden size of 256 as the feature extractor for all methods. Besides, we conduct hyper-parameters tuning, where the best hyper-parameters are selected based on the Recall over the validation set under “all” setting. We tune the learning rate and weight decay in the range of $\{1e^{-4}, 1e^{-3}, 1e^{-2}\}$ and $\{1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}\}$, respectively, for all baselines. And we search other model-specific hyper-parameters with searching scopes and the best ones as follows:

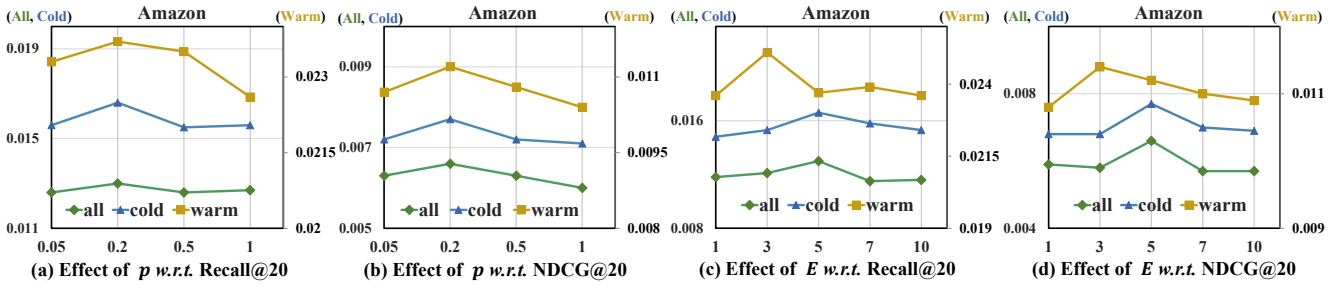


Figure 3: Effect of period importance steepness control factor p for period importance and period number E .

Dataset	#User	#Warm	#Cold	#Int	V	T	Density
Amazon	21,607	75,069	18,686	169,201	64	-	0.01%
Micro-video	21,608	56,712	7,725	276,629	64	768	0.02%
Kwai	7,010	74,470	12,013	298,492	64	-	0.05%

Table 5: Statistics of three datasets. “Int” denotes “Interactions”, “V” and “T” represent the dimension of visual and textual features, respectively.

model, the best K is 5, 3, and 3 on Amazon, Micro-video, and Kwai, respectively. The best E is 5, 3, and 5 on Amazon, Micro-video, and Kwai, respectively. The best λ is 0.3, 0.1, and 0.3 on Amazon, Micro-video, and Kwai, respectively. The best p is 0.2 on the three datasets. As for the GAR backbone model, the best K is 3 on the three datasets. The best E is 3, 3, and 5 on Amazon, Micro-video, and Kwai, respectively. The best λ is 0.9, 0.1, and 0.3 on Amazon, Micro-video, and Kwai, respectively. The best p is 0.2 on the three datasets.

The statistics of three datasets for experiments are summarized in Table 5.

A.6 Additional Analysis of Experimental Results

Overall performance. From Table 1, we may find that as for potential solutions to item feature shifts, InvRL surpasses its backbone model (CLCRec) on cold performance whereas it sacrifices the accuracy of warm item recommendation. This is reasonable since InvRL captures the invariant part of CF and feature representations, which encourages robust cold item recommendation. Nevertheless, it overlooks the variant part, which may be essential for interaction prediction for warm items. In contrast, M2TRec, which leverages the sequential shifting pattern, usually has inferior performance compared to other baselines. The reason might be that it discards both user and item CF representations for prediction, thus being severely affected by the negative impact of feature noise for capturing sequential patterns.

Ablation study. In Table 2, the inferior cold performance of removing the worst-case factor compared to CLCRec is probably due to the different dataset pre-processing for Kwai, where the interactions are randomly split into the training, validation, and testing sets. As such, the shifting trends of Kwai might be less significant (*cf.* Section 5.1).

A.7 Hyper-parameter Analysis

Effect of period importance. To analyze the impact of period importance $\beta_e = \exp(p \cdot e)$ on capturing shifting trend, we vary the steepness control factor p from 0.05 to 1 and report the results on Amazon in Figure 3(a) and (b). The results with similar observations on Micro-video and Kwai are omitted to save space. From the figures, we can find that 1) the performance increases as we enlarge p . This is because a smaller p intends to pay attention to each time period uniformly whereas a larger p encourages to emphasize the later time periods closer to the test phase, pushing the model to capture the shifting trend of item features of subsequent cold items. 2) Due to that only the last time period is considered in TDRO as p approaches infinity, blindly increasing p overlooks other relatively earlier time periods containing useful information to capture shifting patterns, thus hurting the learning of cold-start recommender models.

Effect of time period number. We vary the time period number from 1 to 10 and present the results in Figure 3(c) and (d), where we can find that: 1) Slightly increasing time period numbers yields better performance. This is because a larger time period number leads to a more fine-grained time period division, facilitating TDRO to capture the shifting trend of item features accurately. Nevertheless, 2) the time period number cannot be too large as it will exacerbate the interaction data sparsity for each time period and cause training instability, thus degrading the performance.

Effect of shifting trend strength. We inspect the effect of the shifting factor by changing λ from 0.1 to 0.9. As shown in Figure 4, it is observed that stronger incorporation of the shifting trend typically intends to yield better performance on cold items, indicating the importance of considering shifting patterns in robustness enhancement over cold items. However, the all and warm performance declines if we consider the shifting factor too much (*i.e.*, worst-case factor is weak), which is probably due to the overlook of the minority group of warm items. Therefore, we should carefully choose λ to balance the trade-off between the worst-case group and the groups that are likely to become popular in cold items.

Effect of group number. We change K from 1 to 10, with the results reported in Figure 5. We can observe that the performance achieves the best when $K = 5$ under “all”,

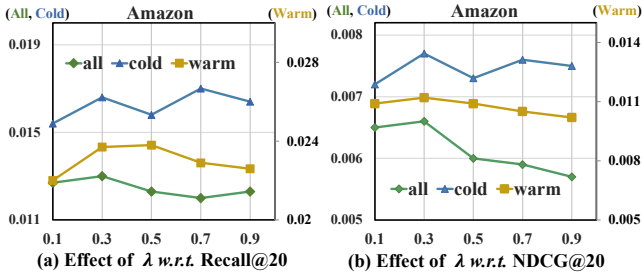


Figure 4: Effect of the strength of shifting factor λ .

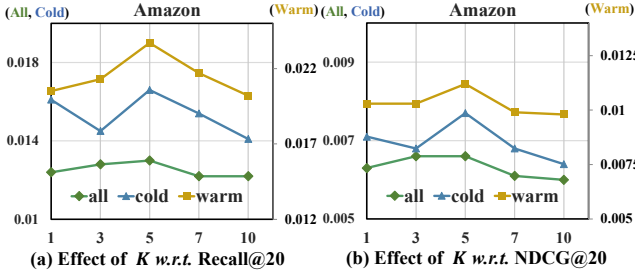


Figure 5: Effect of group number K .

“warm”, and “cold” settings. The fluctuating performance from $K = 1$ to $K = 10$ indicates that the performance is not proportional to the group number. This makes sense since it is non-trivial to recover the group-level preference aligning well with real-world scenarios (e.g., a variety of clothing). A more efficient way of choosing an appropriate group number from interactions can be explored in future works.

A.8 Detailed Related Work

Cold-start Recommendation. Traditional CF methods typically rely on historical user-item interactions to learn CF representations for interaction prediction (Wang et al. 2021b). However, the influx of cold items obstructs traditional CF methods from providing appropriate recommendations due to the lack of historical interactions (Zhao et al. 2022; Rajapakse and Leith 2022; Raziperchikolaei, Liang, and Chung 2021; Pulis and Bajada 2021; Du et al. 2022a), escalating imbalanced impressions on the item side (Huan et al. 2022; Zhu et al. 2021; Sun et al. 2021; Wang et al. 2021a; Chu et al. 2023). Hence, it is crucial to enhance the cold-start item recommendation (Houlsby, Hernández-Lobato, and Ghahramani 2014; Pan et al. 2019; Vartak et al. 2017; Neupane et al. 2022; Wei et al. 2020).

Existing methods address the cold-start problem by aligning the extracted feature representations with user-item interactions (Meng et al. 2020; Shi et al. 2018; Guo et al. 2017; Hao et al. 2021; Togashi, Otani, and Satoh 2021), typically falling into two research lines: 1) Robust training-based methods, where both feature and CF representations are utilized to predict interactions while CF representations are randomly corrupted to encourage robust alignment between feature representations and interactions. For example, (Volkovs, Yu, and Poutanen 2017) randomly drops CF

representations of warm items for robust training. 2) Auxiliary loss-based methods achieve alignment by introducing different auxiliary loss for minimizing the distance between feature representations and CF representations learned from interactions. For example, (Wei et al. 2021) maximizes the mutual information between feature and CF representations spaces via contrastive loss, and (Chen et al. 2022) utilizes adversarial training to effectively bridge the gap between feature and CF spaces.

However, previous methods suffer from temporal feature shifts from warm items to cold items, where the feature representations of cold items may not be well captured by the feature extractor learned from warm items. In this work, we highlight the importance of strengthening the generalization ability of the feature extractor against temporal feature shifts, where the key lies in considering temporally shifted distributions that reflect the shifting trend of item features to achieve robust interaction prediction.

Distributionally Robust Optimization. DRO is an effective approach that aims to achieve uniform performance against distribution shifts by optimizing the worst-case performance over a pre-defined uncertainty set (Rahimian and Mehrotra 2019; Michel, Hashimoto, and Neubig 2022). The most representative line of research works lies in discrepancy-based DRO that defines the uncertainty set as a ball surrounding the training distributions with different discrepancy metrics (e.g., f-divergence (Duchi and Namkoong 2018), MMD ball (Staib and Jegelka 2019), and Wasserstein Distance (Liu et al. 2022)). For example, (Zhai et al. 2021) leverages the Cressie-Read family of Rényi divergence to define the uncertainty set, and (Liu et al. 2022) is developed based on Wasserstein Distance. Due to the fact that discrepancy-based DRO unnecessarily considers implausible distributions (i.e., over-pessimism issue (Oren et al. 2019; Sagawa et al. 2020; Duchi, Hashimoto, and Namkoong 2023)), another line falls into Group-DRO (Zhou et al. 2021; Goel et al. 2021). Works along this line define the uncertainty set as a set of mixtures of subgroups in the training set, which encourages DRO to focus on meaningful distribution shifts. For instance, (Oren et al. 2019) focuses on the distribution shifts between topics for natural language modeling, and (Wen et al. 2022) optimizes the worst-case performance over different user groups in recommender systems. Recently, some works have been proposed to define the uncertainty set based on generative models (Michel, Hashimoto, and Neubig 2021), which inevitably introduces extra parameters, resulting in a burden of high computational costs.

However, applying DRO directly to cold start recommendations has inconsistency issues. The overemphasis on the minority group of warm items may weaken the expressiveness of the groups that are likely to become popular in the upcoming cold items, thus limiting the robustness enhancement for cold item recommendation. In this work, we propose to leverage the temporal shifting trend to guide DRO to improve the generalization ability of the feature extractor against temporal feature shifts.

Other literature on robustness has also been extensively studied. Invariant learning (Arjovsky et al. 2019; Du et al. 2022b; Koyama and Yamaguchi 2020; Liu et al. 2021) considers the invariant part robust to distribution shifts but it overlooks the variant part, which may be essential for prediction. Re-weighting loss (Zhang et al. 2021; Kim et al. 2021; Yang et al. 2023) assigns weights to samples, which however relies heavily on correlations between group density and task difficulty and yields inferior performance than DRO (*cf.* Table 1 in (Wen et al. 2022)).