Temporally and Distributionally Robust Optimization for Cold-Start Recommendation

Xinyu Lin[1], Wenjie Wang[1*], Jujia Zhao[1], Yongqi Li[2], Fuli Feng[3], Tat-Seng Chua[1]

[1]National University of Singapore
[2]The Hong Kong Polytechnic University
[3]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
{xylin1028, wenjiewang96, 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.

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 items[1]). For example, 500 hours of video are uploaded to YouTube every minute[2]. 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 items (Wang et al. 2023b). As illustrated in Figure 1(a), the category features of newly-uploaded items are

*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.

[1]For simplicity, cold-start items and warm-start items are referred to as cold and warm items, respectively.
[2]https://www.statista.com/.

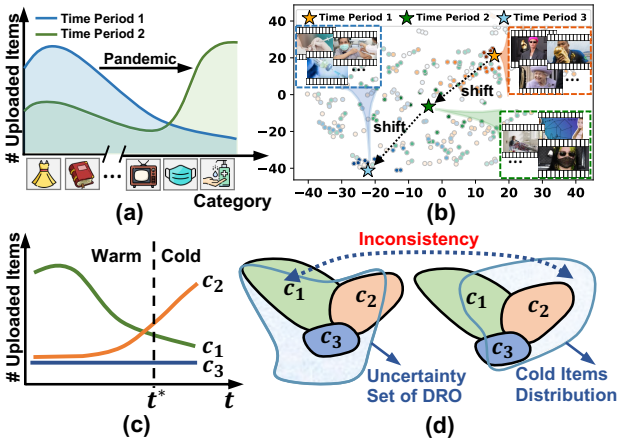


Figure 1: (a) An example of item category feature shifts towards sanitary products. (b) T-SNE visualization of visual features of item thumbnails in three time periods on a Micro-video dataset. The stars represent the average item features in each time period. (c) An example of the shifting trend of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

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 approach[3]. 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 groups[4] in warm items at the expense of other groups' performance (Oren et al. 2019). Due to the fact 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 **T**emporal **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 emphasize the vulnerability of ERM and underscore the necessity of adjusting the learning objective to strengthen the generalization ability of cold-start models under temporal feature shifts.
- 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.

Related Work

- **Cold-start Recommendation.** Traditional CF methods typically rely on CF representations learned from historical interactions (Wang et al. 2022; Li et al. 2019; Sun et al. 2022). However, the influx of cold items hinders traditional

[3]Other potential solutions are discussed in Section .

[4]Minority group usually yields worse performance in recommendation (Wen et al. 2022). In DRO, the training distribution is assumed to be a mixture of subgroups, and the uncertainty set is defined on mixtures of these subgroups (*cf.* Section).

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. 2021; 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. To solve this issue, a concurrent study (Wang et al. 2023b) explores equivariant learning over minority groups of warm items. Differently, we leverage the shifting trend and emphasize the optimization of the potentially popular item groups.

- **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) and 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.

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 $\{\boldsymbol{s}_i | i \in \mathcal{I}_w\}$, 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

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

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, DRO[5] 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 $\{P_i | i = 1, \ldots, K\}$. Then, it optimizes the worst-case performance over the K subgroups for controlling the performance lower bound. Formally,

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

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:

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

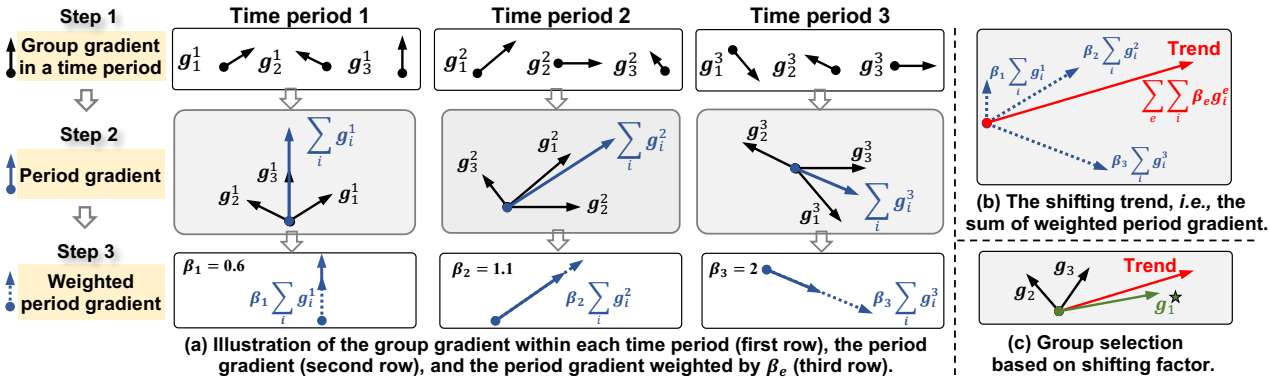
where $\bar{\mathcal{L}}_j = \frac{1}{N_j} \sum_{(u,i,y_{ui}) \sim \tilde{P}_j} \mathcal{L}_j(\theta; (u, i, y_{ui}, \boldsymbol{s}_i))$, $\tilde{P}_j$ is the empirical distribution of group j in dataset $\mathcal{D}$, and N_j is the number of samples in group j. Subsequently, the model parameters θ 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).

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,

[5]We adopt Group-DRO to avoid over-pessimism issue.



(a) Illustration of the group gradient within each time period (first row), the period gradient (second row), and the period gradient weighted by β_e (third row).

(b) The shifting trend, *i.e.*, the sum of weighted period gradient.

(c) Group selection based on shifting factor.

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).

and 2) capturing the temporal shifting trend of item features and emphasizing the optimization of groups that are likely to become popular.

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, $e \in \{1, \ldots, E\}$. 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

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

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.

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

where $\boldsymbol{g}_j = \nabla_\theta \bar{\mathcal{L}}_j(\theta)$ denotes the gradient of the average loss of group j, and $\boldsymbol{g}_i^e = \nabla_\theta \mathcal{L}_i^e(\theta)$ denotes the gradient of group i's average loss in time period e. The $\langle \cdot, \cdot \rangle$ represents the inner product computation. Since $\sum_{e=1}^{E}\sum_{i=1}^{K} \beta_e \boldsymbol{g}_i^e$ 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, $\boldsymbol{g}^e_{i\in\{1,2,3\}}$, for each time period e. We can then obtain the period gradient $\sum_{i=1}^{K} \boldsymbol{g}_i^e$ 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 $\sum_{e=1}^{E}\sum_{i=1}^{K}\beta_e \boldsymbol{g}_i^e$ 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 $\boldsymbol{g}_j$ (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 $\theta^{t+1} = \theta^t - \eta\nabla_\theta \bar{\mathcal{L}}_{j^*}(\theta^t)$.

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 $\boldsymbol{w}$ for groups and regulate the weight dynamic by η_w. Formally,

$$\boldsymbol{w}^{t+1} = \underset{w_{i\in[K]}}{\arg\max} \sum_i w_i[(1-\lambda)\bar{\mathcal{L}}_i(\theta) + \lambda\langle \boldsymbol{g}_i, \sum_{e=1}^{E}\sum_{j=1}^{K}\beta_e \boldsymbol{g}_j^e\rangle] - \frac{1}{\eta_w}\mathrm{KL}(\boldsymbol{w}, \boldsymbol{w}^t), \tag{6}$$

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

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

Thereafter, the model parameters θ are updated through

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

• **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.

Algorithm 1: Training Procedure of TDRO

Input: Number of groups K, number of time periods E, initial model parameters θ^0, initial group weight $\boldsymbol{w} = (\frac{1}{K}, \frac{1}{K}, \ldots, \frac{1}{K})$, initial group loss $\bar{\mathcal{L}}^0_{i\in[K]}$, item features $\{\boldsymbol{s}_i | i \in \mathcal{I}_w\}$, 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.

while not converge **do**
  for all $i \in \{1, \ldots, K\}$ **do**
    Calculate $\bar{\mathcal{L}}_i^t(\theta^t)$ via cold-start loss function.
    $\bar{\mathcal{L}}_i^t(\theta^t) \leftarrow (1-\mu)\bar{\mathcal{L}}_i^{t-1}(\theta^{t-1}) + \mu\bar{\mathcal{L}}_i^t(\theta^t)$
  for all $i \in \{1, \ldots, K\}$ **do**
    $w_i^{t+1} \leftarrow w_i^t \exp(\eta_w[(1-\lambda)\bar{\mathcal{L}}_i^t(\theta^t) + \lambda(\nabla\bar{\mathcal{L}}_i^t(\theta^t)\sum_{e=1}^{E}\sum_{j=1}^{K}\beta_e\nabla\bar{\mathcal{L}}_j^{e,t}(\theta^t))])$
  $w_i^{t+1} \leftarrow w_i^{t+1}/\|\boldsymbol{w}^{t+1}\|_1, \forall i \in \{1, \ldots, K\}$ ▷ Normalize
  $\theta^{t+1} \leftarrow \theta^t - \eta\sum_{i\in[K]} w_i^{t+1}\nabla\bar{\mathcal{L}}_i^t(\theta^t)$ ▷ Update

Output: Optimized model parameters θ.

• **Instantiation.** To instantiate TDRO on cold-start recommender models, we first calculate the group weight $\boldsymbol{w}$ 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.

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?

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 recommendation 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

[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.

and cold sets, where items that do not appear in the training set are regarded as cold items, and the rest as warm items.

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; Pan et al. 2023).

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. Meanwhile, auxiliary loss-based methods decouple the CF and feature representations space, which protects the CF representations from feature noises.

• CLCRec consistently yields impressive performance across the three datasets. This is attributed to the contrastive loss, which maximizes the mutual information between feature and CF representations. 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 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, which justifies the effectiveness of TDRO. 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.

Methods	Amazon All	Amazon Warm	Amazon Cold	Micro-video All	Micro-video Warm	Micro-video Cold	Kwai All	Kwai Warm	Kwai Cold
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 *w.r.t.* Recall@20. The best results are highlighted in bold.

	Amazon G1	Amazon G2	Amazon G3	Micro-video G1	Micro-video G2	Micro-video G3
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.

	All Worst-case	All Popular	Cold Worst-case	Cold 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.

In-depth Analysis

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 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). 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.

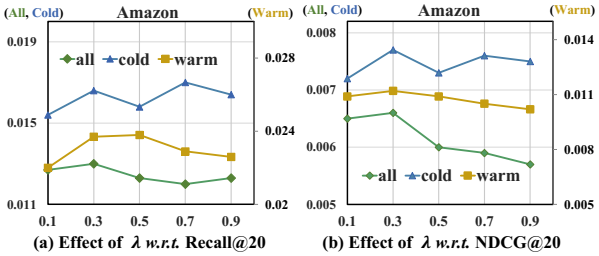


Figure 3: Effect of the strength of shifting factor λ.

Effect of shifting trend strength. We inspect the effect of the shifting factor by changing λ from 0.1 to 0.9. Stronger incorporation of shifting trend intends to yield better performance on cold items as shown in Figure 3, indicating the importance of shifting patterns in robustness enhancement. However, the all and warm performance declines if we consider the shifting factor too much, which is probably due to the overlook of the minority group of warm items.

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. It is also promising to leverage LLM for cold-start recommendation (Wang et al. 2023a; Bao et al. 2023b,a).

References

Bao, K.; Zhang, J.; Wang, W.; Zhang, Y.; Yang, Z.; Luo, Y.; Feng, F.; He, X.; and Tian, Q. 2023a. A bi-step grounding paradigm for large language models in recommendation systems.

Bao, K.; Zhang, J.; Zhang, Y.; Wenjie, W.; Feng, F.; and He, X. 2023b. Large Language Models for Recommendation: Progresses and Future Directions. In *SIGIR-AP*, 306–309.

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.

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.

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.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

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.; 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. 2022. Distributionally robust models with parametric likelihood ratios. In *ICLR*.

Oren, Y.; Sagawa, S.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust language modeling. *arXiv:1909.02060*.

Pan, H.; Chen, J.; Feng, F.; Shi, W.; Wu, J.; and He, X. 2023. Discriminative-invariant representation learning for unbiased recommendation. In *IJCAI*, 2270–2278.

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.

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.

Vapnik, V. 1991. Principles of risk minimization for learning theory. In *NeurIPS*, 831–838. 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. 2021. Privileged graph distillation for cold start recommendation. In *SIGIR*, 1187–1196. ACM.

Wang, W.; Lin, X.; Feng, F.; He, X.; and Chua, T.-S. 2023a. Generative recommendation: Towards next-generation recommender paradigm.

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.

Wang, W.; Lin, X.; Wang, L.; Feng, F.; Wei, Y.; and Chua, T.-S. 2023b. Equivariant Learning for Out-of-Distribution Cold-start Recommendation. In *MM*, 903–914.

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.

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

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 Recommnedation. 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.