Model-Agnostic Causal Embedding Learning for Counterfactually Group-Fair Recommendation

Xiao Zhang, Teng Shi, Jun Xu, *Member, IEEE*, Zhenhua Dong, and Ji-Rong Wen, *Senior Member, IEEE*

***Abstract*—Group-fair recommendation aims at ensuring the equality of recommendation results across user groups categorized by sensitive attributes (e.g., gender, occupation, etc.). Existing group-fair recommendation models traditionally employ original user embeddings for both training and testing, primarily focusing on statistical learning while imposing group fairness constraints under the I.I.D. assumption. However, these models encounter limitations when addressing out-of-distribution (OOD) sensitive attributes. The fundamental issue of unfairness within user embeddings arises from a causal perspective, where each embedding vector comprises an exogenous component devoid of correlations with sensitive attributes and an endogenous component strongly correlated with these attributes. Overlooking the distinction between these two components during model training renders models sensitive to shifts in the distribution of sensitive attributes. This paper introduces the concept of Counterfactual Group Fairness (CGF) along with a corresponding metric to evaluate group fairness in scenarios involving OOD sensitive attributes in recommender systems. Building on this foundation, we propose a model-agnostic causal embedding learning framework named MACE. MACE effectively disentangles user embedding vectors into their exogenous and endogenous parts, thus ensuring group fairness, even in the presence of OOD sensitive attributes in embeddings. Specifically, MACE identifies the exogenous part of each user's embedding using mutual information minimization, treating it as instrumental variables. Subsequently, under the constraint of CGF, MACE reconstructs the endogenous and exogenous parts using the instrumental variable regression, combines the obtained parts into novel user embeddings using deep neural networks, and uses the combined embeddings for fair recommendation. Experimental results demonstrated that MACE can outperform the state-of-the-art baselines in terms of the metric of CGF while maintaining a comparable recommendation accuracy.**

***Index Terms*—Causal learning, fairness, instrumental variables, recommendation.**

I. Introduction

In recent years, the proliferation of recommender systems has raised concerns that the learned recommendation models may be discriminatory with respect to sensitive attributes such as gender, occupation, age, i.e., the issue of group unfairness in recommendation [1]. Group unfairness occurs when recommender systems deliver varying levels of recommendation quality to different user groups defined by these sensitive attributes, leading to potential biases and inequalities. Although deep learning methods for recommender systems can extract abstract representation into embeddings for accurate prediction [2], [3], [4], the user embeddings learned by deep neural networks often contain or are related to sensitive attributes, ultimately compromising group fairness in predicting users' feedback.

Motivated by the differences of the testing and training distributions that widely exist in real-world data [5], [6], [7], [8], [9], it is necessary to ask a counterfactual problem: what the group fairness of a recommendation model would be if the distribution of sensitive attributes on testing data is even slightly different from that on training data? (e.g., how the group fairness metric of a recommendation model would change, if the distribution of users' occupations is replaced by a new distribution on testing data, see Fig. 1).

To validate the presence of the unfairness issue mentioned in the last paragraph, we focus on collaborative filtering (CF) methods for recommendation, which utilize the known preferences of a group of users to predict additional items a new user might like. Specifically, we analyze DMF [2], a classic CF method using neural network-based matrix factorization, in the context of group fairness when there are shifts in the distribution of sensitive attributes in the testing data. For the group fairness metric, we selected ΔDP, which measures the difference in model outputs between different groups defined by sensitive features. In short, ΔDP shows how much a recommendation model's predictions vary between different user groups. A smaller value of ΔDP indicates that the model treats different groups more equally, demonstrating better group fairness. From the empirical results depicted at the top of Fig. 2, we can infer that the metric ΔDP of DMF [2] for movie recommendation is sensitive to the shifts in the distribution of sensitive attributes (occupations) within the testing dataset, where the changed distributions of occupations are illustrated at the bottom of Fig. 2. This effect occurs despite the model being trained and tested without consideration of sensitive attributes. These observations motivate a deeper investigation into achieving group-fair recommendations in scenarios where out-of-distribution (OOD)

Manuscript received 2 September 2023; revised 22 June 2024; accepted 1 July 2024. Date of publication 15 July 2024; date of current version 13 November 2024. This work was supported in part by the National Science and Technology Major Project under Grant 2022ZD0120103, in part by the National Natural Science Foundation of China under Grant 62376275 and under Grant 62377044, in part by Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. Supported in part by fund for building world class universities (disciplines) of Renmin University of China and in part by Public Computing Cloud, Renmin University of China. Recommended for acceptance by M. A. Cheema. *(Xiao Zhang and Teng Shi contributed equally to this work.) (Corresponding author: Jun Xu.)*

Xiao Zhang, Teng Shi, Jun Xu, and Ji-Rong Wen are with the Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China (e-mail: zhangx89@ruc.edu.cn; shiteng@ruc.edu.cn; junxu@ruc.edu.cn; jrwen@ruc.edu.cn).

Zhenhua Dong is with the Huawei Noah's Ark Lab, Shenzhen 518129, China (e-mail: dongzhenhua@huawei.com).

Digital Object Identifier 10.1109/TKDE.2024.3424906

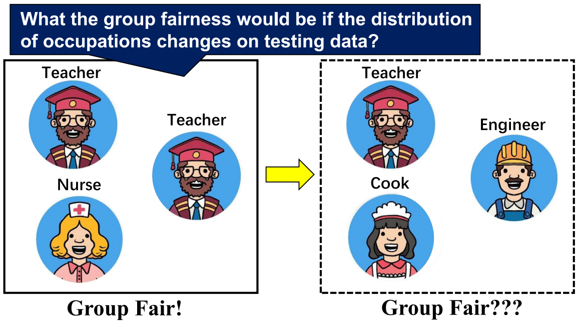
What the group fairness would be if the distribution of occupations changes on testing data?
Teacher
Teacher
Nurse
Group Fair!
Teacher
Engineer
Cook
Group Fair???

Fig. 1. Example of group-fair recommendation when the distribution of occupations (sensitive attributes) changes on testing data. For example, when a teacher working in a university resigns to pursue a career as an engineer in the industry, the distribution of the user's occupation (sensitive attribute) changes. Whether the recommendation algorithm can still ensure group fairness in recommendations in such scenarios remains unexplored.

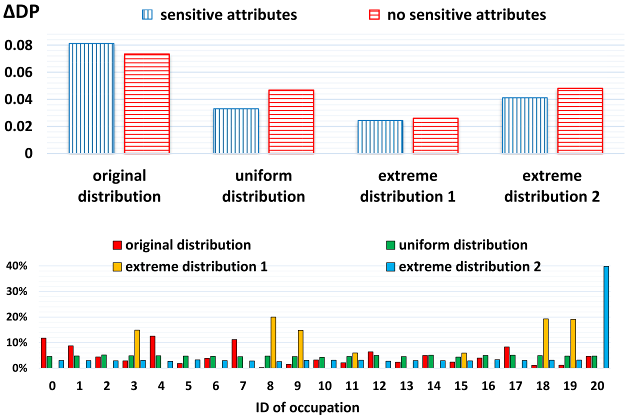

Fig. 2. Empirical study of group fairness on MovieLens-1 M with OOD sensitive attributes. Top: Group fairness on different testing distributions of the occupations (original, uniform, and two extreme distributions), where ΔDP defined in (1) is used to evaluate the group fairness that is a relaxed metric of demographic parity (DP) (a smaller ΔDP indicates a more group-fair model). The results are obtained by DMF [2] trained on the original training data and two versions of DMF are used for testing: using sensitive attributes (vertical stripe bars) and omitting sensitive attributes (horizontal stripe bars) for training and testing. Bottom: Percentage of occupations in different testing distributions.

sensitive attributes arise due to changes in user demographics. This concept is termed as "counterfactually group-fair recommendation" (CGFR), aiming to recommend group-fair results while testing with varying distributions of sensitive attributes compared to those encountered during training.

Recent works focused on achieving group fairness in designing recommendation and ranking algorithms [10], [11], which typically incorporated group fairness constraints into the objective function under the I.I.D. assumption. In addition to considering group fairness, individual fair learning aims to give similar predictions to similar individuals under some similarity metric [12], [13]. To identify and utilize the causal association between attributes and labels, several studies provided metrics of causally motivated individual fairness, called counterfactual individual fairness [14], [15]. But in group-fair recommendation literature, little efforts have been made to consider the causal association between different parts of user embeddings and the labels, and ignored the group fairness in counterfactual worlds where the distribution of sensitive attributes may shift in a causal sense.

Looking into the CGFR problem from the view of causal analysis, even if the sensitive attributes are removed, a user embedding may still contain an endogenous part that is correlated to the sensitive attributes. This observation has been demonstrated in empirical results shown in Fig. 2, where the recommendation model that does not use the sensitive attributes as input is still sensitive to the distribution shift of the sensitive attributes. Since the sensitive attributes usually have a correlation with the user feedback (e.g., user clicks), sensitive attributes can be treated as confounders that confound the causal association between the user embedding and the user feedback. Directly using the mixed user embedding for training and testing may make the recommendation model vulnerable to the OOD sensitive attributes.

To evaluate and improve the group fairness when OOD sensitive attributes exist, we define a notion called counterfactual group fairness (CGF) as well as its metric and propose a model-agnostic causal embedding learning framework to achieve CGF in recommendation. Our definition of CGF captures the intuition that a recommendation model is fair towards different groups if its group fairness metric remains the same in the actual world and a counterfactual world where the distribution of sensitive attributes changes. From a causal view, we treat the sensitive attributes as confounders, and extract the exogenous part and endogenous part of each user embedding under the constraint of CGF. Then, the exogenous part is used to fit the endogenous part using the instrumental variable (IV) regression approach. In this way, the user embedding can be effectively represented by two orthogonal vectors: a causal representation (the values fitted by the IV regression) and a non-causal representation (the regression residuals). Finally, these two parts are combined into a new user embedding using deep neural networks and fed into the recommendation model for prediction.

We summarize the major contributions of the paper:

- A counterfactual group fairness metric to measure the group fairness for recommendation when OOD sensitive attributes exist;
- A novel model-agnostic framework called MACE for learning counterfactually group-fair user embeddings in recommendation using causal learning motivated by instrumental variable method;
- Comprehensive empirical studies showed the effectiveness of MACE in terms of improving the fairness of different recommendation models and its superiority over state-of-the-art baselines.

II. Related Work

Fairness in machine learning has received considerable attention over the past several years, where the metrics of fairness can be group-level [10], [11], individual-level [12], [16], causal [14], [15], [17], etc. Fair learning algorithms can be divided into three categories [18], [19]: (a) Pre-processing algorithms modify the training data for mitigating the effects of underlying discrimination in the data collection policy [20], [21], [22]. Calmon et al. [22] designed a pre-processing transformation approach that trades off group fairness, data utility, and individual distortion. (b) In-processing algorithms incorporate different fairness constraints into the inductive bias of the learning algorithms during training [10], [23], [24], [25]. Mandal et al. [25] proposed a fair online learning approach that is robust to perturbations in the training distribution, where the fair classifiers were obtained by a re-weighted objective function under group fairness constraints. (c) Post-processing algorithms only access to the predictions and sensitive attributes and modify the unfair predictions to improve the prediction fairness [26], [27], [28]. Hardt et al. [28] provided a post-learning framework by constructing a non-discriminating predictor requiring only the learned binary predictor and the aggregate information about the data. Our fair learning framework focuses on group fairness in a causal sense, and draws on the strengths of both the pre-processing and in-processing algorithms, which simultaneously reconstructs user embeddings in training data and incorporates a novel counterfactual group fairness constraint.

Causal machine learning explores modeling approaches in the intersection of machine learning and causal inference [29], [30]. In terms of technique, our embedding reconstruction approach is mostly related to causal representation learning that aims to learn variables and their generation processes in causal graphs [31]. Existing causal representation learning approaches can mitigate bias in training data and help to resist the interventions that change the joint distribution of the variables of interest. Kuang et al. [32] designed a learning-based causal approach to automatically separating confounders and adjustment variables in embeddings, which can be applied to causal effect estimation. interventions that change the joint distribution of the variables of interest. Zheng et al. [33] learned disentangled embeddings for capturing the interest and conformity of users in recommendation, which were trained on cause-specific data from observational interactions. Liu et al. [34] proposed an embedding decomposition method using information bottleneck that learned the biased and unbiased components of an embedding in training, where only the unbiased component was used as input for testing to achieve more accurate recommendation results. Our causal embedding learning framework treats the sensitive attributes as confounders and reconstructs the user embeddings under a new group fairness constraint, which mitigates the confounding bias using an approach motivated by instrumental variables.

Fairness in recommendation has been extensively explored and can be primarily categorized into group fairness and individual fairness. Existing works on group fair recommendation focus on measuring and achieving group fairness under the assumption of an independent and identically distributed (i.i.d.) setting. Yao et al. [35] defined four group fairness metrics in collaborative filtering recommender systems. Li et al. [10] addressed the user-oriented group fairness problem in commercial recommender systems. Rahmani et al. [36] illustrated that the disparity in recommendation accuracy among user groups may vary across different datasets. Zhao et al. [37] learned fair representations from the perspective of mutual information to achieve group fairness in recommendations. Lin et al. [38] formulated the problem of group recommendation as a multiple objective optimization problem and provided a optimization framework to achieve Pareto efficiency. Unlike the above methods for group fairness in recommendations, this paper focuses on group fairness and its metrics and implementation in non-i.i.d. settings. On the other hand, individual fairness posits that similar individuals should receive comparable treatment. Li et al. [15] delved into counterfactual fairness in recommendations, aiming for consistency in recommendation results for each user between the factual and counterfactual worlds. Huang et al. [39] focused on user-side individual fairness for customers in online recommendation, and proposed a fair causal bandit approach for achieving counterfactual individual fairness. Unlike existing works on individual fairness, which often focus on counterfactual individual fairness, this paper focuses on counterfactual group fairness (CGF) in an out-of-distribution (OOD) setting. When there is only one user in each user group, the CGF problem discussed in this paper degenerates into a user-side individual fairness problem in an OOD setting.

III. Problem Formulation

This section introduces the problem of sensitive attributes in user embeddings from a causal view, and define a novel notion called counterfactual group fairness and its metric.

A. Causal View of Sensitive Attributes

When a user $u \in \mathcal{U}$ accesses a recommender system, the system provides a list of items $i \in \mathcal{I}$ with a recommendation model f. Users and items are typically represented by real-valued vectors (i.e., embeddings), denoted by $\mathbf{v}_u \in \mathbb{R}^{d_u}$ and $\mathbf{v}_i \in \mathbb{R}^{d_i}$, respectively, where d_i and d_u are the dimensions of the embeddings for users and items. Although deep learning models can extract abstract representation into embeddings for accurate prediction, user embeddings usually contain sensitive attributes that may hurt the fairness in recommendation tasks. More specifically, a user embedding contains two parts $\mathbf{v}_u = [\mathbf{s}_u^\intercal, \mathbf{x}_u^\intercal]^\intercal$: the sensitive attributes denoted by $\mathbf{s}_u \in \mathcal{S} \subseteq \mathbb{R}^{d_s}$ (gender, occupation, etc.) and the remaining attributes denoted by $\mathbf{x}_u \in \mathbb{R}^{d_x}$ (may also correlate with $\mathbf{s}_u$), where $\mathcal{S}$ denotes the space of sensitive attributes which is a set consisting of a finite number of vectors, d_s is the dimension of the sensitive attributes and $d_x = d_u - d_s$ is the dimension of $\mathbf{x}_u$. In deep learning-based recommender systems, the recommendation model f is trained with the user-system interaction histories and the training dataset can be represented by $\mathcal{D}_{\text{train}} = \{(\mathbf{s}_u, \mathbf{x}_u, \mathbf{v}_i, y_{u,i})\}$ that is drawn from an unknown distribution $\mathbb{P}$. Each tuple $(\mathbf{s}_u, \mathbf{x}_u, \mathbf{v}_i, y_{u,i}) \in \mathcal{D}_{\text{train}}$ means that the item i was exposed to the user u and the interaction was $y_{u,i} \in \{0, 1\}$, where $y_{u,i}$ denotes the user feedback (e.g, $y_{u,i} = 1$ means clicked and $y_{u,i} = 0$ otherwise).

Following the causal inference framework [40], a causal graph can be constructed for recommendation when sensitive attributes exist in user embeddings (see the left side of Fig. 3). Specifically, since both the user embedding and item embedding have an influence on the user feedback, we treat the user and item embeddings $\mathbf{v}_u$ and $\mathbf{v}_i$ as the treatment, and the user feedback $y_{u,i}$ as the outcome. Moreover, the sensitive attributes $\mathbf{s}_u$of user

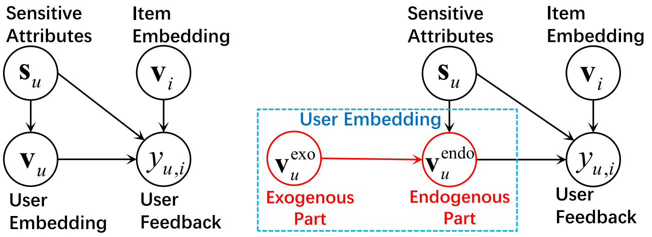

Fig. 3. Left: Causal graph of recommendation when sensitive attributes exist in user embeddings. Right: Decomposition of user embedding in the proposed MACE.

u can be seen as confounders that confound the causal association between the user embedding $\mathbf{v}_u$ (treatment) and the user feedback $y_{u,i}$ (outcome), where the path $\mathbf{s}_u \to \mathbf{v}_u$ exists since $\mathbf{s}_u$ is part of $\mathbf{v}_u$, and the path $\mathbf{s}_u \to y_{u,i}$ exists since user profiles usually contribute to the prediction in recommendation. Thus, directly using the confounded user embedding for training and testing will make the recommendation model unfair especially in the case that OOD sensitive attributes exist in testing data. As shown on the right side of Fig. 3, in this paper, the proposed MACE extracts two parts from the original user embedding: an *exogenous part* that satisfies the unconfounded condition of instrumental variables (i.e., has no correlation with the confounders $\mathbf{s}_u$) [41], [42], and an *endogenous part* constructed under a novel fairness constraint that may be correlated to $\mathbf{s}_u$. Motivated by instrumental variable (IV) methods [43], [44], by regressing the endogenous part on the exogenous part, we can reconstruct the user embeddings by balancing the fairness and recommendation accuracy in an end-to-end manner. Using the reconstructed user embeddings, we can perform group-fair recommendation without obvious accuracy degradation even when OOD sensitive attributes exist in a counterfactual world. Compared with counterfactual learning methods for debias such as inverse propensity score [45] and doubly robust estimator [46], IV methods can handle high-dimensional treatment variables (e.g., user's embedding vectors).

B. Counterfactual Group Fairness (CGF)

The most widely used definitions of group fairness are Demographic Parity (DP) and Equalized Odds (EO) [47], which enforce the statistical independence between the model prediction and the sensitive attributes (EO requires this independence condition holds conditioned on the user feedbacks). Different from the DP distance and EO distance defined on the training distribution $\mathbb{P}$ for two groups [23], [24], given a testing dataset $\mathcal{D}_{\text{test}}$ drawing from any distribution, we introduce the following relaxed metrics of group fairness for multiple user groups (corresponding to different sensitive attributes): for a recommendation model f, letting $\hat{y}_{u,i} = f(\mathbf{s}_u, \mathbf{x}_u, \mathbf{v}_i)$ be the prediction of a user's preference score,

$$\begin{aligned}\Delta\text{DP}(f, \mathcal{D}_{\text{test}}) &= \sup_{\mathbf{s}_1,\mathbf{s}_2 \in \mathcal{S}, \mathbf{s}_1 \neq \mathbf{s}_2} \left| \frac{\sum_{u \in \mathcal{U}_{\mathbf{s}_1}} \hat{y}_{u,i}}{|\mathcal{U}_{\mathbf{s}_1}|} - \frac{\sum_{u \in \mathcal{U}_{\mathbf{s}_2}} \hat{y}_{u,i}}{|\mathcal{U}_{\mathbf{s}_2}|} \right|, \\ \Delta\text{EO}(f, \mathcal{D}_{\text{test}}) &= \sup_{\mathbf{s}_1,\mathbf{s}_2 \in \mathcal{S}, \mathbf{s}_1 \neq \mathbf{s}_2} \sum_{y \in \{0,1\}} \left| \frac{\sum_{u \in \mathcal{U}^y_{\mathbf{s}_1}} \hat{y}_{u,i}}{|\mathcal{U}^y_{\mathbf{s}_1}|} - \frac{\sum_{u \in \mathcal{U}^y_{\mathbf{s}_2}} \hat{y}_{u,i}}{|\mathcal{U}^y_{\mathbf{s}_2}|} \right|, \end{aligned} \tag{1}$$

where $\mathcal{U}_{\mathbf{s}} = \{u \mid (\mathbf{s}_u = \mathbf{s}, \mathbf{x}_u, \mathbf{v}_i, y_{u,i}) \in \mathcal{D}_{\text{test}}\}, \mathbf{s} \in \mathcal{S}$, and $\mathcal{U}^y_{\mathbf{s}} = \{u \mid (\mathbf{s}_u = \mathbf{s}, \mathbf{x}_u, \mathbf{v}_i, y_{u,i} = y) \in \mathcal{D}_{\text{test}}\}, y \in \{0, 1\}$, are the subgroups of users in the testing dataset $\mathcal{D}_{\text{test}}$.[1] The proposed metrics ΔDP and ΔEO can measure the difference of the model outputs across multiple user subgroups on any testing dataset (may distribute differently from the training data).

Next, we introduce the formal definition of counterfactual group fairness, which enforces that, for a recommendation model f trained on $\mathcal{D}_{\text{train}}$, a distribution over the group fairness metrics should be consistent with that in a "counterfactual world" where the distribution of sensitive attributes had been changed in a causal sense. We represent this counterfactual world by a counterfactual testing dataset $\mathcal{D}^{\mathbb{Q}}_{\text{test}} = \{(\tilde{\mathbf{s}}_u, \mathbf{x}_u, \mathbf{v}_i, y\}$, where the tuple $(\mathbf{x}_u, \mathbf{v}_i, y)$ is drawn from the same distribution $\mathbb{P}$ with the training dataset but $\tilde{\mathbf{s}}_u \in \mathbb{R}^{d_s}$ denote the out-of-distribution (OOD) sensitive attributes that are drawn from another counterfactual distribution $\mathbb{Q}$.

Definition 1 (Counterfactual Group Fairness, CGF). Let Δ_s be the set of all possible distributions of sensitive attributes. Given a group fairness metric called GF (e.g., ΔDP, ΔEO), a recommendation model f trained on $\mathcal{D}_{\text{train}} = \{(\mathbf{s}_u, \mathbf{x}_u, \mathbf{v}_i, y)\}$ satisfies counterfactual group fairness if for any two distributions of sensitive attributes $\mathbb{Q}_1, \mathbb{Q}_2 \in \Delta_s$ and for all possible values of GF,

$$\Pr\left\{\text{GF}\left(f, \mathcal{D}^{\mathbb{Q}_1}_{\text{test}}\right) \mid \mathcal{D}_{\text{train}}\right\} = \Pr\left\{\text{GF}\left(f, \mathcal{D}^{\mathbb{Q}_2}_{\text{test}}\right) \mid \mathcal{D}_{\text{train}}\right\}. \tag{2}$$

Since the distribution condition (2) in Definition 1 is a rather strict constraint, we focus instead on expectation and variance w.r.t. the counterfactual distribution $\mathbb{Q} \in \Delta_s$ of sensitive attributes, and will target a relaxation of CGF as well as its metric using the second-order central moment, defined below.

Definition 2 (ε-Counterfactually Group-Fair Recommendation). Given a group fairness metric called GF and a fairness threshold $\varepsilon > 0$, we call a recommendation model f is ε-counterfactually group-fair if the following metric of CGF satisfies

$$\begin{aligned}\text{CGF}(f) &:= \text{E}_{\mathbb{Q}}\left[\text{GF}\left(f, \mathcal{D}^{\mathbb{Q}}_{\text{test}}\right)\right]^2 \\ &= \text{Var}_{\mathbb{Q}}\left[\text{GF}\left(f, \mathcal{D}^{\mathbb{Q}}_{\text{test}}\right)\right] + \left\{\text{E}_{\mathbb{Q}}\left[\text{GF}\left(f, \mathcal{D}^{\mathbb{Q}}_{\text{test}}\right)\right]\right\}^2 \leq \varepsilon, \end{aligned} \tag{3}$$

where the expectation and variance are taken over the random selection of distribution $\mathbb{Q}$ in Δ_s. In particular, we denote the CGF metric in (3) by CGF-DP if ΔDP is used as the GF metric, and CGF-EO for the case that ΔEO is used as GF.

C. An Empirical Implementation of CGF

One problem with the CGF metric in (3) is that it is impossible to access to each counterfactual distribution $\mathbb{Q}$ of sensitive attributes in testing data. Therefore, (3) cannot be directly applied

[1] In (1), ΔDP and ΔEO are only computed using the nonempty sets $\mathcal{U}_{\mathbf{s}}$ and $\mathcal{U}^y_{\mathbf{s}}$ where $\mathbf{s} \in \mathcal{S}$, and $|\mathcal{U}|$ denotes the cardinality of the set $\mathcal{U}$.

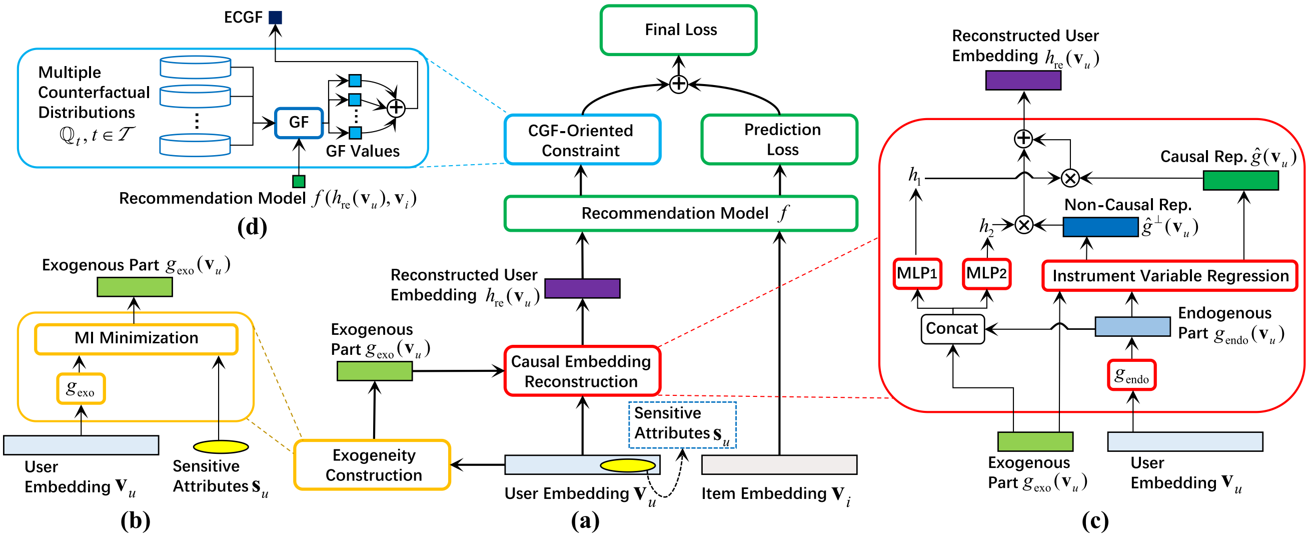
ECGF
Multiple Counterfactual Distributions
$\mathbb{Q}_t, t \in \mathcal{T}$
GF
GF Values
Recommendation Model $f(h_{re}(\mathbf{v}_u), \mathbf{v}_i)$
(d)
Final Loss
CGF-Oriented Constraint
Prediction Loss
Recommendation Model f
Reconstructed User Embedding $h_{re}(\mathbf{v}_u)$
Exogenous Part $g_{exo}(\mathbf{v}_u)$
Causal Embedding Reconstruction
Sensitive Attributes $\mathbf{s}_u$
Exogeneity Construction
User Embedding $\mathbf{v}_u$
Item Embedding $\mathbf{v}_i$
(a)
Exogenous Part $g_{exo}(\mathbf{v}_u)$
MI Minimization
g_{exo}
User Embedding $\mathbf{v}_u$
Sensitive Attributes $\mathbf{s}_u$
(b)
Reconstructed User Embedding $h_{re}(\mathbf{v}_u)$
h_1
h_2
Causal Rep. $\hat{g}(\mathbf{v}_u)$
Non-Causal Rep. $\hat{g}^{\perp}(\mathbf{v}_u)$
MLP1
MLP2
Instrument Variable Regression
Concat
Endogenous Part $g_{endo}(\mathbf{v}_u)$
g_{endo}
Exogenous Part $g_{exo}(\mathbf{v}_u)$
User Embedding $\mathbf{v}_u$
(c)

Fig. 4. (a): The overall architecture of the proposed MACE framework. (b): The module of exogeneity construction, where MI denotes mutual information. (c): The module of causal embedding reconstruction, where rep. stands for representation. (d): The module of CGF-oriented constraint, where GF denotes the metric of group fariness (e.g., ΔDP, ΔEO in (1)), the metric ECGF and the counterfactual distribution $\mathbb{Q}_t$ are defined in (5) and (4), respectively.

to the model training phase. One solution is mimicking the distribution shift of the sensitive attributes by sampling a series of counterfactual distributions that are formulated by adding uniform perturbations to the training data, yielding an empirical version of CGF metric named ECGF. Specifically, given a set of perturbation parameters $\mathcal{T} = \{t | t \in [0, 1]\}$, the *counterfactual distribution* of sensitive attributes is defined by

$$\mathbb{Q}_t := t\mathbb{P}_s + (1 - t)\mathbb{U}_s, \quad t \in \mathcal{T}, \tag{4}$$

where $\mathbb{P}_s$ denotes the distribution of sensitive attributes on training data that can be estimated by simply counting users with different sensitive attributes in training data, and $\mathbb{U}_s$ denotes a uniform distribution over sensitive attributes. In particular, if $t = 1$, then the counterfactual distribution $\mathbb{Q}_t$ degenerates to the original training distribution of sensitive attributes; if $t = 0$, then $\mathbb{Q}_t$ is equivalent to a uniform distribution over sensitive attributes. Finally, based on the proposed counterfactual distribution in (4), the following ECGF can be used to estimate the CGF metric in (3):

$$\mathrm{ECGF}(f, \mathcal{T}) := \sum_{t \in \mathcal{T}} \left[\mathrm{GF}\left(f, \mathcal{D}_{\text{train}}^{\mathbb{Q}_t}\right)\right]^2, \tag{5}$$

where, similarly to the definition of $\mathcal{D}_{\text{test}}^{\mathbb{Q}}$, $\mathcal{D}_{\text{train}}^{\mathbb{Q}}$ denotes a training dataset whose sensitive attributes change that are drawn from a counterfactual distribution $\mathbb{Q}$.

IV. MACE: The Proposed Framework

We propose a Model-Agnostic Causal Embedding learning framework named MACE which reconstructs the user embeddings for counterfactually group-fair recommendation.

A. Framework Overview

Fig. 4(a) illustrates the overall architecture of the proposed MACE framework. MACE consists of three ingredients: (1) the *exogeneity construction* tries to identify the exogenous part that is unconfounded by the sensitive attributes in user embeddings; (2) the *causal embedding reconstruction* treats the exogenous part as instrumental variables (IVs), decomposes a user embedding into the causal and non-causal representations using IV regression, and combines these representations for fair recommendation; (3) the *CGF-oriented constraint* is incorporated into the final loss for training the embeddings and recommendation model.

Algorithm 1 shows the detailed training procedure. At each training iteration, after B training instances are sampled into the buffer $\mathcal{V}$, a network g_{exo} is trained by performing mutual information minimization with an update cycle ρ, which is used for exogeneity construction. Then, the parameters of the recommendation model f and the networks $h_{\text{re}}, g_{\text{endo}}, g_{\text{exo}}$ for causal embedding reconstruction, are updated on $\mathcal{V}$ with a CGF-oriented constraint.

Next, we specify the ingredients of MACE with details.

B. Causal Embedding Learning

The main idea behind the causal embedding learning module is to construct instrumental variables that are unrelated to sensitive attributes (i.e., the exogenous part). These instrumental variables are then utilized to decompose and reconstruct the original user embeddings, distinguishing between the parts related and unrelated to sensitive attributes. This module aims to ensure fairness while maintaining recommendation performance, comprising two components: (1) *Exogeneity construction*, shown in Fig. 4(b), extracts the exogenous part of user embeddings and ensures that the exogenous part is uncorrelated with users' sensitive attributes by minimizing the mutual information; (2) *Causal embedding reconstruction*, shown in Fig. 4(c), involves regressing the endogenous part of user embeddings onto the exogenous part. This process disentangles the causal part, unconfounded

Algorithm 1: Model-Agnostic Causal Embedding Learning (MACE).

```
Require: Training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{s}_u, \mathbf{x}_u, \mathbf{v}_i, y_{u,i})\}$,
         batch size $B$, update cycle $\rho$, maximum
         number of iterations $N$,
Ensure: Neural networks $f, h_{\text{re}}, g_{\text{exo}}, g_{\text{endo}}$
Initialize $f, h_{\text{re}}, g_{\text{exo}}, g_{\text{endo}}$
for $n = 1, 2, \ldots, N$ do
  Sample $B$ instances from $\mathcal{D}_{\text{train}}$ and store them in $\mathcal{V}$
  if $n \mod \rho = 0$ then
    Update the parameters $\boldsymbol{\theta}_{\text{exo}}$ of $g_{\text{exo}}$ by mutual
    information minimization on $\mathcal{V}$   {(7)}
  end if
  Update the parameters $\boldsymbol{\theta}_{\text{all}}$ of $f, h_{\text{re}}, g_{\text{endo}}, g_{\text{exo}}$ by
  minimizing the final loss on $\mathcal{V}$   {(10)}
end for
return $\boldsymbol{\theta}_{\text{all}}$
```

by the sensitive attributes, and the non-causal part, confounded by the sensitive attributes. Subsequently, these two parts are combined to get the reconstructed user representations for the fair recommendation.

1) Exogeneity Construction: To identify the exogeneity and endogeneity in the original user embedding shown in Fig. 3, we introduce two deep neural networks g_{exo} and g_{endo} to extract the exogenous and endogenous parts, respectively. Specifically, given a user embedding $\mathbf{v}_u$ with sensitive attributes $\mathbf{s}_u$, the exogenous part can be represented by $g_{\text{exo}}(\mathbf{v}_u) \in \mathbb{R}^{d_{\text{exo}}}$ whose dimension is d_{exo}. As shown in Fig. 4(b), the model parameters $\boldsymbol{\theta}_{\text{exo}}$ of the network g_{exo} are obtained using mutual information minimization:

$$\boldsymbol{\theta}_{\text{exo}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\text{exo}}}{\arg\min}\, \mathrm{I}\left(g_{\text{exo}}(\mathbf{v}_u|\boldsymbol{\theta}); \mathbf{s}_u\right), \tag{6}$$

where $\mathrm{I}(\mathbf{a};\mathbf{b})$ denotes the mutual information (MI) between vectors $\mathbf{a}$ and $\mathbf{b}$, $\boldsymbol{\Theta}_{\text{exo}}$ denotes the trainable parameter space of g_{exo}, and $g_{\text{exo}}(\cdot|\boldsymbol{\theta})$ denotes the network parameterized by $\boldsymbol{\theta}$. Since directly minimizing MI is intractable, we employ an upper bound estimation of MI instead, called CLUB [48], which is formulated using the probability log-ratio of conditional distributions between the user embedding and the sensitive attributes. Then, at each training iteration, after B user embeddings are sampled into the buffer $\mathcal{V}$, $\boldsymbol{\theta}_{\text{exo}}$ in (6) can be estimated by minimizing the objective function:

$$\sum_{(\mathbf{s}_u,\mathbf{v}_u)\in\mathcal{V}} \left[\log q\left(\mathbf{s}_u|g_{\text{exo}}(\mathbf{v}_u|\boldsymbol{\theta})\right) - \log q\left(\tilde{\mathbf{s}}|g_{\text{exo}}(\mathbf{v}_u|\boldsymbol{\theta})\right)\right], \tag{7}$$

where $\tilde{\mathbf{s}}$ is a vector of sensitive attributes uniformly sampled from the buffer $\mathcal{V}$, and q is a stochastic encoder implemented in a Gaussian variational family as in [48].

The use of MI minimization aims to reduce the correlation between the exogenous part $g_{\text{exo}}(\mathbf{v}_u)$ and the sensitive attributes $\mathbf{s}_u$, enforcing this exogenous part to meet the unconfounded condition of instrumental variables. Besides, we extract the endogenous part from the original user embedding $\mathbf{v}_u$ using another neural network g_{endo}, denoted by $g_{\text{endo}}(\mathbf{v}_u) \in \mathbb{R}^{d_{\text{endo}}}$, where d_{endo} is the dimension of $g_{\text{endo}}(\mathbf{v}_u)$. Obviously, $g_{\text{exo}}(\mathbf{v}_u)$ has correlation with $g_{\text{endo}}(\mathbf{v}_u)$ since they are both extracted from the same user embedding $\mathbf{v}_u$. Next, we will show how to reconstruct user embedding through causal embedding learning on the exogenous and endogenous parts, and how to update g_{endo} and g_{exo} under fairness constraint.

2) Causal Embedding Reconstruction: The causal embedding reconstruction module first regresses the endogenous part onto the exogenous part, disentangling the causal part unconfounded by sensitive attributes and the non-causal part confounded by sensitive attributes. Directly using the causal part, which is unconfounded by sensitive attributes, can ensure the fairness of recommendations. However, to maintain recommendation accuracy, we also retain the non-causal part. We employ two Multi-Layer Perceptrons (MLPs) to adaptively learn the weights for these two parts, then combine them through weighted summation. This approach enables the model to enhance fairness while maintaining accuracy for recommendations.

Specificall, after decomposing the original user embedding $\mathbf{v}_u$ into the exogenous part $g_{\text{exo}}(\mathbf{v}_u)$ and the endogenous part $g_{\text{endo}}(\mathbf{v}_u)$, using these two parts, we can reconstruct the user embedding through causal learning, shown in Fig. 4(c).

Formally, we denote the reconstructed user embedding of $\mathbf{v}_u$ by $h_{\text{re}}(\mathbf{v}_u)$ and represent it using a combination of two orthogonal vectors $\hat{g}(\mathbf{v}_u)$ and $\hat{g}^{\perp}(\mathbf{v}_u)$:

$$h_{\text{re}}(\mathbf{v}_u) = h_1\hat{g}(\mathbf{v}_u) + h_2\hat{g}^{\perp}(\mathbf{v}_u), \tag{8}$$

where: (a) the vector $\hat{g}(\mathbf{v}_u)$ is the projection of the endogenous part $g_{\text{endo}}(\mathbf{v}_u)$ onto the exogenous part $g_{\text{exo}}(\mathbf{v}_u)$, which can be seen as a *causal representation* since it does not depend on the confounders $\mathbf{s}_u$ in Fig. 3; (b) another vector $\hat{g}^{\perp}(\mathbf{v}_u)$ is the residual of projecting $g_{\text{endo}}(\mathbf{v}_u)$, which is a *non-causal representation* that is orthogonal to $\hat{g}(\mathbf{v}_u)$; (c) $h_1, h_2 \in \mathbb{R}$ are the combination coefficients in the mapping h_{re}, which are obtained using two different MLPs (multilayer perceptrons)

$$h_j = \mathrm{MLP}_j(g_{\text{exo}}(\mathbf{v}_u), g_{\text{endo}}(\mathbf{v}_u)), \quad j = 1 \text{ or } 2,$$

where their inputs are both the concatenations of the exogenous and endogenous parts of the user embedding $\mathbf{v}_u$.

Next, we provide a rigorous derivation of the orthogonal vectors $\hat{g}(\mathbf{v}_u)$ and $\hat{g}^{\perp}(\mathbf{v}_u)$ in (8). The causal representation $\hat{g}(\mathbf{v}_u)$ aims to remove the path from the confounders (sensitive attributes) to the treatment (user embedding) on the left side of Fig. 3, for achieving fair recommendation through embedding learning. Inspired by the instrumental variable (IV) regression [44], [49], we formulate the causal representation $\hat{g}(\mathbf{v}_u)$ as a linear projection of the exogenous and unconfounded $g_{\text{exo}}(\mathbf{v}_u)$ as $\hat{g}(\mathbf{v}_u) = \widehat{\boldsymbol{\Gamma}}g_{\text{exo}}(\mathbf{v}_u)$, where $\widehat{\boldsymbol{\Gamma}}$ is a parameter matrix that is estimated by regressing the endogenous part on the exogenous part. More specifically, at each training iteration, B user embeddings are sampled from $\mathcal{D}_{\text{train}}$ and stored in a buffer $\mathcal{V}_{\text{ue}} = \{\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \ldots, \bar{\mathbf{v}}_B\}$, and then $\widehat{\boldsymbol{\Gamma}}$ is obtained by a ridge regression

$$\widehat{\boldsymbol{\Gamma}} = \underset{\boldsymbol{\Gamma}\in\mathbb{R}^{d_{\text{endo}}\times d_{\text{exo}}}}{\arg\min} \sum_{\mathbf{v}\in\mathcal{V}_{\text{ue}}} \|g_{\text{endo}}(\mathbf{v}) - \boldsymbol{\Gamma}g_{\text{exo}}(\mathbf{v})\|_2^2 + \tau\|\boldsymbol{\Gamma}\|_{\text{F}}^2$$

$$= \mathbf{G}_{\text{endo}}^{(B)} \mathbf{G}_{\text{exo}}^{(B)\intercal} \left(\mathbf{G}_{\text{exo}}^{(B)} \mathbf{G}_{\text{exo}}^{(B)\intercal} + \tau \mathbf{I}_{d_{\text{exo}}} \right)^{-1}, \tag{9}$$

where $\tau > 0$ is a regularization parameter, and $\mathbf{G}_{\text{endo}}^{(B)}$, $\mathbf{G}_{\text{exo}}^{(B)}$ are constructed using $\mathcal{V}_{\text{ue}}$: $\mathbf{G}_{\text{endo}}^{(B)} = [g_{\text{endo}}(\bar{\mathbf{v}}_1), \ldots, g_{\text{endo}}(\bar{\mathbf{v}}_B)] \in \mathbb{R}^{d_{\text{endo}} \times B}$,$\mathbf{G}_{\text{exo}}^{(B)} = [g_{\text{exo}}(\bar{\mathbf{v}}_1), \ldots, g_{\text{exo}}(\bar{\mathbf{v}}_B)] \in \mathbb{R}^{d_{\text{exo}} \times B}$. Please note that classic IV regression approaches typically regress the endogenous variables on the instrumental variables using the whole dataset. Here we regard the exogenous part as the instrumental variables and calculate the solution of IV regression problem using mini-batches of training data due to computational efficiency requirements. Besides, to avoid overfitting the batch size should satisfy $B > d_{\text{exo}}$.

After obtaining the causal representation using the IV regression, we can easily express the non-causal representation of user embedding $\mathbf{v} \in \mathcal{V}_{\text{ue}}$ as the regression residual in (9) $\hat{g}^{\perp}(\mathbf{v}) = g_{\text{endo}}(\mathbf{v}) - \widehat{\mathbf{\Gamma}} g_{\text{exo}}(\mathbf{v})$. The intuition is that the non-causal associations can also capture the user preference, which can contribute to the recommendation accuracy and user satisfaction [50]. The observation motivates us that the non-causal representation $\hat{g}^{\perp}(\mathbf{v})$ can be leveraged to improve the recommendation performance although the non-causal part is often discarded for model parameter estimation in traditional causal inference. Furthermore, the combination $h_{\text{re}}(\mathbf{v}_u)$ in (8) can be seen as a learning-based mixup of the causal and non-causal representations of the user embedding $\mathbf{v}_u$, which balances the fairness and recommendation accuracy.

C. Fairness-Oriented Model Learning

1) CGF-Oriented Constraint: To predict the user's preference score, the reconstructed user embedding $h_{\text{re}}(\mathbf{v}_u)$ is fed into a recommendation model f, and the predicted user's preference score for item $i \in \mathcal{I}$ can be represented by a composite function $\hat{y}_{u,i} = f(h_{\text{re}}(\mathbf{v}_u), \mathbf{v}_i)$. Obviously, the model parameters of f and h_{re} need to be trained alternately. To further improve the counterfactual group fairness (CGF) of f, it is necessary to incorporate a CGF-oriented constraint into the model learning process.

As shown in Fig. 4(d), we use ECGF in (5) as an empirical implementation of CGF-oriented constraint, which needs to be computed using the outputs of the function $f(h_{\text{re}}(\cdot), \cdot)$. Given a set of perturbation parameters $\mathcal{T} = \{t | t \in [0, 1]\}$, we specify the notion of ECGF by ECGF$(f, h_{\text{re}}, \mathcal{T})$, which can be seen as a distributionally robust fairness constraint. Next, we will incorporate ECGF into the optimization problem as a fairness constraint, in helping to make counterfactually group-fair recommendations.

2) Model Optimization: We formulate an optimization problem with the CGF-oriented constraint, for trading-off the recommendation performance on observational data and the group fairness on potentially shifted distributions of sensitive attributes. Specifically, letting $\boldsymbol{\theta}_{\text{all}}$ be all the trainable parameters of $f, h_{\text{re}}, g_{\text{endo}}, g_{\text{exo}}$, as shown in Algorithm 1, we alternatively optimize the parameters $\boldsymbol{\theta}_{\text{exo}}$ of the network g_{exo} for exogeneity construction, and $\boldsymbol{\theta}_{\text{all}}$ for minimizing the final loss (i.e., prediction loss with a CGF-oriented constraint). More formally, after the exogeneity construction, model optimization amounts to minimizing the following *final loss* on training data $\mathcal{D}_{\text{train}}$:

$$\mathcal{L}(\boldsymbol{\theta}_{\text{all}}) := \sum_{(\mathbf{s}_u, \mathbf{x}_u, \mathbf{v}_i, y_{u,i}) \in \mathcal{D}_{\text{train}}} \underbrace{\ell\left[f(h_{\text{re}}(\mathbf{v}_u), \mathbf{v}_i), y_{u,i}\right]}_{\text{Prediction Loss}} + \underbrace{\lambda \cdot \text{ECGF}(f, h_{\text{re}}, \mathcal{T})}_{\text{CGF-Oriented Constraint}} + \gamma \|\boldsymbol{\theta}_{\text{all}}\|_2^2, \tag{10}$$

where ℓ is a loss function for the recommendation task (cross entropy loss is adopted in this paper) and $\lambda > 0$ denotes the fair weight, $\gamma > 0$ is the regularization parameter. In Algorithm 1, MACE updates the parameters $\boldsymbol{\theta}_{\text{all}}$ in an end-to-end way using the mini-batch gradient descent. The optimization process in Algorithm 1 can be viewed as a multi-objective optimization problem, including two optimization objectives: (a) minimizing mutual information (MI) for exogenous part construction, and (b) minimizing the final loss in (10) for the recommendation task under the CGF-oriented constraint. We employ a sequential optimization method (also known as the lexicographic method) in multi-objective optimization [51], sequentially optimizing the objectives (a) and (b). Since objective (a) is solely related to the parameters of the exogenous part g_{exo}, during MI minimization, only the parameters of the exogenous part are optimized. On the other hand, objective (b) is related to all parameters, necessitating a full parameter updating. Thus, although it may seem that the parameters of the exogenous part are optimized twice in certain steps, they serve different optimization objectives. Besides, MACE is a model-agnostic framework, where the recommendation model f can be implemented over existing recommendation models (specified in Section VI-A2).

V. Discussion

Relation to Existing Fairness Metrics: To measure the group fairness in the presence of OOD sensitive attributes, we introduce a novel concept named counterfactual group fairness (CGF) as well as its empirical metric ECGF defined in (5). Our aim in part is to work toward a more unified view of existing concepts of group fairness and counterfactual fairness:

(1) *Relation to group fairness:* Existing metrics of group fairness (e.g., DP distance, EO distance [23], [24]) aim to measure the difference of the model outputs between two groups (e.g., two user subgroups) on a *fixed* distribution of sensitive attributes. The proposed ECGF can be regarded as an extension of the existing metrics of group fairness in counterfactual worlds, which can be used to give counterfactual estimations of the group fairness in multiple *shifted* distributions of sensitive attributes. As an empirical implementation, we estimate ECGF on the counterfactual distributions $\mathbb{Q}_t, t \in \mathcal{T}$ defined in (4). In particular, letting the set of perturbation parameters be $\mathcal{T} = \{1\}$, the counterfactual distribution $\mathbb{Q}_t$ turns to be the original training distribution $\mathbb{P}_s$, and then the proposed ECGF degenerates into the classic group fairness.

(2) *Relation to counterfactual fairness:* Counterfactual fairness (CF) is typically defined at the individual level [14], [15], which requires that the predicted results for an individual (e.g., a user) should be the same when the sensitive attributes of the user change. Efforts have been made to generalize the metric

of individual CF to a more general formula, by introducing the average causal effects of the sensitive attributes on the outcomes [17]. But in these metrics of CF, the assignment of sensitive features is *deterministic* for each individual, ignoring the potential distribution shift of the sensitive attributes of the whole population. The proposed CGF and the corresponding empirical metric ECGF assume that the assignments of sensitive features are *random*, and provide a way to measures the effect of the changes in the population distribution of sensitive attributes on group fairness. In particular, the proposed CGF in Definition 1 will degenerate into the classic individual CF, if GF is replaced with the model outputs and the distributions $\mathbb{Q}_1$ and $\mathbb{Q}_2$ on sensitive features are set to the deterministic indicator vectors (e.g., $(0, 1, 0, \ldots, 0)$ that indicates the second attribute is assigned).

Difference With IV Methods: MACE has made the following fundamental modifications for adapting IV regression to counterfactually group-fair recommendation:

1) *Representation of embeddings*. In classic IV methods [52], [53], instrumental variables (IVs) are predefined by domain experts. However, effective IVs are usually hard to find. To avoid the challenge of choice of IVs, in the exogeneity construction module of MACE (Section IV-B1), exogenous parts in user embeddings are extracted and represented through deep learning, which are treated as the proxy of IVs. Besides, the treatments (i.e., the endogenous parts in user embeddings) are also represented using neural networks and updated in an end-to-end manner. In this way, useful information is extracted from the user embeddings for fair prediction of outcomes.

2) *Reconstruction of embeddings:* After estimating the model parameters using the IV regression, traditional IV methods use the original treatments as inputs for prediction. However, to achieve a favorable trade-off between the fairness and accuracy of recommendation, MACE learns the combination coefficients of the causal and non-causal user representations ((8) in Section IV-B2). The new reconstructed user embeddings were used as the input for recommendation. The module of causal embedding reconstruction in MACE can be seen as a flexible extension of IV methods to the case where the non-causal representation of the original treatment is also helpful in enhancing the recommendation accuracy. In particular, setting the combination coefficients $h_1 = h_2 = 1$ in (8) for prediction, the prediction process in the proposed causal embedding reconstruction degenerates into that in classic IV methods.

VI. Experiments

We conducted experiments to test the performance of MACE. The source code, dataset description and implementation details have been shared at https://anonymous.4open.science/r/MACE-9568/.

A. Experimental Settings

1) Dataset Description: The experiments were conducted on 3 publicly available recommendation benchmarks: (a) `MovieLens-1 M`contains 1,000,209 user-system interactions from 6,040 users on 3,952 movies;[2] (b) `Insurance` is an insurance recommendation dataset on Kaggle,[3] containing 5,382 interactions from 1,231 users on 21 insurances; (c) `Rent-TheRunWay` [54] contains 192,544 user-system interactions from 105,508 customers on 5,850 products.

2) Baselines: The proposed MACE is model-agnostic, which can help enhance the counterfactual group fairness for the following *base models* for recommendation:

DMF [2] is a matrix factorization model for recommendation, which uses deep learning to capture the low-dimensional representations of users and items from the user-item interaction matrix. **DIN** [3] learns the representations of user interests by considering the sequential historical behaviors given a candidate item. **DeepModel** [55], [56] is a base model for recommendation which concatenates the embeddings of users and items together and feeds them into a DNN to learn the nonlinear relations among features.

MACE was compared to baselines that focus on the group fairness and optimize the fairness constraint with different strategies, including:

Mixup [24] optimizes the loss function on paths of interpolated examples between different groups to improve the generalization for both accuracy and fairness, where the interpolated sample is convex combinations of examples. **GapReg** [24] incorporates the group fairness constraint into loss functions, and directly optimizes the obtained loss. **FairMI** [37] aims to enhance group fairness in recommender systems by minimizing the mutual information between embeddings and sensitive information while simultaneously maximizing the mutual information between embeddings and non-sensitive information.

MACE was also compared to another category of baseline that focuses on the individual counterfactual fairness, called **AdvLearning** [15]. It generates user embeddings that are independent of sensitive attributes via adversary learning for fair recommendation.

The proposed MACE is a model-agnostic framework, which can be applied to different base models such as DMF, DIN and DeepModel. In the experiments, the recommendation model f in MACE was set as the base models (DMF, DIN or DeepModel), and the prediction modules in other baselines were set as the same base models. Besides, the constraint ECGF in MACE can integrate any optimization strategies such as Mixup and GapReg, yielding two versions of MACE: **MACE-Mixup** and **MACE-GapReg**. Mixup [24] utilizes the Mixup technique for interpolation. It involves forward propagation of the output obtained by interpolating two samples with different sensitive features, followed by backward propagation to compute gradients. GapReg [24] directly incorporates the counterfactual group fairness constraints computed from Eq.(5) into the final loss function.

3) Evaluation Metrics: Area Under Curve (AUC) was adopted to measure the recommendation accuracy. To measure the counterfactual group fairness, we used the metric of CGF

[2] [Online]. Available: https://grouplens.org/datasets/movielens/1m/

[3] [Online]. Available: https://www.kaggle.com/mrmorj/insurance-recommendation

defined in (3), including CGF-DP (selecting GF as ΔDP) and CGF-EO (selecting GF as ΔEO). Specifically, to assess AUC and CGF over different testing distributions of sensitive attributes, we employed four distinct testing distributions. These distributions are as follows. (1) *The original distribution:* reflects the natural distribution of sensitive attributes in the original testing data, where we did not alter the proportions of sensitive attributes while testing. (2) *A uniform distribution:* we adjusted the testing data such that the proportions of instances for each sensitive attribute were set to be nearly identical. For example, if the sensitive attribute was gender and the original dataset had an imbalanced gender ratio, we resampled the testing data and modified the gender of some users to ensure an equal number of male and female instances. (3) *Two extreme distributions:* For the `MovieLens-1 M` dataset, we depicted the extreme distributions of occupations in histograms of Fig. 2. For the `RentTheRunWay` dataset, we focused on the age attribute and constructed extreme distributions by allocating only 10% of the users to two specific age groups, significantly deviating from the original age distribution. For the `Insurance` dataset, we created two distinct extreme distributions based on gender. In one distribution, 90% of the users were female and the remaining 10% were male. In the other distribution, 10% of the users were female while the remaining 90% were male. This drastic alteration allowed us to observe the model's performance under highly imbalanced gender distributions.

The above 4 sensitive attribute distributions were utilized in computing the CGF metrics on testing data defined in (3). Specifically, we used the mean of CGF over these four sensitive attribute distributions as an estimate of the expectation in (3), to evaluate the counterfactual group fairness metrics of different algorithms. According to the definition of CGF, the lower CGF score is better, and an ideal result to meet the counterfactual group fairness requirement is a CGF score closed 0.

4) Implementation Details: All the baselines and base models were trained on a single NVIDIA Tesla P100 GPU, with the batch size tuned among $\{64, 128, 256, 512, 1024\}$ and the learning rate tuned in $\{1\text{E-}1, 1\text{E-}2, 1\text{E-}3, 1\text{E-}4\}$. In the implementation of MACE, we set the networks g_{exo}, g_{endo}, MLP$_1$ and MLP$_2$ to 3-layer fully connected neural networks, respectively, where the activation functions were tanh and sigmoid. The dimensions of the exogenous part $g_{\text{exo}}(\mathbf{v}_u)$ and the endogenous part $g_{\text{endo}}(\mathbf{v}_u)$ were set to $d_{\text{exo}} = d_{\text{exo}} = 32$, and the dimensions of the user and item embeddings were all set to 128. The embeddings are randomly initialized and are continuously updated during training. In the training process of MACE, the batch size B and the maximum number of iteration N were set to 128 and 20 × sample size/128, respectively, the update cycle ρ was set according to the sample size of training data ensuring that the MI minimization was executed 20 times, the fair weight λ in the final loss 10 was tuned among $[0 : +0.1 : 5]$, the regularization parameter γ in the final loss was set to 0.001, and the regularization parameter τ in the IV regression 9 was set to 0.9. For the baselines Mixup and MACE-GapReg, we chose Δ DP as the fairness constraint. For fair comparisons, in the CGF-oriented constraint of MACE, we chose a weighted version of Δ DP as the GF metric in ECGF, where the set of perturbation parameters in ECGF was set to $\mathcal{T} = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

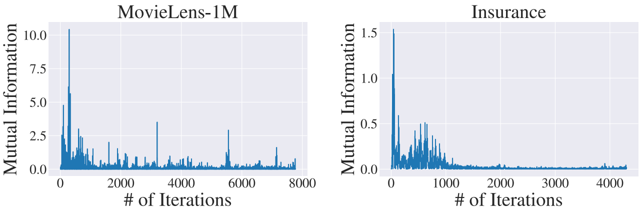

Fig. 5. Mutual information between the sensitive attributes and exogenous part in MACE-Mixup w.r.t. number of iterations, where MACE-Mixup was equipped with DMF.

B. Experimental Results

Tables I, II and III report the performances of the proposed two versions of MACE and the baselines on `MovieLens-1 M`, `Insurance` and `RentTheRunWay`, in terms of the AUC and the fairness metrics CGF-DP and CGF-EO. From the results, we can conclude that MACE consistently achieved the lowest CGF-DP and CGF-EO, verifying the effectiveness of MACE in terms of enhancing the counterfactual group fairness in recommendation. More specifically, MACE can help significantly improve the group fairness of the base models in counterfactual worlds that out-of-distribution (OOD) sensitive attributes exist. Besides, the two versions of MACE combined with the constraint optimization strategies Mixup and GapReg obtained lower CGF scores than the baselines that apply these strategies to the traditional group fairness constraint alone.

On the other hand, the results of AUC in Tables I, II and III revealed an interesting phenomenon: MACE even achieved a higher accuracy for recommendation compared to the base models in most cases, indicating that MACE enhanced not only the OOD generalization of group fairness constraints but also the overall accuracy. Similar experimental results were also obtained by the baseline Mixup, which were consistent with its empirical studies in the fairness literature [24]. Looking into the phenomenon, the CGF-oriented constraint in MACE can not only be regarded as a fairness constraint, but also a stability constraint of the training process for achieving a more accurate learner. Besides, the decomposition and reconstruction of the embeddings in MACE can also help achieve the improvements of both the fairness and the accuracy.

From the results in Tables I, II, and III, it is evident that MACE-Mixup often exhibits higher fairness metrics while having slightly lower recommendation accuracy. This suggests that incorporating Mixup, an effective optimization technique, during MACE optimization can further enhance CGF without significant sacrifice in recommendation accuracy. In practical applications, MACE-Mixup offers a better solution for scenarios with higher fairness requirements.

C. Analysis

1) Effectiveness of Exogeneity Construction: Fig. 5 illustrates the mutual information between the sensitive attributes

TABLE I

PERFORMANCE COMPARISON OF MACE AND BASELINES ON `MovieLens-1 M`, WHERE BASE MODEL REPRESENTS DMF (COLUMNS 2–4) OR DIN (COLUMNS 5–7), AND USER'S OCCUPATION WAS SELECTED AS THE SENSITIVE ATTRIBUTE

Algorithm	DMF			DIN		
	AUC▲	CGF-DP▼	CGF-EO▼	AUC▲	CGF-DP▼	CGF-EO▼
Base Model	0.7260	0.2490E-2	0.1534E-1	0.7445	1.1988E-2	0.5888E-1
FairMI	0.7270	0.2068E-2	0.2287E-1	0.7380	0.7480E-2	0.4997E-1
AdvLearning	0.6586	0.2161E-2	0.2364E-1	0.7308	0.5385E-2	0.3438E-1
Mixup	0.7274	0.1669E-2	0.1677E-1	0.7374	1.1875E-2	0.6207E-1
GapReg	0.7236	0.2399E-2	0.1634E-1	0.7346	1.2022E-2	0.4281E-1
MACE-Mixup	**0.7275**	0.1581E-2	0.1381E-1	0.7388	**0.4869E-2***	**0.2914E-1***
MACE-GapReg	0.7259	**0.1077E-2***	**0.0889E-1***	**0.7396**	0.7152E-2	0.4233E-1

'▲': larger is better; '▼': smaller is better; bold: best results except for the base models; '*': improvements over baselines are statistical significant (t-test, p-value< 0.05).

TABLE II

PERFORMANCE COMPARISON OF MACE AND BASELINES ON `Insurance`, WHERE THE BASE MODEL REPRESENTS DMF (COLUMNS 2–4) OR DEEPMODEL (COLUMNS 5–7), AND USER'S GENDER WAS SELECTED AS THE SENSITIVE ATTRIBUTE

Algorithm	DMF			DeepModel		
	AUC▲	CGF-DP▼	CGF-EO▼	AUC▲	CGF-DP▼	CGF-EO▼
Base Model	0.9075	2.7759E-6	6.9387E-5	0.8703	11.9092E-6	67.9759E-5
FairMI	0.8138	0.1609E-6	0.2707E-5	0.9071	4.7449E-6	7.9277E-5
AdvLearning	0.8336	0.0670E-6	0.3247E-5	0.8816	1.4116E-6	5.9599E-5
Mixup	0.9096	0.9188E-6	0.8719E-5	0.9011	1.3171E-6	5.6571E-5
GapReg	0.9010	1.3603E-6	6.4156E-5	0.8685	1.8850E-6	6.6218E-5
MACE-Mixup	0.9117	**0.0456E-6***	**0.2063E-5***	0.9020	**1.2442E-6***	5.4251E-5
MACE-GapReg	**0.9180***	1.2443E-6	5.3975E-5	**0.9230***	1.1485E-6	**4.5155E-5***

Since the dataset Insurance did not record the timestamps of user behaviors that are essential for DIN, we use DeepModel as a baseline instead of DIN on Insurance.

TABLE III

PERFORMANCE COMPARISON OF MACE AND BASELINES ON `RentTheRunWay`, WHERE THE BASE MODEL REPRESENTS DMF (COLUMNS 2–4) OR DEEPMODEL (COLUMNS 5–7), AND USER'S AGE WAS SELECTED AS THE SENSITIVE ATTRIBUTE

Algorithm	DMF			DeepModel		
	AUC▲	CGF-DP▼	CGF-EO▼	AUC▲	CGF-DP▼	CGF-EO▼
Base Model	0.6965	7.2216E-3	4.8071E-2	0.6078	13.1790E-4	11.5365E-3
FairMI	0.7234	6.0873E-3	2.0625E-2	0.6070	2.2471E-4	8.8665E-3
AdvLearning	0.7223	5.5384E-3	1.9763E-2	0.5930	8.1088E-4	8.3435E-3
Mixup	0.7065	4.4912E-3	3.4078E-2	0.6075	9.1795E-4	8.8028E-3
GapReg	0.7003	5.5505E-3	3.9213E-2	0.6080	9.8532E-4	8.7682E-3
MACE-Mixup	**0.7864***	4.9261E-3	**1.9484E-2***	0.6129	**0.5512E-4***	**5.3810E-3***
MACE-GapReg	0.7747	**4.3917E-3***	2.5537E-2	**0.6144***	0.8726E-4	7.5784E-3

Similarly to the reasons for dataset `Insurance`, we also use DeepModel as a baseline instead of DIN on RentTheRunWay.

and the constructed exogenous part $g_{\text{exo}}(\mathbf{v}_u)$ of user embedding $\mathbf{v}_u$ in MACE w.r.t. number of iterations. On both datasets, the mutual information decreased steadily with the training went on. The results indicate that the network g_{exo} in MACE effectively extracted the exogenous part from user embeddings and it is reasonable to treat the exogenous part as instrumental variables (IVs). To verify the importance of the exogenous part, we compared the original MACE with its variant that used a completely random network g_{exo}, denoted by "MACE w/ random IV". The results in Table IV indicate that exogenous parts contributed to the recommendation performances, especially in terms of the CGF.

2) Impact of Causal Embedding Reconstruction: The module of causal embedding reconstruction in MACE made

TABLE IV
ABLATION STUDY OF MACE, WHERE THE BASE MODEL WAS SELECTED AS DIN ON `MovieLens-1 M` AND DMF ON `Insurance`, AND USER'S OCCUPATION AND GENDER WERE SELECTED AS THE SENSITIVE ATTRIBUTES ON `MovieLens-1 M` AND `Insurance`, RESPECTIVELY

Section	Algorithm	DIN on `MovieLens-1M`			DMF on `Insurance`		
		AUC▲	CGF-DP▼	CGF-EO▼	AUC▲	CGF-DP▼	CGF-EO▼
Section 6.3.1	MACE w/ random IV	0.7382	0.5594E-2	0.3768E-1	0.8858	1.1577E-6	1.1867E-5
Section 6.3.2	MACE w/ classic IV	0.7375	0.5427E-2	0.3027E-1	0.7020	0.7447E-6	0.2854E-5
	MACE w/o residual	0.7371	**0.4688E-2**	**0.2750E-1**	0.8288	**0.0003E-6**	**0.0003E-5**
	MACE w/o IV regression	0.7373	0.5257E-2	0.3308E-1	0.8827	0.1621E-6	0.2543E-5
Section 6.3.3	MACE w/o ECGF	**0.7407**	0.7579E-2	0.4875E-1	0.9057	2.7093E-6	1.7153E-5
	MACE w/ GF	0.7369	0.5831E-2	0.3515E-1	0.9044	0.5066E-6	0.2787E-5
Section 6.3.4	MACE ($\mathcal{T} = \{1\}$)	0.7383	0.5520E-2	0.3534E-1	0.9076	0.2594E-6	0.4932E-5
	Original MACE	0.7388	0.4869E-2	0.2914E-1	**0.9117**	0.0456E-6	0.2063E-5

Original MACE adopted the base model for recommendation and the Mixup for constraint optimization, and its results were identical to that of in Table 1 and Table 2.

several fundamental modifications compared with classic IV methods. To further understand the role of these modifications, we designed the following variants of MACE for comparison: (a) "MACE w/ classic IV" set the combination coefficients in (8) as $h_1 = 1, h_2 = 0$ for training and $h_1 = h_2 = 1$ for testing, which conformed to the setting of classic IV methods; (b) "MACE w/o residual" directly discarded the IV regression residuals that may be helpful for recommendation accuracy, which focused on the effectiveness of the proposed fairness constraint; (c) "MACE w/o IV regression" only used the exogenous part obtained by g_{exo} as the reconstructed user embedding (i.e., without using the IV regression). From the results in Table IV we have the following three observations: (1) The significant decrease in recommendation accuracy when employing "MACE w/ classic IV" suggests that traditional instrumental variable methods prioritize estimating causal effects rather than enhancing prediction accuracy. Therefore, they cannot be directly applied to recommendation models. (2) The advantage of "MACE w/o residual" in fairness metrics indicates that the causal part in the causal embedding reconstruction can effectively improve recommendation fairness, aligning with the motivation outlined in Section IV-B2. (3) The decline in performance with "MACE w/o IV regression" validates the effectiveness of introducing instrumental variables in counterfactually group-fair recommendation. In summary, the proposed fairness constraint was effective for achieving CGF, and the modifications of IV methods in MACE were helpful for improving counterfactual group fairness while retaining recommendation accuracy.

3) Impact of CGF-Oriented Constraint: We conducted an ablation study to show the performances of the proposed CGF-oriented constraint called ECGF. We compared MACE with its two variants: "MACE w/o ECGF" that removed the ECGF term from the final loss (10), and "MACE w/ GF" that used the traditional group fairness constraint (i.e., a fairness constraint defined on a fixed distribution). The results reported in Table IV show that, although good accuracy can be achieved by these two variants, lower CGF scores were obtained, verifying the importance of the ECGF term for fair recommendation on counterfactual distributions. Besides, Fig. 6 illustrated the impact of the fair wight λ in (10), from which we can observe that a larger λ leads to a lower CGF score with a sacrifice on AUC, and a suitable choice of λ will significantly decrease the CGF score with a comparable accuracy.

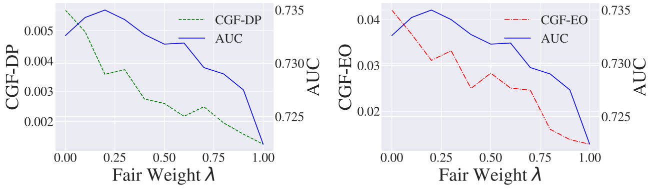

Fig. 6. Impact of the fair weight λ in MACE, measured by AUC, CGF-DP (Left), and CGF-EO (Right) on `MovieLens-1 M`. Results are obtained by MACE-Mixup equipped with DMF.

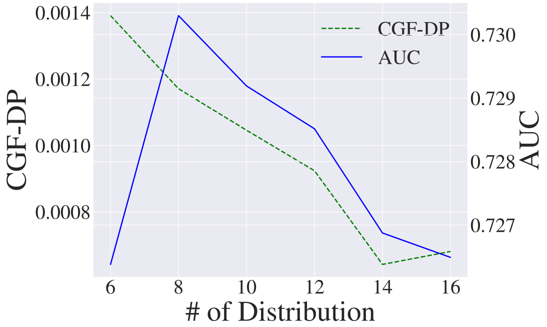

Fig. 7. Impact of the number of counterfactual distributions in (4) (i.e., the size of the set of perturbation parameters $\mathcal{T}$, denoted by "# of Distributions"), measured by AUC, CGF-DP on `MovieLens-1 M`. Results were obtained by MACE-Mixup equipped with DMF.

4) Impact of the Number of Counterfactual Distributions: To verify the effectiveness of the empirical version of CGF metric, we conducted experiments by varying the number of counterfactual distributions in Eq. (4). From the results in Fig. 7 we can observe that larger number of counterfactual distributions $\{\mathbb{Q}_t\}_{t\in\mathcal{T}}$ can significantly decrease the CGF score while slightly hurting the recommendation accuracy, indicating that the counterfactual distributions can well mimic the distribution drift of sensitive attributes. We also tested the performance of MACE when $\mathcal{T} = \{1\}$, as shown in Table IV. In this case, the training process only considers group fairness, like classic group fairness methods. It can be observed that the counterfactual group fairness and accuracy achieved in a single distribution were not as good as those obtained by training MACE on multiple distributions. This highlights the significance of incorporating multiple distributions of sensitive attributes during training to enhance the robustness of fairness. The conclusion validates the effectiveness of the proposed MACE in ensuring the counterfactual group fairness of recommendation models.

VII. Conclusion

This paper aims to measure and enhance the group fairness of recommendation in counterfactual worlds with out-of-distribution (OOD) sensitive attributes. Specifically, we define a novel notion called counterfactual group fairness (CGF) from a unified view of group fairness and counterfactual fairness. From a causal view, we treat the sensitive attributes as confounders, extract and reconstruct the exogenous and endogenous parts of user embeddings under the constraint of CGF. The proposed causal embedding learning framework is model-agnostic, which can help improve CGF for existing recommendation models. Experimental results demonstrated the effectiveness of MACE in counterfactually group-fair recommendation.

References

[1] D. Pedreschi, S. Ruggieri, and F. Turini, “Measuring discrimination in socially-sensitive decision records,” in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 581–592.

[2] H. Xue, X. Dai, J. Zhang, S. Huang, and J. Chen, “Deep matrix factorization models for recommender systems,” in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3203–3209.

[3] G. Zhou et al., “Deep interest network for click-through rate prediction,” in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2018, pp. 1059–1068.

[4] J. Xu, X. He, and H. Li, “Deep learning for matching in search and recommendation,” *Found. Trends Inf. Retrieval*, vol. 14, no. 2-3, pp. 102–288, 2020.

[5] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

[6] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, “A theory of learning from different domains,” *Mach. Learn.*, vol. 79, no. 1, pp. 151–175, 2010.

[7] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18583–18599.

[8] J. Yu et al., “Influence function for unbiased recommendation,” in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1929–1932.

[9] Z. Shen et al., “Towards out-of-distribution generalization: A survey,” 2021, *arXiv:2108.13624*.

[10] Y. Li, H. Chen, Z. Fu, Y. Ge, and Y. Zhang, “User-oriented fairness in recommendation,” in *Proc. Web Conf.*, 2021, pp. 624–632.

[11] M. Morik, A. Singh, J. Hong, and T. Joachims, “Controlling fairness and bias in dynamic learning-to-rank,” in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 429–438.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel, “Fairness through awareness,” in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, 2012, pp. 214–226.

[13] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, “Fairness in learning: Classic and contextual bandits,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 325–333.

[14] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4066–4076.

[15] Y. Li, H. Chen, S. Xu, Y. Ge, and Y. Zhang, “Towards personalized fairness based on causal notion,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1054–1063.

[16] A. J. Biega, K. P. Gummadi, and G. Weikum, “Equity of attention: Amortizing individual fairness in rankings,” in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2018, pp. 405–414.

[17] A. Khademi, S. Lee, D. Foley, and V. G. Honavar, “Fairness in algorithmic decision making: An excursion through the lens of causality,” in *Proc. World Wide Web Conf.*, 2019, pp. 2907–2914.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[19] S. Caton and C. Haas, “Fairness in machine learning: A survey,” 2020, *arXiv: 2010.04053*.

[20] F. Kamiran and T. Calders, “Data preprocessing techniques for classification without discrimination,” *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.

[21] B. d’Alessandro, C. O’Neil, and T. LaGatta, “Conscientious classification: A data scientist’s guide to discrimination-aware classification,” *Big Data*, vol. 5, no. 2, pp. 120–134, 2017.

[22] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, “Optimized pre-processing for discrimination prevention,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3992–4001.

[23] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel, “Learning adversarially fair and transferable representations,” in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 3381–3390.

[24] C. Chuang and Y. Mroueh, “Fair mixup: Fairness via interpolation,” in *Proc. 9th Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=DNl5s5BXeBn

[25] D. Mandal, S. Deng, S. Jana, J. M. Wing, and D. J. Hsu, “Ensuring fairness beyond the training data,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 18445–18456.

[26] F. Kamiran, A. Karim, and X. Zhang, “Decision theory for discrimination-aware classification,” in *Proc. IEEE 12th Int. Conf. Data Mining*, 2012, pp. 924–929.

[27] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti, “Discrimination- and privacy-aware patterns,” *Data Mining Knowl. Discov.*, vol. 29, no. 6, pp. 1733–1782, 2015.

[28] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3315–3323.

[29] J. Pearl, “The seven tools of causal inference, with reflections on machine learning,” *Commun. ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[30] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, “Invariant risk minimization,” 2019, *arXiv: 1907.02893*.

[31] B. Schölkopf et al., “Toward causal representation learning,” *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.

[32] K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang, and F. Wang, “Treatment effect estimation with data-driven variable decomposition,” in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 140–146.

[33] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, and D. Jin, “Disentangling user interest and conformity for recommendation with causal embedding,” in *Proc. Web Conf.*, 2021, pp. 2980–2991.

[34] D. Liu et al., “Mitigating confounding bias in recommendation via information bottleneck,” in *Proc. 15th ACM Conf. Recommender Syst.*, 2021, pp. 351–360.

[35] S. Yao and B. Huang, “Beyond parity: Fairness objectives for collaborative filtering,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2925–2934.

[36] H. A. Rahmani, M. Naghiaei, M. Dehghan, and M. Aliannejadi, “Experiments on generalizability of user-oriented fairness in recommender systems,” in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2022, pp. 2755–2764.

[37] C. Zhao, L. Wu, P. Shao, K. Zhang, R. Hong, and M. Wang, “Fair representation learning for recommendation: A mutual information perspective,” in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4911–4919.

[38] L. Xiao, Z. Min, Z. Yongfeng, G. Zhaoquan, L. Yiqun, and M. Shaoping, “Fairness-aware group recommendation with Pareto-efficiency,” in *Proc. 11th ACM Conf. Recommender Syst.*, 2017, pp. 107–115.

[39] W. Huang, L. Zhang, and X. Wu, “Achieving counterfactual fairness for causal bandit,” in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 6952–6959.

[40] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[41] V. Chernozhukov, G. W. Imbens, and W. K. Newey, "Instrumental variable estimation of nonseparable models," *J. Econometrics*, vol. 139, no. 1, pp. 4–14, 2007.

[42] B. A, C. D, C. V, and H. C, "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, vol. 80, no. 6, pp. 2369–2429, 2012.

[43] J. D. Angrist and J.-S. Pischke, *Mostly Harmless Econometrics*. Princeton, NJ, USA: Princeton Univ. Press, 2008.

[44] L. Xu, Y. Chen, S. Srinivasan, N. de Freitas, A. Doucet, and A. Gretton, "Learning deep features in instrumental variable regression," in *Proc. 9th Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=sy4Kg_ZQmS7

[45] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proc. 10th ACM Int. Conf. Web Search Data Mining*, 2017, pp. 781–789.

[46] M. Dudík, J. Langford, and L. Li, "Doubly robust policy evaluation and learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1097–1104.

[47] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning," fairmlbook.org, 2019. [Online]. Available: http://www.fairmlbook.org

[48] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "CLUB: A contrastive log-ratio upper bound of mutual information," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1779–1788.

[49] J. S. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1414–1423.

[50] Y. Zhang et al., "Causal intervention for leveraging popularity bias in recommendation," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 11–20.

[51] Y. Collette and P. Siarry, *Multiobjective Optimization: Principles and Case Studies*. Berlin, Germany: Springer Science & Business Media, 2013.

[52] J. D. Angrist, G. W. Imbens, and D. B. Rubin, "Identification of causal effects using instrumental variables," *J. Amer. Stat. Assoc.*, vol. 91, no. 434, pp. 444–455, 1996.

[53] J. H. Stock and F. Trebbi, "Retrospectives: Who invented instrumental variable regression?," *J. Econ. Perspectives*, vol. 17, no. 3, pp. 177–194, 2003.

[54] R. Misra, M. Wan, and J. McAuley, "Decomposing fit semantics for product size recommendation in metric spaces," in *Proc. 12th ACM Conf. Recommender Syst.*, 2018, pp. 422–426.

[55] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for youtube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, 2016, pp. 191–198.

[56] H.-T. Cheng et al., "Wide & deep learning for recommender systems," in *Proc. 1st Workshop Deep Learn. Recommender Syst.*, 2016, pp. 7–10.

Xiao Zhang is an assistant professor with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests include online learning, trustworthy machine learning, and information retrieval. He has published more than 40 papers on top-tier conferences and journals in artificial intelligence, e.g., NeurIPS, ICML, KDD, SIGIR, AAAI, IJCAI, ICDE, WWW, VLDB, etc.

Teng Shi is currently working toward the PhD degree of artificial intelligence with the Gaoling School of Artificial Intelligence, Renmin University of China (RUC). His current research interests lie at the intersection of trustworthy machine learning and causal learning, especially their applications in recommender systems and information retrieval.

Jun Xu (Member, IEEE) is a professor with the Gaoling School of Artificial Intelligence, Renmin University of China. His research interests focus on learning to rank and semantic matching in web search. He served or is serving as SPC for SIGIR, WWW, and AAAI, editorial board member for *Journal of the Association for Information Science and Technology*, and associate editor for *ACM Transactions on Intelligent Systems and Technology*. He has won the Test of Time Award Honorable Mention in SIGIR (2019), Best Paper Award in AIRS (2010) and Best Paper

Zhenhua Dong received the PhD degree in computer science from Nankai University, China, in 2012. He is currently a principal researcher of Huawei Noah's Ark Lab. His research interests include recommender system, counterfactual learning, causal information retrieval, and their applications. He has published papers in refereed conferences and journals such as AAAI, CIKM, ICDE, RecSys, SIGIR, WWW, *IEEE Transactions on Knowledge and Data Engineering*, etc.

Ji-Rong Wen (Senior Member, IEEE) is a professor of the Renmin University of China (RUC). He is also the dean of the School of Information and executive dean of the Gaoling School of Artificial Intelligence with RUC. His main research interests include information retrieval, data mining, and machine learning. He was a senior researcher and group Manager of the Web Search and Mining Group with Microsoft Research Asia (MSRA).