Evaluating, Understanding, and Mitigating Unfairness in Recommender Systems

Sirui Yao

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Bert Huang, Chair
Naren Ramakrishnan, Co-Chair
B.Aditya Prakash, Member
Chandan K. Reddy, Member
Alex Beutel, Member

May 4, 2021
Blacksburg, Virginia

Keywords: Algorithmic Fairness, Recommender Systems, Matrix Factorization,
Personalized Regularization, Long-term Equality

Evaluating, Understanding, and Mitigating Unfairness in Recommender Systems

Sirui Yao

(ABSTRACT)

Recommender systems are information filtering tools that discover potential matchings between users and items and benefit both parties. This benefit can be considered a social resource that should be equitably allocated across users and items, especially in critical domains such as education and employment. Biases and unfairness in recommendations raise both ethical and legal concerns. In this dissertation, we investigate the concept of unfairness in the context of recommender systems. In particular, we study appropriate unfairness evaluation metrics, examine the relation between bias in recommender models and inequality in the underlying population, as well as propose effective unfairness mitigation approaches.

We start with exploring the implication of fairness in recommendation and formulating unfairness evaluation metrics. We focus on the task of rating prediction. We identify the insufficiency of demographic parity for scenarios where the target variable is justifiably dependent on demographic features. Then we propose an alternative set of unfairness metrics that measured based on how much the average predicted ratings deviate from average true ratings. We also reduce these unfairness in matrix factorization (MF) models by explicitly adding them as penalty terms to learning objectives.

Next, we target a form of unfairness in matrix factorization models observed as disparate model performance across user groups. We identify four types of biases in the training data that contribute to higher subpopulation error. Then we propose personalized regularization learning (PRL), which learns personalized regularization parameters that directly address the data biases. PRL poses the hyperparameter search problem as a secondary learning task. It enables back-propagation to learn the personalized regularization parameters by leveraging the closed-form solutions of alternating least squares (ALS) to solve MF. Furthermore, the learned parameters are interpretable and provide insights into how fairness is improved.

Third, we conduct a theoretical analysis on the long-term dynamics of inequality in the underlying population, in terms of the fitting between users and items. We view the task of recommendation as solving a set of classification problems through threshold policies. We mathematically formulate the transition dynamics of user-item fit in one step of recommendation. Then we prove that a system with the formulated dynamics always has at least one equilibrium, and we provide sufficient conditions for the equilibrium to be unique. We also show that, depending on the item category relationships and the recommendation policies, recommendations in one item category can reshape the user-item fit in another item category.

To summarize, in this research, we examine different fairness criteria in rating prediction and recommendation, study the dynamic of interactions between recommender systems and users, and propose mitigation methods to promote fairness and equality.

Evaluating, Understanding, and Mitigating Unfairness in Recommender Systems

Sirui Yao

(GENERAL AUDIENCE ABSTRACT)

Recommender systems are information filtering tools that discover potential matching between users and items. However, a recommender system, if not properly built, may not treat users and items equitably, which raises ethical and legal concerns. In this research, we explore the implication of fairness in the context of recommender systems, study the relation between unfairness in recommender output and inequality in the underlying population, and propose effective unfairness mitigation approaches.

We start with finding unfairness metrics appropriate for recommender systems. We focus on the task of rating prediction, which is a crucial step in recommender systems. We propose a set of unfairness metrics measured as the disparity in how much predictions deviate from the ground truth ratings. We also offer a mitigation method to reduce these forms of unfairness in matrix factorization models

Next, we look deeper into the factors that contribute to error-based unfairness in matrix factorization models and identify four types of biases that contribute to higher subpopulation error. Then we propose personalized regularization learning (PRL), which is a mitigation strategy that learns personalized regularization parameters to directly addresses data biases. The learned per-user regularization parameters are interpretable and provide insight into how fairness is improved.

Third, we conduct a theoretical study on the long-term dynamics of the inequality in the fitting (e.g., interest, qualification, etc.) between users and items. We first mathematically formulate the transition dynamics of user-item fit in one step of recommendation. Then we discuss the existence and uniqueness of system equilibrium as the one-step dynamics repeat. We also show that, depending on the relation between item categories and the recommendation policies (unconstrained or fair), recommendations in one item category can reshape the user-item fit in another item category.

In summary, we examine different fairness criteria in rating prediction and recommendation, study the dynamics of interactions between recommender systems and users, and propose mitigation methods to promote fairness and equality.

Dedication

This dissertation is wholeheartedly dedicated to my beloved parents and family, for their endless love, support and encouragement.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Professor Bert Huang for dedicating his time and energy throughout the past six years to help me grow both academically and personally. Not only is he a tremendous mentor but also a great friend who always has my best interest at heart. I am truly fortunate and honored to have the opportunity to work with him.

I also very much appreciate Professor Naren Ramakrishnan, Professor Chandan K. Reddy, Professor B.Aditya Prakash, and Dr.Alex Beutel for generously serving as my committee members. I am grateful for their guidance that helps me deepen my knowledge in this dissertation.

Furthermore, I would like to thank my lab mates for their support and company. I cherish the time that we learn and have fun together. I am also lucky to have been surrounded by my sweet friends who always fill my days with laughter and warmth. Thank you all for bringing so much joy into my life.

Last but not least, a huge thank you to my parents, grandparents, uncles, aunts, and cousins back in China, for always being there for me despite the distance. They give me the strength and purpose to become who I am today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Recommender Systems

We introduce recommender systems in this section. Specifically, we discuss the roles and purposes of recommender systems, their success in various applications. Recommender systems are one of the most widely applied machine learning technique. A recommender system identifies the potential matching between users and items, then show the relevant items to users. For example, users receive product recommendations on Amazon, playlist recommendations on music or video platforms such as Netflix [1, 2], YouTube [3] and Spotify [4], as well as content or friend recommendations on social media such as Twitter and Facebook[5]. Recommender systems have also been applied in critical areas such as health care, education and employment to recommend treatments to patients [6], majors and courses to students [7], and job opportunities to job seeker [8, 9, 10].

Recommenders systems are of great significance in the age of information overload where there are often thousands if not millions of users and items [11]. From the customer's point of view, recommender systems help users find what they need from a large item set. From the service providers' point of view, recommender systems help them present their items to the target users who are likely to be interested. Jannach and Adomavicius [12] provide a full list of purposes of building a recommenders sytem, ranging from helping users explore and make decisions, to changing user behavior in desired directions, to increasing user engagement.

It is indisputable that recommender systems have been successful in promoting business[13]. For example, a blog post disclosed by Netflix [1] states that "75 percent of what people watch is from some sort of recommendation", and YouTube reports that 60 percent of the clicks on the home screen are on the recommendations [3]. Netfix also reveals that recommendations led to a measurable increase in user engagement, and that over the years, customer churn decreases by several percentage points with the help of personalization and recommendation service [2].

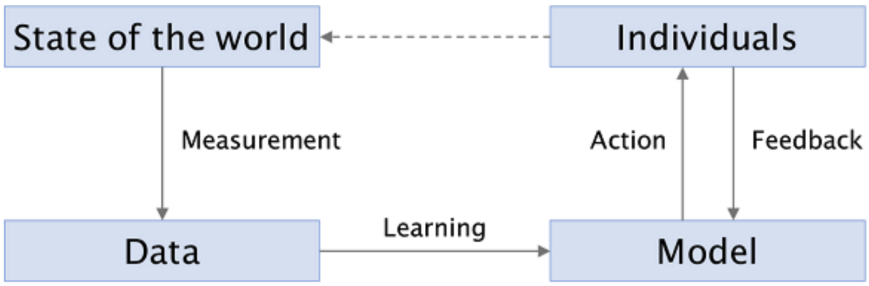

Figure 1.1: The machine learning loop.

1.2 Fairness in Recommender Systems

Fairness is a generic term that describes the quality of being fair. It is often closely related to concepts such as justice and equality, and they all have been extensively studied in philosophy and sociology. In this section, we discuss how fairness is perceived and addressed in the context of machine learning and recommender systems.

1.2.1 The Machine Learning Pipeline

We start with examining the pipeline of implementing a machine learning model. Barocas et al. [14] summarize different stages of a machine learning system as in Figure 1.1.

The first stage is measurement, which is to collect data that reflects the state of the world. In this stage, the real world is characterized as a set of rows, columns, and values. The measurements are based on the observations as well as human decisions in terms of what to measure.

The next stage is learning. This is the process of building a model that learns from the collected data. This process is not simply memorizing the existing examples, but rather to draw general rules and summarize the patterns in the training data, so that the model can make induction and generalize to future unseen cases.

The third stage is taking actions upon individuals based on the output of a trained model. For example, after a model learns to predict user preference towards different items (e.g., rating), the corresponding action can be recommending a list of items to a user, ranked by the predicted ratings. It is worth noting that these actions can also change the individuals and subsequently alter the state of the world. In the same example, after a user receives the recommendations, his or her preference may shift because of the exposure to previously unknown items.

The fourth stage is feedback. Some machine learning systems refine the model depending on how users react to actions. For example, whether a user clicks on the recommended items is

often used as an indicator of the quality of recommendations, which is later used to modify the model.

1.2.2 Disparities in Machine Learning

Barocas et al. [14] also pointed out that disparity can penetrate the entire machine learning pipeline. In the following paragraphs, we discuss how unfairness exist in each of the above mentioned stages.

State of the world. Demographic disparities exist in our society. For example, gender imbalances are often observed in education and employment. In the fields of STEM, Why do these disparities exist? There are many potentially contributing factors, including a history of explicit discrimination, implicit attitudes and stereotypes about gender, or even innate differences in the distribution of certain characteristics by gender. If we're building a machine learning system that screens university applicants or job candidates, we should be aware that the model learns from data that is likely to encode the disparities in the real world. Note that it does not necessarily mean that the outputs of our system will be inaccurate or discriminatory.

Measurement. The process of measurement involves first defining the target variable of interest as well as the predictor variables, then collecting examples and assigning values to the variables based on observations. However, the definition of variables can be subjective, which opens up the chance for human bias to slip in. For example, to build a hiring system that predicts "a good employee", what is an objective measurement? If we rely on the existing performance reviews, then the target variable will inevitably inherit the biases present in managers' evaluations.

Learning. We have discussed that training data reflects the disparities from the real world and the measurement process. These patterns are blended with the knowledge that we wish to learn using machine learning. However, the learning algorithms are not taught to distinguish between them. Therefore, when we train a model using biased training data, machine learning may also extract biases in the same way that it extracts knowledge. Furthermore, the model may even exacerbate unfairness by introducing disparities that are not in the training data.

Action. Since actions are the subsequent decisions based on the output of a machine learning model, when the model is biased, so will the actions. For example, if a rating predictor has disparate error rates for different groups, then the quality of recommendations will also be different for these groups. Further, as we mentioned previously that the user

preferences can be changed by recommendations, if users groups change differently, new disparities will be introduced in the underlying population.

Feedback. It is important to consider the bias in user feedback when we collect use it to refine a machine learning system. For example, in a recommender system, users are more likely to click on items that appeal to them. Therefore, a recommender is likely to receive positive feedback if it recommends items that are similar to the users current preference. Subsequently, the recommender is more encouraged to do so given the positive feedback. As a consequence, the model validates and enhances the stereotypes it has already learned from existing disparities.

1.2.3 Fairness Criteria

Many fairness criteria have been proposed to formalize different forms of disparities. These different intuitions are distinguished by whether the observed disparities can be considered discrimination. Specifically, it boils down to two key questions: whether the disparities are justified and whether they are harmful. Here we discuss the formal definitions of two representative fairness criteria, and how they are mathematically formulated. These two criteria are widely adopted and serve as the foundation of many domain-specific fairness criteria.

Independence. Independence is also known as demographic parity. Let Y be the output of machine learning model and S be the sensitive attributes. Suppose the model makes binary decision and two user groups are involved, i,e., $Y \in \{0, 1\}, S \in \{0, 1\}$, independence is mathematically formulated as

$$\mathbb{P}(Y = 1|S = 0) = \mathbb{P}(Y = 1|S = 1).$$

Independence expresses the belief about equality in human nature. For example, some may believe the qualification for a job should be independent of demographic attributes. However, for scenarios where two groups are believed to be heterogeneous, decisions that satisfies independence are undesirable.

Separation Separation aligns with the principle of equalized opportunity (or equalized odds). We further denote $\hat{Y} \in \{0, 1\}$ to be the true value of the target variable, then separation is mathematically formulated as

$$\mathbb{P}(Y = 1|S = 0, \hat{Y} = 0) = \mathbb{P}(Y = 1|S = 1, \hat{Y} = 0),$$

$$\mathbb{P}(Y = 1|S = 0, \hat{Y} = 1) = \mathbb{P}(Y = 1|S = 1, \hat{Y} = 1).$$

Separation is appropriate for scenarios where the sensitive attributes may be justifiably correlated with the target variable. The separation criterion allows correlation between the target variable and the sensitive attribute to the extent that it is justifiable by the target variable. For example, if one group indeed has a higher default rate on loans than another, A bank might justify different lending rates for these groups for business necessity.

1.2.4 Disparities in Recommendation

Now we talk about the forms of disparities that are specific to recommendations. In particular, fairness in recommendations can be considered from both the user side or the item side.

User unfairness We first discuss how a recommender can unfairly treat users. One form of user unfairness is the unjustified disparity in the recommendations that different groups of users receive. For example, a job recommender may be more likely to recommend lower-paying jobs to woman than man even when they are equally-qualified [15]. This may due to the male dominance in most high-paying jobs, male users might be more likely to click optimistically on high-paying jobs [16]. This form of unfairness will reinforce the existing stereotypes in the real world. Another form of user unfairness refers to the disparity in the quality of recommendations [17]. For example, a recommender may provide worse recommendations for one user group than the others. This often happens when a user subgroup is underrepresented, therefore, a model trained to minimize a global prediction loss pays less attention to learning the preference of the minority. Further, this form of unfairness will reinforce over time because high prediction error drives user departure, so the minority group will further shrink, leading to even greater disparity in model performance.

Item unfairness The two forms of user unfairness can also occur in the items, or more specifically, the service providers of the items. First, a recommender may be biased against some items by a) being less likely to recommend them to users, or b) reduce their exposure to users by placing them at lower rank on a sorted list of recommendations [18, 19, 20]. For example, in 2019, a group of content creators sued YouTube was sued by a group of content creators for suppressing the reach of LGBT-focused videos [21]. Second, we may also observe the recommendations to be less accurate for some items [22, 23]. The items that are ill-modeled are unfairly treated by the recommender because they are less likely to reach their target users.

1.2.5 Fairness Interventions

It may be tempting to think that we can ensure the impartiality of the resulting classifier by simply removing or ignoring sensitive attributes. For example, Suppose we're building a model for making loan approval decisions. Can we withhold gender from the data so that the model decisions can't be gender biased? Unfortunately, it's not that simple. This is due to the problem of proxies or redundant encoding, which means typically we have many features in data that are more or less correlated with the sensitive attribute.

Generally, there are three main strategies to mitigate unfairness in a machine learning model that interfere different stages of modeling.

- Pre-processing. Pre-processing directly adjust the training data to remove the dependency from the target to sensitive features. The advantage of this approach is that once the data is projected to the new feature space, it is generally agnostic to downstream applications.
- Mid-processing. Mid-processing enforces fairness constraints through optimization to alter the trained machine learning model. This technique requires access to the raw data as well as the training pipeline. It is also usually tied to a specific model and application.
- Post-processing. Post-processing changes the output of a model after it is already trained so as to satisfy certain fairness criterion. This technique works for any black-box classifier and does not require knowledge of the inner workings of the training pipeline. However, since the post-processing step is performed after modeling and usually for the sole purpose of satisfying fairness criteria, this strategy is likely to hurts utility the most.

1.3 Research Goals

This goal of this research is to study methods for measuring, understanding, and mitigating unfairness in recommender systems. We approach the problem from the following aspects.

1. Definition and evaluation. As we previously discussed in the task of classification, fairness can be interpreted differently therefore its definition is not unique [24, 25, 26]. This is also true for the task of recommendation. The selection of fairness criteria requires thorough understanding of the specific applications. We will examine different notions of fairness and propose metrics that promote more comprehensive evaluation of unfairness in recommender systems.

2. Mitigation and intervention. Unfair recommendations raise both ethical and legal concerns. Thus we seek to provide effective strategies that reduces unfairness in recommendation models and maintain the system utility at the same time.

3. Comprehension. A recommender system and the entities it hosts (users, items, etc.,) form an environment that evolves over time throughout the interaction among the components [27, 28, 29, 30]. We believe it is important to understand the dynamic of this environment because it provides valuable insight for anticipating the future and possibly shape it towards a more fair state.

1.4 Contributions

The main contributions of this research are as follows.

1. We focus on the task of rating prediction and discuss the drawback of enforcing demographic parity when user interest is justifiably different across demographic groups. As alternatives, we offer a set of fairness metrics that are measured as the disparity in the (signed or unsigned) difference between average prediction and average true ratings. We also construct corresponding fairness objectives with fairness penalty terms to optimize fairness in matrix factorization models. We run experiments on both synthetic and real datasets. Results show that unfairness in the predicted ratings can be effectively decreased through optimization.

2. We investigate different types of data biases that cause discrepancy in prediction accuracy across user subgroups matrix factorization models. We then identify the insufficiency of a global optimal regularization parameter in those situations and introduced personalized regularization learning (PRL) to promote fairer allocation of error in prediction. Results on real dataset with five user splits show that PRL outperforms existing methods in reducing error-based unfairness. Moreover, we interpret the learned personalized regularization parameters to understand how fairness is improved.

3. We study the long-term dynamics of inequality in the fit between users and items in recommender systems. We characterize the transition of user-item fit as it is affected by recommender decisions as a set of probabilities. Then we examine how the fit changes depending on different recommendation policies and the relationship between item categories. We also validate our theoretical analysis using simulation on synthetic users and items.

1.5 Outline

The remaining chapters are organized as follows.

- Chapter 2 reviews the prior literature related to fairness and recommender systems.
- Chapter 3 explains the motivation and implications of error-based fairness, shows with experiments on synthetic and real data that error-based fairness can be reduced through optimizing learning objectives with unfairness penalty terms.
- Chapter 4 shows four types of bias that lead to disparity in model performance and introduces the personalized regularization learning technique to reduce unfairness.
- Chapter 5 analyzes the long-term dynamics of inequality in the underlying population and discusses how the fit between users and items changes depending on the relationship between item categories and different recommendation policies.
- Chapter 6 summarizes our work, discusses how different projects connected with each other, and describes direction for future research.

Chapter 2

Literature Review

In this chapter, we first review the prior work on core concepts such as matrix factorization and algorithm fairness. Then we discuss the work related to each chapter in separate sections.

2.1 Recommender System

A plethora of methods have been proposed for modeling recommender systems. Collaborative filtering (CF) [31, 32] methods is a frequently practiced approach which makes recommendations based on the ratings or behavior of other users in the system. The fundamental assumption behind collaborative filtering is that other users' opinions can be selected and aggregated in such a way as to provide a reasonable prediction of the active user's preference. Matrix factorization [33, 34, 35, 36] is an popular approach for collaborative filtering. Matrix factorization projects users and items to a lower dimensional latent feature space through low-rank or low-norm approximation [37]. This factorization is often solved by minimizing a regularized squared reconstruction error. This objective function is non-convex. One approach for optimizing this objective is through gradient descent [38], which has been made especially convenient with the advance of automatic differentiation tools [39]. Another commonly used approach is alternating least squares (ALS) [40, 41]. ALS solves MF by alternating between computing user features while fixing item features and computing item features with user features fixed. Each sub-step is a convex, quadratic minimization, so it can be solved with an closed-form solution. In each iteration, the closed-form solution computes each individual user or item's latent feature independently therefore making ALS easily parallelizable. The other classes of recommendation algorithms are content-based [42], hybrid approaches [43], etc.

2.2 Fairness

Fairness Criteria Demographic parity [44] is a frequently adopted fairness criteria for encouraging fairness, the goal is to achieve statistical parity among groups, in other words, to ensure that the overall proportion of members in the protected group receiving positive (or negative) classifications is identical to the proportion of the population as a whole. In the case of a binary decision $\hat{Y} \in \{0, 1\}$ and a binary protected attribute $A \in \{0, 1\}$, the true label $Y \in \{0, 1\}$. Demographic parity can be formulated as $\Pr\{\hat{Y} = 1|A = 0\} = \Pr\{\hat{Y} = 1|A = 1\}$. Another family of commonly adopted group fairness criteria are equalized odds and equal opportunity proposed by [45]. These criteria allow the predictions $\hat{Y}$ to be independent on A but only through Y. In classification, equal opportunity requires parity in true positive rate across demographic groups, it can be formulated as $\Pr\{\hat{Y} = 1|A = 0, Y = 1\} = \Pr\{\hat{Y} = 1|A = 1, Y = 1\}$; equalized odds is a stronger constraint that requires both true positive and false positive rate to be the same, this means $\Pr\{\hat{Y} = 1|A = 0, Y = y\} = \Pr\{\hat{Y} = 1|A = 1, Y = y\}, y \in \{0, 1\}$. These criteria are measured based on group statistics. Individual fairness criteria [46, 47] instead requires that similar individuals are treated similarly, the decision for an individual should be the same if only the demographic was changed.

Mitigation Methods Generally, fairness mitigation methods can fall into three categories depending on which stage of modeling is intervened: pre-processing, in-processing, and post-processing. Pre-precessing is a data transformation step before training to reduce bias in the data so that the outcome of downstream tasks can be fairer [48, 49]. A typical example of pre-processing is the work by Zemel et al.[44], which is to search for new representations that encodes the original data and simultaneously remove information about group membership. In-processing methods [50, 51, 52] modify the training algorithm, usually by changing the objective function or adding fairness-related constraints. Post-processing [45, 53] modifies the output of a model and does not require access to the training process. A variety of methods have been proposed for different learning tasks such as classification [45], regression [54], structured prediction [55], dimension reduction [56], etc. Mehrabi et al. [57] published a survey that summarize these methods.

Fairness in Recommender Systems Kamishima et al.[58] applies demographic parity in the context of rating prediction as the difference between the mean prediction across user subgroups, and modify matrix factorization objective by adding the difference as a neutrality constraint. They later update the constraint to measure difference in higher moments [59]. Zhu et al. [60] also adopt statistical parity as fairness criteria and propose a tensor-based fairness-award recommendation framework that aims to remove sensitive information from the latent representations. On item fairness, Beutel et al.[22] address the discrepancy in model performance across different items, they split items into focused and unfocused set and regularize them differently. Fairness-aware ranking algorithms [61, 62] are also studied to ensure fair allocation of attention to items. Since recommendation usually involve multiple

stakeholder, such as users, items, recommendation service providers, etc., another line of research studies the complex problem of multi-sided fairness [16, 23, 63], which considers fairness for these parties simultaneously. Furthermore, diversity in recommendation [64, 65, 66, 67] have been extensively studied. Diversity is a related concept with fairness, though it is a natural outcome of removing certain forms of bias such as stereotyping, these studies are usually driven by different motivations. Multi-sidedness [68, 63]. The ecosystem of recommendation service consists of not only users but also items. In some cases, fairness for both parties needs to be considered.

2.3 Error-based Fairness

Bias can arise in algorithms due to undesirable properties in the training data, such as sampling bias. Researchers have shown that sampled ratings have markedly different properties from the users' true preferences [69, 70]. Sampling is heavily influenced by social bias, which results in more missing ratings in some cases than others. This non-random pattern of missing and observed rating data is a potential source of unfairness. For the purpose of improving recommendation accuracy, there are collaborative filtering models [69, 22, 71] that use side information to address the problem of imbalanced data. Schnabel et al. [72] handle selection bias in data through estimation techniques from causal inference.

One type of model bias is bias amplification. Leino et al. [73] define bias amplification as "...a machine learning model learns to predict classes with a greater disparity than the underlying ground truth", they propose to mitigate the overestimation of the importance of weak features through targeted feature selection. Zhao et al. [55] found that in an image recognition dataset, women are 33% more likely to be in cooking images, and a model trained on this dataset increases the disparity to 68%. Bias can also arise in the form of disparity in accuracy. Beutel et al. [22] show that rating prediction in recommender systems have higher accuracy for some items than the others. It has been further observed that the disparities in model accuracy can amplify through repeated use of the model. [74, 75]

2.4 Personalized Regularization

Echoing our observation, Mansoury et al. [76] listed three factors that lead to discrepancy in model performance across users: profile anomaly, profile entropy and profile size. Hashimoto et al. [74] show that through empirical risk minimization, disparity in representation and model performance can enhance each other, and cause a minority group shrinking over time. In the context of recommendations, Beutel et al. [22] observe disparity in rating prediction accuracy across items. They propose a hyperparameter optimization task called focused learning which separate the ill-modeled items and regularize items differently from other items. They optimize the hyperparameters through grid search, and show that this approach

improves accuracy for badly-model items. Our study can be viewed as a direct extension of this work. The difference is that our parameter space is much bigger since we regularization parameters to individual users instead of groups; also we search for the hyperparameters in continuous space through optimization instead of from limited preset discrete values.

Another line of research related is hyperparameter tuning, which can have large impact on model performance [77]. In practice, grid search is often used when the space of hyperparameters is small. Random search can be more efficient because it randomly samples hyperparameter values instead of trying all combinations of the candidate hyperparameters [78, 79]. Bayesian hyperparameter optimization speeds up random search by taking into consideration the past evaluations to decide what areas of hyperparameter space to search next [80]. Gradient-based hyperparameter search is challenging for many machine learning methods when learning is a complex optimization scheme. maclaurin et al. [81] devised a method to compute hyperparameter gradients by analyzing the dynamics of gradient descent. Our approach also uses gradient-based hyperparameter tuning, but, in contrast with this advanced approach, our learning algorithm leverages alternating least square method for matrix factorization and only requires back-propagating through closed-form updates.

Besides, a number of methods have been developed to address the challenging task of minimizing unfairness when demographic information about users is unavailable. Many of these approaches incorporate a search over possible subgroups and mitigating unfairness on these derived groups [53, 74, 82]. We adopt a similar approach in our method's variant designed to handle such settings.

2.5 Long-term Inequality

Our work is related to existing theoretical analysis on the outcome of fairness interventions and how they influencing the underlying population. The work that we are most related to is by Zhang et al. [83] and their analysis are centered around the task of classification. Coate et al. [84] study the effectiveness of affirmative action policies in labor-market under the compound influence of employer beliefs and worker productivity. Mouzannar et al. [85] and Liu et al. [29] explore the possibility and conditions of different outcome with or without fairness intervention, the former studies one-step feedback while the latter studies over long time horizons. Hu et al. [30] is a domain-specific study on labor market under two hiring models, they discuss a stable and self-sustaining market equilibrium and fairness intervention become unnecessary. These work are all theoretical based on high level abstraction of entities and impose strong assumptions. The proposed work distinct from these prior work in that we provide a concrete view of the dynamic with realistic user and model specifications by tracking the system trajectories through simulation. As for other empirical studies, D'Amour et al. [86] simulate the repeated decision making processing and measure fairness in the outcome, they consider the changes in how a model perceives the underlying population. Our simulation instead focus on the dynamic of internal properties of users and enable

analysis with user-level granularity.

There are other simulation work that studies concepts related to fairness such as Diversity and homogeneity. For example, Shi et al. [87] conduct simulations of user-recommender interaction and demonstrate the trade-off between short-term accuracy and long-term diversity of recommendations. Chaney et al. [27] study how feedback loop reinforce recommendation homogeneity, assuming user-item utilities are static. But they are not user-centered and assume users are static. Studies on recommender models that consider user long-term and short-term preference has been studied extensively [88, 89, 28, 90] upholds our assumption that user preference are dynamic. However, these work are oblivious to the cause of user preference change while we focus on interest shift directly stimulated by consumption driven by recommendations. We build user models based on work on user behavior, for example, Zajonc et al. [91] study the mechanism of mere exposure effect on human attitude. Schnabel et al. [92] investigate the relationship between the amount of exploration in recommendation and user satisfaction.

Chapter 3

Error-based Unfairness in Rating Prediction

Enforcing demographic parity on recommendations may strongly conflict with accuracy due to the differences in user preference across demographic groups. Therefore, to respect the pre-existed differences, we propose an alternative set of fairness criteria for recommendations that are error-based and consider unfairness as the disparity introduced through algorithms. Specifically, we focus on rating prediction as the recommendation task, measure unfairness as the difference in the distribution of deviation of predictions from "ground truth" labels, which we assume reflects user true preference. We then discuss two forms of data imbalance that may lead to unfair rating predictions. The error-based unfairness criteria can be optimized by being added as penalty terms to the learning objective. We demonstrate with experiments on synthetic and real data that each fairness metric can be optimized without much degradation in prediction accuracy.

3.1 Introduction

Demographic parity is a common notion of fairness but obviously have its limitations. It aims to achieve statistical equivalence in predictions across demographic groups and disregards the difference already existed. When the true label is indeed dependent on demographic features, one has to trade off between fairness and accuracy. We argue that this trade-off is especially critical in assisting-tools such as recommender systems because users may easily abandon the service if they are unsatisfied with the recommendations. Improper pursuit of Demographic parity can bring irreversible damage to the utility of recommender systems, leading to user departure and loss in profit.

Deciding the justification of observed difference across demographic groups and whether they should be mitigated require deep understanding and careful reasoning of the societal and

historical factors behind them. Therefore, instead of targeting those pre-existed differences, some research have instead focused on preventing bias amplification which refers to the layer of disparity directly introduced through algorithms [45, 55, 93, 94].

Inspired by this line of research in classification, we propose an alternative set of fairness criteria for rating prediction, which is common recommendation task. The common goal of the proposed fairness criteria is to avoid higher directional or unidirectional deviation in predictions from "ground truth" ratings in one group than the others. We make the assumption that the observed ratings accurately reflects users' true preferences, so we expect a fair recommender model to maintain a similar distribution in model predictions as that in training data. This way, we can still allow dependency of predictions on demographic features but only through the true labels.

We consider a running example of unfair recommendation in education, and unfairness that may occur in areas with current gender imbalance, such as science, technology, engineering, and mathematics (STEM) topics. Due to societal and cultural influences, fewer female students currently choose careers in STEM. For example, in 2010, women accounted for only 18% of the bachelor's degrees awarded in computer science [95]. The underrepresentation of women causes historical rating data of computer-science courses to be dominated by men. Consequently, the models trained on these data may be biased toward men by further underestimating women's preferences compared to the ratings provided by students, which we assume accurately reflect their true preferences.

The remainder of this chapter is organized as follows. First, in Section 3.3 we discuss two forms of data imbalance that are likely to lead to unfair recommendations. In Section 3.4, we introduce four new error-based unfairness metrics and give justifications and examples. In Section 3.5, we show that unfairness occurs as data gets more imbalanced, we present results that successfully minimize each form of unfairness, and show improvement with a modified metric. Finally, Section 3.6 concludes the chapter and discuss the limitation of this work.

3.2 Preliminaries

In this section we discuss introduce matrix factorization models. Suppose dataset $\mathcal{D}$ contains n ratings by M users on N items, denoted as $(u_j, i_j, r_j)_{j=1}^{n}$, $u_j \in 1, \ldots, M, i_j \in 1, \ldots, N$. Users can be split into subpopulations by a property S, such as a demographic feature (e.g., gender, race, or age). Subpopulation memberships are denoted with $\Sigma = \{\sigma_u\}_{u=1}^{M}$. The ratings in $\mathcal{D}$ can be represented as a $M \times N$ matrix R where each observed entry r_{ui} represents the rating user u gives to item i. For modeling, R is split into training and testing set, i.e., R^{Train} and R^{Test}. With a specified dimensionality d and regularization parameter λ^*, a matrix factorization model decomposes R^{Train} into user and item matrices $P \in \mathbb{R}^{M \times d}$ and $Q \in \mathbb{R}^{N \times d}$. The rows of these matrices can be viewed as user and item coordinates in

a d-dimensional latent feature space. The uth row of P, denoted as p_u, is the latent feature of user u. Likewise, the ith row of Q, denoted as q_i, is the latent feature of item i. With a matrix factorization model, ratings are predicted as $\hat{r}_{ui} = p_u q_i^\intercal$. The parameters p_u and q_i are obtained by minimizing the regularized squared reconstruction error

$$J(P,Q) = \sum_{r_{ui} \in R^{\text{Train}}} (r_{ui} - \hat{r}_{ui})^2 + \frac{\lambda^*}{2} \left(\sum_{u=1}^{M} p_u p_u^\intercal + \sum_{i=1}^{N} q_i q_i^\intercal \right). \tag{3.1}$$

3.3 Data Imbalance

In this section, we describe a process through which matrix factorization leads to unfair recommendations. Such unfairness can occur with imbalanced data. We identify two forms of underrepresentation: *population imbalance* and *observation bias*. We later demonstrate that either leads to unfair recommendation, and both forms together lead to worse unfairness. In our discussion, we use a running example of course recommendation, highlighting effects of underrepresentation in STEM education.

Population imbalance occurs when different types of users occur in the dataset with varied frequencies. For example, we consider four types of users defined by two aspects. First, each individual identifies with a gender. For simplicity, we only consider binary gender identities, though in this example, it would also be appropriate to consider men as one gender group and women and all non-binary gender identities as the second group. Second, each individual is either someone who enjoys and would excel in STEM topics or someone who does and would not. Population imbalance occurs in STEM education when, because of systemic bias or other societal problems, there may be significantly fewer women who succeed in STEM (WS) than those who do not (W), and because of converse societal unfairness, there may be more men who succeed in STEM (MS) than those who do not (M). This four-way separation of user groups is not available to the recommender system, which instead may only know the gender group of each user, but not their proclivity for STEM.

Observation bias is a related but distinct form of data imbalance, in which certain types of users may have different tendencies to rate different types of items. This bias is often part of a feedback loop involving existing methods of recommendation, whether by algorithms or by humans. If an individual is never recommended a particular item, they will likely never provide rating data for that item. Therefore, algorithms will never be able to directly learn about this preference relationship. In the education example, if women are rarely recommended to take STEM courses, there may be significantly less training data about women in STEM courses.

We simulate these two types of data bias with two stochastic block models [96]. We create one block model that determines the probability that an individual in a particular user group likes an item in a particular item group. The group ratios may be non-uniform, leading to

population imbalance. We then use a second block model to determine the probability that an individual in a user group rates an item in an item group. Non-uniformity in the second block model will lead to observation bias.

Formally, let matrix $\boldsymbol{L} \in [0,1]^{|g|\times|h|}$ be the block-model parameters for rating probability. For the ith user and the jth item, the probability of $r_{ij} = +1$ is $L_{(g_i,h_j)}$, and otherwise $r_{ij} = -1$. Morever, let $\boldsymbol{O} \in [0,1]^{|g|\times|h|}$ be such that the probability of observing r_{ij} is $O_{(g_i,h_j)}$.

3.4 Fairness Metrics and Objectives

In this section, we present four new unfairness metrics for preference prediction, all measuring a discrepancy between the prediction behavior for disadvantaged users and advantaged users. Each metric captures a different type of unfairness that may have different consequences. We describe the mathematical formulation of each metric, its justification, and examples of consequences the metric may indicate. We consider a binary group feature and refer to disadvantaged and advantaged groups, which may represent women and men in our education example. We list all used notations in Table 3.1.

Table 3.1: Table of notation in Chapter 3.

Variable	Definition
$\mathcal{D}$	dataset of (user, item, rating) tuples
M	Number of users
N	Number of items
R	Groundtruth rating matrix
R^{Train}	Groundtruth rating matrix for training
R^{Test}	Groundtruth rating matrix for testing
$\hat{R}$	Predicted rating matrix
d	Matrix factorization model dimensionality
P	User features
Q	Item features
b_u	User bias
b_i	Item bias
E	Overall RMSE
E_g	RMSE of group g
Σ	Subpopulations users belong to
λ	Global regularization weight

3.4.1 Fairness Metrics

Let g be the disadvantaged group and $\neg g$ be the advantaged group. The first metric is *value unfairness*, which measures inconsistency in signed estimation error across the user types, computed as

$$U_{\text{val}} = \frac{1}{N} \sum_{j=1}^{N} \left| \left(\mathrm{E}_g \left[\hat{r}\right]_j - \mathrm{E}_g \left[r\right]_j \right) - \left(\mathrm{E}_{\neg g} \left[\hat{r}\right]_j - \mathrm{E}_{\neg g} \left[r\right]_j \right) \right| \ , \tag{3.2}$$

where $\mathrm{E}_g \left[\hat{r}\right]_j$ is the average predicted score for the jth item from disadvantaged users, $\mathrm{E}_{\neg g} \left[\hat{r}\right]_j$ is the average predicted score for advantaged users, and $\mathrm{E}_g \left[r\right]_j$ and $\mathrm{E}_{\neg g} \left[r\right]_j$ are the average ratings for the disadvantaged and advantaged users, respectively. Precisely, the quantity $\mathrm{E}_g \left[\hat{r}\right]_j$ is computed as

$$\mathrm{E}_g \left[\hat{r}\right]_j := \frac{1}{|\{i : (r_{ui} \in R) \wedge g_i\}|} \sum_{i:(r_{ui} \in R) \wedge g_i} \hat{r}_{ij} \ , \tag{3.3}$$

and the other averages are computed analogously.

Value unfairness occurs when one class of user is consistently given higher or lower predictions than their true preferences. If the errors in prediction are evenly balanced between overestimation and underestimation or if both classes of users have the same direction and magnitude of error, the value unfairness becomes small. Value unfairness becomes large when predictions for one class are consistently overestimated and predictions for the other class are consistently underestimated. For example, in a course recommender, value unfairness may manifest in male students being recommended STEM courses even when they are not interested in STEM topics and female students not being recommended STEM courses even if they are interested in STEM topics.

The second metric is *absolute unfairness*, which measures inconsistency in absolute estimation error across user types, computed as

$$U_{\text{abs}} = \frac{1}{N} \sum_{j=1}^{N} \left| \left| \mathrm{E}_g \left[\hat{r}\right]_j - \mathrm{E}_g \left[r\right]_j \right| - \left| \mathrm{E}_{\neg g} \left[\hat{r}\right]_j - \mathrm{E}_{\neg g} \left[r\right]_j \right| \right| \ . \tag{3.4}$$

Absolute unfairness is unsigned, so it captures a single statistic representing only the magnitude but not the direction of systematic deviation for each user type. For example, if female students are given predictions 0.5 points below their true preferences and male students are given predictions 0.5 points above their true preferences, there is no absolute unfairness. Absolute unfairness is high when error is not evenly distributed in user types, which means one type of user has the unfair advantage of good recommendation, while the other user type has poor recommendation.

The third metric is *underestimation unfairness*, which measures inconsistency in how much the predictions underestimate the true ratings:

$$U_{\text{under}} = \frac{1}{N}\sum_{j=1}^{N}\left|\max\{0, \mathrm{E}_g\left[r\right]_j - \mathrm{E}_g\left[\hat{r}\right]_j\} - \max\{0, \mathrm{E}_{\neg g}\left[r\right]_j - \mathrm{E}_{\neg g}\left[\hat{r}\right]_j\}\right| \ . \tag{3.5}$$

Underestimation unfairness is important in settings where missing recommendations are more critical than extra recommendations. For example, underestimation could lead to a top student not being recommended to explore a topic they would excel in.

Conversely, the fourth new metric is *overestimation unfairness*, which measures inconsistency in how much the predictions overestimate the true ratings:

$$U_{\text{over}} = \frac{1}{N}\sum_{j=1}^{N}\left|\max\{0, \mathrm{E}_g\left[\hat{r}\right]_j - \mathrm{E}_g\left[r\right]_j\} - \max\{0, \mathrm{E}_{\neg g}\left[\hat{r}\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\}\right| \ . \tag{3.6}$$

Overestimation unfairness may be important in settings where users may be overwhelmed by recommendations, so providing too many recommendations would be especially detrimental. For example, if users must invest large amounts of time to evaluate each recommended item, overestimating essentially costs the user time. Thus, uneven amounts of overestimation could cost one type of user more time than the other.

We restate the *parity* unfairness measure introduced by Kamishima et al. [58], it can be computed as the absolute difference between the overall average ratings of disadvantaged users and those of advantaged users:

$$U_{\text{par}} = \left(\mathrm{E}_g\left[\hat{r}\right] - \mathrm{E}_{\neg g}\left[\hat{r}\right]\right)^2 \ . \tag{3.7}$$

3.4.2 Fairness Objectives

Each of these metrics has a straightforward subgradient and can be optimized by various subgradient optimization techniques. We augment the learning objective by adding a smoothed variation of a fairness metric based on the Huber loss [97], where the outer absolute value is replaced with the squared difference if it is less than 1. We solve for a local minimum, i.e,

$$\min_{\boldsymbol{P},\boldsymbol{Q}} \ J(P, Q) + \gamma U \ . \tag{3.8}$$

The smoothed penalty helps reduce discontinuities in the objective, making optimization more efficient. γ is a scalar trade-off term to weight the fairness against the loss. In our experiments, we use equal weighting, so $\gamma = 1$.

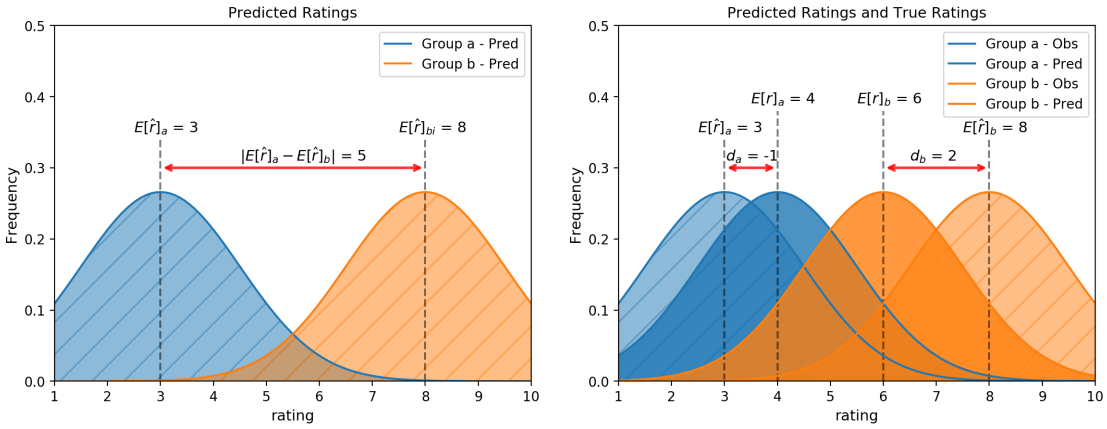

Figure 3.1: Illustration of parity (left) and error-based unfairness (right). Parity is measured based on the difference in predicted ratings. Error-based unfairness is measured based on the deviation of predicted ratings from true ratings.

3.5 Experiments

We run experiments on synthetic data based on the simulated course-recommendation scenario and real movie rating data [98]. For each experiment, we investigate whether the learning objectives augmented with unfairness penalties successfully reduce unfairness.

3.5.1 Synthetic Data

In our synthetic experiments, we generate simulated course-recommendation data from a block model as described in Section 3.3. We consider four user groups $g \in \{\text{W}, \text{WS}, \text{M}, \text{MS}\}$ and three item groups $h \in \{\text{Fem}, \text{STEM}, \text{Masc}\}$. The user groups can be thought of as women who do not enjoy STEM topics (W), women who do enjoy STEM topics (WS), men who do not enjoy STEM topics (M), and men who do (MS). The item groups can be thought of as courses that tend to appeal to most women (Fem), STEM courses, and courses that tend to appeal to most men (Masc). Based on these groups, we consider the rating block model

$$\boldsymbol{L} = \left[\begin{array}{c|ccc} & \text{Fem} & \text{STEM} & \text{Masc} \\ \hline \text{W} & 0.8 & 0.2 & 0.2 \\ \text{WS} & 0.8 & 0.8 & 0.2 \\ \text{MS} & 0.2 & 0.8 & 0.8 \\ \text{M} & 0.2 & 0.2 & 0.8 \end{array}\right]. \tag{3.9}$$

We also consider two observation block models: one with uniform observation probability across all groups $\boldsymbol{O}^{\text{uni}} = [0.4]^{4\times 3}$ and one with unbalanced observation probability inspired

by how students are often encouraged to take certain courses

$$\boldsymbol{O}^{\text{bias}} = \left[\begin{array}{c|ccc} & \text{Fem} & \text{STEM} & \text{Masc} \\ \hline \text{W} & 0.6 & 0.2 & 0.1 \\ \text{WS} & 0.3 & 0.4 & 0.2 \\ \text{MS} & 0.1 & 0.3 & 0.5 \\ \text{M} & 0.05 & 0.5 & 0.35 \end{array}\right]. \tag{3.10}$$

We define two different user group distributions: one in which each of the four groups is exactly a quarter of the population, and an imbalanced setting where 0.4 of the population is in W, 0.1 in WS, 0.4 in MS, and 0.1 in M. This heavy imbalance is inspired by some of the severe gender imbalances in certain STEM areas today.

For each experiment, we select an observation matrix and user group distribution, generate 400 users and 300 items, and sample preferences and observations of those preferences from the block models. Training on these ratings, we evaluate on the remaining entries of the rating matrix, comparing the predicted rating against the true expected rating, $2L_{(g_i,h_j)} - 1$.

Unfairness from different types of underrepresentation

Using standard matrix factorization, we measure the various unfairness metrics under the different sampling conditions. We average over five random trials and plot the average score in Figure 3.2. We label the settings as follows: uniform user groups and uniform observation probabilities (U), uniform groups and biased observation probabilities (O), biased user group populations and uniform observations (P), and biased populations and biased observations (P+O).

The statistics demonstrate that each type of underrepresentation contributes to various forms of unfairness. For all metrics except parity, there is a strict order of unfairness: uniform data is the most fair; biased observations is the next most fair; biased populations is worse; and biasing the populations and observations causes the most unfairness. The squared rating error also follows this same trend. In contrast, non-parity behaves differently, in that it is heavily amplified by biased observations but seems unaffected by biased populations. Note that though non-parity is high when the observations are imbalanced, because of the imbalance in the observations, one should actually expect non-parity in the labeled ratings, so it a high non-parity score does not necessarily indicate an unfair situation. The other unfairness metrics, on the other hand, describe examples of unfair behavior by the rating predictor. These tests verify that unfairness can occur with imbalanced populations or observations, even when the measured ratings accurately represent user preferences.

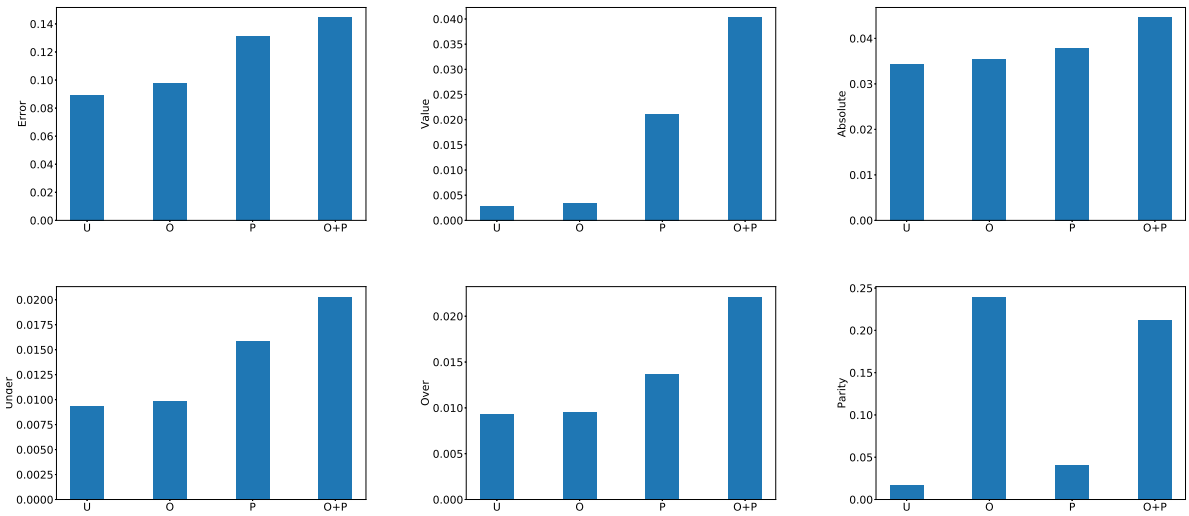

Figure 3.2: Average unfairness scores for standard matrix factorization on synthetic data generated from different underrepresentation schemes. For each metric, the four sampling schemes are uniform (U), biased observations (O), biased populations (P), and both biases (O+P). The reconstruction error and the first four unfairness metrics follow the same trend, while non-parity exhibits different behavior.

Optimization of unfairness metrics

As before, we generate rating data using the block model under the most imbalanced setting: The user populations are imbalanced, and the sampling rate is skewed. We provide the sampled ratings to the matrix factorization algorithms and evaluate on the remaining entries of the expected rating matrix. We again use two-dimensional vectors to represent the users and items, a regularization term of $\lambda = 10^{-3}$, and optimize for 250 iterations using the full gradient. We generate three datasets each and measure squared reconstruction error and the six unfairness metrics.

The results are listed in Table 3.2. For each metric, we print in bold the best average score and any scores that are not statistically significantly distinct according to paired t-tests with threshold 0.05. The results indicate that the learning algorithm successfully minimizes the unfairness penalties, generalizing to unseen, held-out user-item pairs. And surprisingly, reducing any unfairness metric even leads to a significant decrease in reconstruction error. However, since decrease in error is not observed in experiment results on real data, we are hesitant to draw any conclusions.

While optimizing each metric leads to improved performance on itself (see the highlighted entries in Table 3.2), a few trends are worth noting. Optimizing any of our new unfairness metrics almost always reduces the other forms of unfairness. An exception is that optimizing absolute unfairness leads to an increase in underestimation. Value unfairness is closely related

Table 3.2: Average error and unfairness metrics for synthetic data using different fairness objectives. Each row represents a different unfairness penalty, and each column is the measured metric on the expected value of unseen ratings. The best scores and those that are statistically indistinguishable from the best are printed in bold. The unfairness value with the corresponding unfairness penalty are highlighted in yellow.

Unfairness	Error	Value	Absolute	Underestimation	Overestimation
None	0.317 ± 1.3e-02	0.649 ± 1.8e-02	0.443 ± 2.2e-02	0.107 ± 6.5e-03	0.544 ± 2.0e-02
Value	**0.130 ± 1.0e-02**	**0.245 ± 1.4e-02**	0.177 ± 1.5e-02	**0.063 ± 4.1e-03**	**0.199 ± 1.5e-02**
Absolute	0.205 ± 8.8e-03	0.535 ± 1.6e-02	0.267 ± 1.3e-02	0.135 ± 6.2e-03	0.400 ± 1.4e-02
Under	0.269 ± 1.6e-02	0.512 ± 2.3e-02	0.401 ± 2.4e-02	**0.060 ± 3.5e-03**	0.456 ± 2.3e-02
Over	**0.130 ± 6.5e-03**	0.296 ± 1.2e-02	**0.172 ± 1.3e-02**	0.074 ± 6.0e-03	0.228 ± 1.1e-02

to underestimation and overestimation, since optimizing value unfairness is as effective at reducing underestimation and overestimation as directly optimizing them. Also, optimizing value and overestimation are more effective in reducing absolute unfairness than directly optimizing it.

The complexity of computing the unfairness metrics is similar to that of the error computation, which is linear in the number of ratings, so adding the fairness term approximately doubles the training time. In our implementation, learning with fairness terms takes longer because loops and backpropagation introduce extra overhead. For example, with synthetic data of 400 users and 300 items, it takes 13.46 seconds to train a matrix factorization model without any unfairness term and 43.71 seconds for one with value unfairness.

3.5.2 Real Data

We use the Movielens 1M Dataset [98], which contains 1 million ratings (from 1 to 5) by 6,040 users of 3,883 movies. The users are annotated with demographic variables including gender, and the movies are each annotated with a set of genres. We manually selected genres that feature different forms of gender imbalance and only consider movies that list these genres. Then we filter the users to only consider those who rated at least 50 of the selected movies.

The genres we selected are *action*, *crime*, *musical*, *romance*, and *sci-fi*. We selected these genres because they each have a noticeable gender effect in the data. Women rate musical and romance films higher and more frequently than men. Women and men both score action, crime, and sci-fi films about equally, but men rate these film much more frequently. Table 3.3 lists these statistics in detail. After filtering by genre and rating frequency, we have 2,953 users and 1,006 movies in the dataset.

We run five trials in which we randomly split the ratings into training and testing sets, train each objective function on the training set, and evaluate each metric on the testing set. The average scores are listed in Table 3.4, where bold scores again indicate being statistically indistinguishable from the best average score. On real data, the results show that optimizing

Table 3.3: Gender-based statistics of movie genres in Movielens data.

	Romance	Action	Sci-Fi	Musical	Crime
Count	325	425	237	93	142
Ratings per female user	54.79	52.00	31.19	15.04	17.45
Ratings per male user	36.97	82.97	50.46	10.83	23.90
Average rating by women	3.64	3.45	3.42	3.79	3.65
Average rating by men	3.55	3.45	3.44	3.58	3.68

each unfairness metric leads to the best performance on that metric without a significant change in the reconstruction error. As in the synthetic data, optimizing value unfairness leads to significant decrease on under- and overestimation.

Table 3.4: Average error and unfairness metrics for movie-rating data using different fairness objectives. The best scores and those that are statistically indistinguishable from the best are printed in bold. The unfairness value with the corresponding unfairness penalty are highlighted in yellow.

Unfairness	Error	Value	Absolute	Underestimation	Overestimation
None	0.887 ± 1.9e-03	0.234 ± 6.3e-03	0.126 ± 1.7e-03	0.107 ± 1.6e-03	0.153 ± 3.9e-03
Value	0.886 ± 2.2e-03	**0.223 ± 6.9e-03**	0.128 ± 2.2e-03	0.102 ± 1.9e-03	0.148 ± 4.9e-03
Absolute	0.887 ± 2.0e-03	0.235 ± 6.2e-03	**0.124 ± 1.7e-03**	0.110 ± 1.8e-03	0.151 ± 4.2e-03
Under	0.888 ± 2.2e-03	0.233 ± 6.8e-03	0.128 ± 1.8e-03	**0.102 ± 1.7e-03**	0.156 ± 4.2e-03
Over	0.885 ± 1.9e-03	0.234 ± 5.8e-03	0.125 ± 1.6e-03	0.112 ± 1.9e-03	**0.148 ± 4.1e-03**

3.6 Discussion

We discussed various types of unfairness that can occur in collaborative filtering. We demonstrate that these forms of unfairness can occur even when the observed rating data is correct, in the sense that it accurately reflects the preferences of the users. We identify two forms of data bias that can lead to such unfairness. We then demonstrate that augmenting matrix-factorization objectives with these unfairness metrics as penalty functions enables a learning algorithm to minimize each of them. Our experiments on synthetic and real data show that minimization of these forms of unfairness is possible with no significant increase in reconstruction error. However, no single objective was the best for all unfairness metrics, so it remains necessary for practitioners to consider precisely which form of fairness is most important in their application and optimize that specific objective.

Future Work While our work in this chapter focused on improving fairness among users so that the model treats different groups of users fairly, we did not address fair treatment of different item groups. The model could be biased toward certain items, e.g., performing

better at prediction for some items than others in terms of accuracy or over- and underestimation. Achieving fairness for both users and items may be important when considering that the items may also suffer from discrimination or bias, for example, when courses are taught by instructors with different demographics.

Our experiments demonstrate that minimizing empirical unfairness generalizes to unseen data. However, it's worth noting that this generalization is dependent on data density. When ratings are extremely sparse, per-item unfairness estimations are likely to be inaccurate because of limited sampling, thus the empirical fairness does not always generalize well to held-out predictions. We are investigating methods that are more robust to data sparsity in future work.

Finally, our fairness metrics assume that users rate items according to their true preferences. This assumption is likely to be violated in real data, since ratings can also be influenced by various environmental factors. E.g., in education, a student's rating for a course also depends on whether the course has an inclusive and welcoming learning environment. However, addressing this type of bias may require additional information or external interventions beyond the provided rating data.

Chapter 4

Personalized Regularization Learning

As discussed in the previous chapter, matrix factorization is a canonical method for modeling user preferences for items. Regularization of matrix factorization models often uses a single hyperparameter tuned globally based on metrics evaluated on all data. However, due to the differences in the structure of per-user data, a globally optimal value may not be locally optimal for each individual user, leading to an unfair disparity in performance. Therefore, we propose to tune individual regularization parameters for each user. Our approach, *personalized regularization learning* (PRL), solves a secondary learning problem of finding the per-user regularization parameters by back-propagating through alternating least squares. Experiments on a benchmark dataset with different user group splits show that PRL outperforms existing methods in improving model performance for disadvantaged groups. We also analyze the learned parameters, finding insights into the effect of regularization on subpopulations with varying properties.

4.1 Introduction

Matrix factorization is an important and widely adapted collaborative filtering technique for training recommender systems to predict ratings. However, MF has been found to be easily influenced by data biases and becomes unfair [99, 59]. For example, demographic groups for whom training data is less frequently available can suffer less accurate predictions of their preferences [99]. This phenomenon is a form of *error-based unfairness* where users may receive lower quality service because of a demographic attribute that ideally should not affect their experience. Even worse, the group of users who receive less accurate recommendations are more likely to abandon the service, leading to an even more biased environment and more unfair models in the future [74].

Collecting more and better quality data for the disadvantaged groups will help a model better learn these users' preferences. However, this approach is usually expensive or even

infeasible. For example, a recommender service provider cannot request users to change how they interact with the service. Therefore, the more important questions is, can we handle these biases more appropriately to make better use of the available data, and build a model with improved accuracy for the ill-served users?

In this work, we first consider different types of data biases, which all refer to a certain form of divergence in the structure of per-user or per-group data. We verify on synthetic datasets that these biases can lead to one subgroup experiencing higher error than the others. We then consider the connection between prediction error and the role of regularization. If we acknowledge the difference in per-user data, then instead of tuning a global hyperparameter, a matrix factorization could benefit from personalized regularization, which better accommodates each individual user. This strategy not only directly addresses the cause of error disparity, but also provides more interpretability compared to directly manipulate the latent features since regularization is a comparatively well-understood concept.

Since personalized regularization drastically increases the number of hyperparameters, commonly used hyperparameter searching procedures—such as grid search and random search—become prohibitively expensive. It is also challenging to derive the parameters from heuristics because, in joint embedding models like matrix factorization, the effect of personalized regularization parameters are not independent of each other. Therefore, we propose a learning problem, *personalized regularization learning* (PRL), to learn the optimal set of hyperparameters that minimizes a secondary objective, in our case, the error of the disadvantage groups. We consider the secondary objective as a function of the personalized regularization parameters. To enable direct back-propagation and facilitate efficient learning, we leverage the closed-form solutions of *alternating least squares* (ALS) to solve MF.

The main contributions of this work are as follows:

1. We identify the insufficiency of global regularization for matrix factorization in dealing with complexity or sparsity imbalance across users, and conduct validation on synthetic data with explicitly injected biases;

2. We propose *personalized regularization learning* (PRL), an interpretable algorithm for learning personalized regularization by back-propagating through the closed-form computation of ALS;

3. We demonstrate the effectiveness of the proposed approach with experiments on a benchmark dataset with different user group splits, comparing against three baseline models.

4.2 Problem Definition

Given a dataset $\mathcal{D}$ that contains ratings by M users on N items. which can be represented as an $M \times N$ sparse matrix R where each observed entry r_{ui} represents the rating user u gives to item i. Suppose each user is associated with a set of properties S, based on one or more such properties $s \in S$, we can split users into a set of subgroups G.

A rating prediction model predicts the missing values in the sparse rating matrix. We randomly split all observed ratings into R^{Train} and R^{Test}. We train a model on R^{Train}, and use root mean squared error (RMSE) to measure prediction error on R^{Test} as

$$RMSE = \sqrt{\frac{1}{|R^{\text{Test}}|} \sum_{r_{ui} \in R^{\text{Test}}} (r_{ui} - \hat{r}_{ui})^2)} \tag{4.1}$$

where $\hat{r}_{ui}$ is the predicted value of r_{ui}. With a matrix factorization model, users and items are projected as matrices $P \in \mathbb{R}^{M \times d}$ and $Q \in \mathbb{R}^{N \times d}$. The uth row of P, denoted as p_u, is the latent feature of user u; the ith row of Q, denoted as q_i, is the latent feature of item i. The ratings are predicted as $\hat{r}_{ui} = p_u q_i^{\intercal}$.

Alternating Least Squares Alternating least squares (ALS) solves the above optimization task by computing the parameters alternately. At each iteration, ALS first holds Q and b_i fixed and then computes P and b_u via a closed-form solution for the minimization

$$\begin{bmatrix} b_u \\ p_u \end{bmatrix} \leftarrow \left(\sum_{i:(u,i) \in \mathcal{D}} \tilde{q_i} \tilde{q_i}^T + \lambda^* I_d \right)^{-1} \sum_{i:(u,i) \in \mathcal{D}} (r_{ui} - b_i) \tilde{q_i}, \tag{4.2}$$

where $\tilde{q_i} = [1, q_i]$ and I_d is an identity matrix of rank d. Then ALS holds P and b_u fixed and computes Q and b_i similarly as

$$\begin{bmatrix} b_i \\ q_i \end{bmatrix} \leftarrow \left(\sum_{u:(u,i) \in \mathcal{D}} \tilde{p_u} \tilde{p_u}^T + \lambda^* I_d \right)^{-1} \sum_{u:(u,i) \in \mathcal{D}} (r_{ui} - b_u) \tilde{p_u}, \tag{4.3}$$

where $\tilde{p_u} = [1, p_u]$.

Problem Formulation Given a user subgroup of concern $\hat{g} \in G$, which has higher prediction error and is considered to be the disadvantage population. The goal is to find a model that reduces error for this subgroup. The error of a subgroup $\hat{g} \in G$ is denoted and measured as

$$RMSE_{\hat{g}} = \sqrt{\frac{1}{|R_{\hat{g}}^{\text{Test}}|} \sum_{r_{ui} \in R_{\hat{g}}^{\text{Test}}} (r_{ui} - \hat{r}_{ui})^2)} \tag{4.4}$$

where $R_{\hat{g}}^{\text{Train}}$ and $R_{\hat{g}}^{\text{Test}}$ denote the training and test data of $\hat{g}$ respectively.

4.3 Data Biases and Regularization

In this section, we discuss four types of data biases that contribute to higher prediction error in disadvantaged subgroups, and empirically show the consequences of these biases with synthetic datasets. We also discuss how these data biases are related to regularization and imply the need for personalized regularization.

4.3.1 Data Biases

We first consider a group-level bias called *population bias*, which refers to the discrepancy among the size of subgroups. The subgroups with smaller populations are more likely to be compromised in modeling, especially when the data of these minority groups have a very different structure from the other groups. We also consider three individual-level biases. The first one is *sparsity bias*, which refers to the difference in per-user data sparsity. A model with a particular complexity requires a corresponding amount of data to overcome the curse of dimensionality, which creates a disadvantage for users who are new or less active. The second one is *rank bias*, it refers to the situation that some users' preferences are more complicated than others. Therefore, a higher-dimensional model is required to capture their preferences. The third one is *noise bias*, which suggests different levels of data quality and the situation where some users' data is more noisy than the others. We believe these four types of data biases lead to increased prediction error for the subgroups that are being biased against.

4.3.2 Validation

We validate our heuristics on the effect of data bias on synthetic datasets where we explicitly inject two types of data bias among users. We create the synthetic data by first generating a twenty-dimensional user and item feature matrices for 100 users and 600 items. Then we compute the rating matrix as their dot product. To make the datasets more realistic, we also add Gaussian noise to these ratings and clamp them within the range of 1.0 to 5.0.

We assign users to two subgroups A and B. To inject data biases, without loss of generosity, we choose group B to be the disadvantaged group and the users from group B to be the disadvantaged users. For population bias, we lower the population of group B to be the minority group; for rank bias, we force some columns of the latent features of users from group A to be zero, so that group B has higher dimension than group A; for sparsity bias, we mask more ratings from group B than group A; for noise bias, we add a higher level of Gaussian noise to the rating of group B. Specifically, to create the biased settings, we set $|B| = 30, |A| = 70$; $d_A = 5, d_B = 20$; $5\times$ amount of ratings are observed from group A than group B; $2\times$ amount of noise is added to the ratings of group B[1].

[1]Note that to avoid the effect of irrelevant factors, we normalize the ratings so that the ratings of group

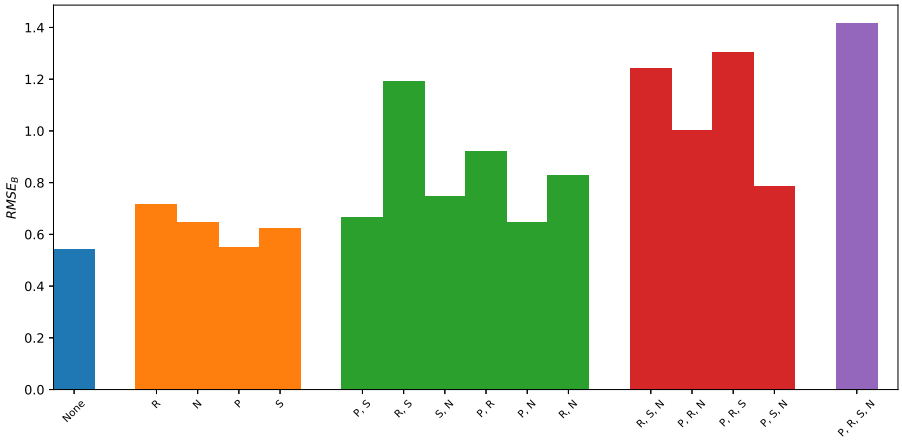

Figure 4.1: The measured $RMSE_B$ of models trained on synthetic datasets with different data biases injected. Here we use R, N, P, S to indicate rank bias, noise bias, population bias, and sparsity bias respectively. The models are grouped based on the number of injected data biases and are presented in different colors.

We train matrix factorization models on these datasets and measure $RMSE_B$, the error of group B. The results are shown in Figure 4.1. The blue bar represents a bias-free dataset; the orange bars represent datasets with each individual type of bias; the green, red, and purple bars represent datasets with 2, 3, and 4 types of biases respectively. R, N, P, S are short for rank bias, noise bias, population bias, and sparsity bias respectively. First, by comparing the fair setting (the blue bar) and the settings with each individual biases (the orange bars), we observe that, except population bias, each of the discussed biases alone directly leads to higher $RMSE_B$. Population bias is the exception because when the other three biases are not present, the two subpopulations have exactly the same data structure. Second, in general, $RMSE_B$ increases as more types of biases are injected, suggesting a compound impact of data biases. Third, the effect of these data biases are not independent but can enhance each other. For example, we observe a noticeable increase in $RMSE_B$ from setting "R, S, N" to "R, S, N, P" when population bias is added, which by itself does not have the same effect.

4.3.3 Relation to Regularization

The data biases discussed above are directly related to data properties that, if not carefully handled when building a model, will lead to overfitting or underfitting. Overfitting or underfitting are two forms of mistakes that a model could make to mishandle training data and increase error. Overfitting happens when a model attends to too much detail and noise,

A and B follow the same distribution. We also keep the overall level of sparsity and noise unchanged by rescaling the configuration within each subgroup.

and is more likely to occur when data is insufficient due to increase variance; underfitting, on the other hand, happens when a model oversimplifies and fail to capture the underlying structure of the data.

An important component for balancing underfitting and overfitting in machine learning models is regularization. The strength of regularization needs to be tuned to best fit the learning task and the data. Since we discussed that data properties such as quality, sparsity, and complexity may not be universal across all users, tuning a global regularization parameter λ^*, as is done in standard matrix factorization models, becomes insufficient to accommodate the important differences in per-user data, leading to poor model performance on certain user subgroups.

4.4 Personalized Regularization Learning

We have discussed that a globally tuned regularization parameter λ^* neglects the differences in per-user data. Therefore, we believe a model could benefit from a set of personalized regularization parameters. With this expanded set of hyperparameters, the objective function of training a matrix factorization model is modified by replacing the globally tuned regularization hyperparameter λ^* with user-personalized regularization parameters $\Lambda = \{\lambda_u\}_{u=1}^{M}$. The user and item latent features are learned as

$$P^*, Q^* = \min_{P,Q} \sum_{r_{ui} \in R^{\text{Train}}} (r_{ui} - \hat{r}_{ui})^2 + \frac{1}{2} \left(\sum_{u=1}^{M} \boldsymbol{\lambda_u} (p_u p_u^\intercal) + \sum_{i=1}^{N} \lambda^* (q_i q_i^\intercal) \right) \tag{4.5}$$

Since personalized regularization grows the space of hyperparameters from $\mathbb{R}$ to $\mathbb{R}^M$, traditional tuning procedures such as grid search become insufficient in such a high-dimensional space. Also, in joint embedding models like matrix factorization, the personalized regularization parameters are not independent of each other, therefore, it is challenging to derive the parameters from heuristics.

4.4.1 Personalized Regularization Learning

To efficiently search for the optimal personalized hyperparameters, we propose *personalized regularization learning* (PRL), which poses the hyperparameter search problem as a secondary learning task. We denote the primary learning problem in Equation (4.5) as L, which returns the learned P, Q for a given hyperparameter set Λ

$$P, Q = L(\Lambda) \tag{4.6}$$

We then make predictions using the learned latent features P and Q through a predictor function H,

$$\hat{R} = H(P, Q) \tag{4.7}$$

and evaluate a secondary objective, which in our running example, is the subgroup error $RMSE_{\hat{g}}$ through function E,

$$RMSE_{\hat{g}} = E(\hat{R}) \tag{4.8}$$

Combining equation Equations (4.7), (4.8) and (5.5), we get $RMSE_{\hat{g}} = E(H(L(\Lambda))) = F(\Lambda)$ where $F = E(H(L))$. The secondary learning problem is formulated as

$$\Lambda^* = \min_{\Lambda \in \mathbb{R}^M} F(\Lambda) \tag{4.9}$$

F is a differentiable function if E, H, and L are all differentiable. Then we can directly backpropagate through F to compute gradients of the secondary objective with respect to Λ.

4.4.2 Leveraging ALS

Solving the factorization problem L involves minimizing a non-convex regularized squared reconstruction error. One approach for optimizing this objective is through gradient descent [38], which has been made especially convenient with the advance of automatic differentiation tools [39]. However, the gradient of hyperparameters are usually unavailable [81]. Therefore, we instead use alternating least squares (ALS) [40] to solve L, which alternates between optimizing P and Q by iteratively applying a closed-form solution

$$\begin{aligned} P_u &\leftarrow \left(\sum_{i:(u,i)\in\mathcal{D}} \tilde{q_i}\tilde{q_i}^T + \lambda^* I_d \right)^{-1} \sum_{i:(u,i)\in\mathcal{D}} r_{ui}\tilde{q_i} \\ Q_i &\leftarrow \left(\sum_{u:(u,i)\in\mathcal{D}} \tilde{p_u}\tilde{p_u}^T + \lambda^* I_d \right)^{-1} \sum_{u:(u,i)\in\mathcal{D}} (r_{ui} - b_u)\tilde{p_u} \end{aligned} \tag{4.10}$$

The closed-form updates are differentiable, so we can conveniently back-propagate through F to compute gradients of the secondary objective with respect to Λ and learn Λ with a standard gradient-based optimizer. Since the time complexity of computing partial derivative is the same as forward passing [100], the time it takes to back-propagate to Λ is the same as forward ALS, thus the time complexity of PRL is $\mathcal{O}(T)$, where T is the number of epochs ALS takes to converge. This is on par with the state-of-the-art hyperparameter optimization techniques [81].

4.4.3 Data Split

During learning, we must use different datasets for training the MF model and measuring subpopulation error. This is because our goal is to decrease generalization error, which needs to be evaluated on data unseen by the training algorithm. If we measure subpopulation error on the same data that the MF model is trained on, we may simply incentivize the learning optimization to overfit the data as much as possible.

Therefore, after we split data $\mathcal{R}$ into $\mathcal{R}^{\text{Train}}$ and $\mathcal{R}^{\text{Test}}$, we further split $\mathcal{R}^{\text{Train}}$ into $\mathcal{R}^{\text{Train-Primary}}$ and $\mathcal{R}^{\text{Train-Secondary}}$. In each PRL iteration, we train a matrix factorization model on $\mathcal{R}^{\text{Train-Primary}}$ and compute $RMSE_{\hat{g}}$ on $\mathcal{R}^{\text{Train-Secondary}}$, which is used to update Λ. After we obtained Λ^*, we apply it to train a final matrix factorization model on the full training set $\mathcal{R}^{\text{Train}}$. We then evaluate $RMSE_{\hat{g}}$ on $\mathcal{R}^{\text{Test}}$. The full algorithm is listed as Algorithm 1. In practice, we recommend creating multiple primary-secondary splits of $\mathcal{R}^{\text{Train}}$ so that the learned Λ^* is not overfitted to one particular split.

Algorithm 1: Personalized Regularization Learning

Given dataset $\mathcal{R}$, global optimal lambda λ^*, MF model L, error metric E, disadvantaged subpopulation $\hat{g}$. Split $\mathcal{R}$ into $\mathcal{R}^{\text{Train}}$ and $\mathcal{R}^{\text{Test}}$, further split $\mathcal{R}^{\text{Train}}$ into $\mathcal{R}^{\text{Train-Primary}}$ and $\mathcal{R}^{\text{Train-Secondary}}$.

Initialize $\Lambda \leftarrow \{\lambda_i = \lambda^*\}_{i=1}^{N}$, randomly initialize P and Q

while *not converged* **do**

- $P^*, Q^* \xleftarrow{\mathcal{R}^{\text{Train-Primary}}} L(\Lambda)$
- $\hat{R} \xleftarrow{\mathcal{R}^{\text{Train-Secondary}}} H(P^*, Q^*)$
- $RMSE_{\hat{g}} \xleftarrow{\mathcal{R}^{\text{Train-Secondary}}} E(\hat{R}_{\hat{g}})$
- compute gradient $\nabla_\Lambda RMSE_{\hat{g}}$ through backpropagation
- update Λ with $\nabla_\Lambda RMSE_{\hat{g}}$

end

Re-initialize P and Q

$P^*, Q^* \xleftarrow{\mathcal{R}^{\text{Train}}} L(\Lambda^*)$

4.4.4 Interpretability

A key advantage of PRL is that it provides interpretable feedback in the magnitude of the learned per-user regularization parameters. Compared to regularization-based methods that directly manipulate user and item latent representations, PRL's learned parameters indicate the level of regularization, which is comparatively well-understood and can help us understand how the model is improved. We can interpret them by comparing their values against the globally tuned value. If a user's parameter increases, it suggests that this user

would have been overfitted. Conversely, if a user receives lower regularization from PRL, they were prone to underfit and needed a more complex model.

4.5 Experiments

4.5.1 Datasets

The choice of public real datasets that provide user demographic information is very limited. We use the benchmark MovieLens 100k dataset [98], which contains 100,000 ratings from 1,000 users on 1,700 movies, and conveniently provides multiple user demographic features. Specifically, we consider demographic information such as gender, age, zip code; we also consider user degree–the number of ratings each user has, and user error–the error of each user with a vanilla matrix factorization model. For gender, we split users by category and create two subgroups (female and male users); for zipcode, we split by the first digit of zip code and create 10 subgroups, representing users from different regions in the US; for age, degree, and error, we split by percentile and each split creates 10 equal size subgroups.

We randomly sample 10% of data as holdout set for testing and use the rest as training set. Then we do 10-fold cross-validation on the training set to select the best global regularization weight λ and the rank d. The optimal combination we found is $d^* = 30$ and $\lambda^* = 1.0$. We train a standard matrix factorization model and measure the subgroup errors under all user splits. For each split, we pick the subgroup with the highest error as the disadvantaged group, denoted as $\hat{g}$ and seek to reduce $RMSE_{\hat{g}}$. The disadvantaged subgroups are listed in the second row of Table 4.1.

4.5.2 Baselines

Focused learning (FL) FL [22] assigns users to only two subgroups, a focused set, and an unfocused set. The two sets of users are regularized differently to optimize the model performance on the focused set of users. The optimal hyperparameter pair is searched via grid search.

Differentiated regularization (DR) DR [101] is motivated to alleviate the cold-start problem and regularize every user differently. The regularization parameters are computed from three functions (one linear and two logarithmic) of user degree. We denote the three formulas as DR-linear, DR-Log-1, and DR-Log-2.

Unfairness-regularized matrix factorization (URMF) URMF [99] is designed to optimize a secondary fairness in matrix factorization models. The strategy is to add the opti-

mized secondary objective as a penalty term to the standard matrix factorization objective, weighted by a weight parameter. URMF directly manipulates the fitted latent embeddings instead of through regularization.

4.5.3 Specifications and Results

For DR, we directly apply the three formula (Equation 5 in [101]) to compute personalized regularization parameters. For FL, we follow the same procedure as proposed by the authors and try a range of regularization values on the focused and unfocused set, $\{0.001, 0.01, 0.1, 1, 1, 10, 20, 30, 50, 100\}$, which gives 100 combinations. For URMF, we try 10 different unfairness pernalty weights $\{1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 5, 10, 20\}$. For FL and UR, we identify the optimal setting or weight via cross-validation, then apply the same setting or weight to train a final model on the full training set. For all trained models, we measure $RMSE_{\hat{g}}$ on the holdout test set.

We show the results of all compared models on different user splits in Table 4.1. We first compare the performance of PRL against the standard matrix factorization model. We observed that PRL successfully reduces $RMSE_{\hat{g}}$ on all user splits, and achieves more than 10% improvement on subgroups split by Zip Code and Error. We also observe that PRL outperforms all baseline models by a convincing margin. Focused learning is the second-best method, this further suggests that fitting different regularization is effective in optimizing subpopulation error. We believe PRL wins over Focused learning due to the expanded set of hyperparameters and smart search through optimization. We observe a big fluctuation in the performance of URMF, it is possibly because URMF still can easily overfit to the training data since it measures both primary and secondary objectives on the same data. Lastly, DR performs the worst. It rarely reduces $RMSE_{\hat{g}}$ and even when it does, the improvements are trivial. This pattern aligns with the results and conclusion in the original paper that DR sometimes makes things worse and especially so on the MovieLens dataset. The poor performance of DR suggests it is challenging to find a one-fits-all heuristic for setting the personalized regularizations.

We next examine the personalized regularization parameter fitted through PRL to understand how it reduces $RMSE_{\hat{g}}$. We compute the mean and standard deviation of the regularization parameters in each subgroup throughout PRL optimization. We show the plot in gender subgroups as an example in Figure 4.2. In this case, female users is the disadvantaged subgroup. We observe that female users, on average, have been assigned lower regularization, and male users' regularization has been increased. This provides an interesting insight that the complexity of the originally tuned global model was lower than what is need for the disadvantaged group. PRL allows these users to enjoy a more complex model that better captures their preferences. We also noticed an increased variance in the regularization parameter, and surprisingly even more so in the advantaged group. As we discussed in Section 4.4, the personalized regularization parameters are not independent of each other

Table 4.1: Comparison of all model performance in reducing $RMSE_{\hat{g}}$ and the percentage of change compared to MF. Bold values are the most significant improvement in each column.

User Split	Gender	Age	Zip Code	Degree	Error
$\hat{g}$	F	52-59	0	0%-10%	90%-100%
MF	1.029	1.045	1.062	1.118	1.612
PRL	**0.967(-6.0%)**	**0.983(-5.9%)**	**0.952(-10.4%)**	**1.043(-6.7%)**	**1.367(-15.2%)**
FL	0.998(-3.0%)	1.001(-4.2%)	0.989(-6.9%)	1.094(-2.1%)	1.479(-8.2%)
URMF	1.013(-1.6%)	1.089(+4.2%)	0.956(-10.0%)	1.102(-1.4%)	1.579(-2.0%)
DR-Linear	1.041(+1.2%)	1.127(+7.8%)	1.047(-1.4%)	1.114(-0.4%)	1.601(-0.7%)
DR-Log-1	1.067(+3.7%)	1.113(+6.5%)	1.082(+1.9%)	1.109(-0.8%)	1.641(+1.7%)
DR-Log-2	1.059(+2.9%)	1.114(+6.6%)	1.098(+3.9%)	1.112(-0.5%)	1.632(+1.2%)

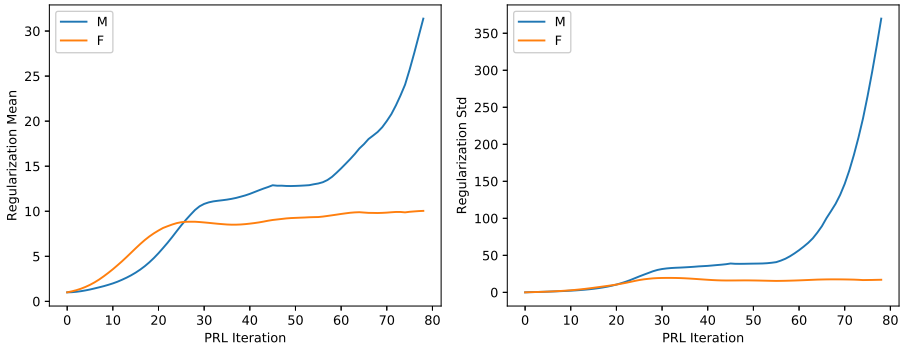

Figure 4.2: The curve of mean and standard deviation of personalized regularization values in different gender groups during PRL. On average, PRL assigns lower regularization to female users (the disadvantaged group). The regularization parameters of female users also have lower variance.

in joint embedding models, therefore, the regularization of all users are shifted even though the objective is to only optimize the prediction error of the disadvantaged group. Further, we found the same direction of change in the pair of regularization parameters identified via focused learning ($\lambda_F = 10$ and $\lambda_M = 30$). This suggests an alignment between the two methods in reducing high subpopulation error through adjusted regularization.

4.6 Discussion

In this work, we address the problem of error disparity in matrix factorization models. We discuss four types of biases that contribute to higher subpopulation error and validate their effect on synthetic datasets. We presented *personalized regularization learning* (PRL), a method that learns to regularize users differently to improve prediction performance for

disadvantaged subgroups of users. PRL solves a secondary learning problem to minimize validation unfairness by back-propagating through alternating least squares. In experiments, PRL outperforms existing methods for reducing error disparity in recommendations. Moreover, the learned per-user regularization parameters are interpretable and provide insight into how fairness is improved. For future work, we are interested in investigating the effectiveness of PRL in other variants of matrix factorization, such as SVD++, factorization machine. We are also interested in further exploring the learned regularization parameters to uncover richer group structures.

Chapter 5

Long-term Equality

In this chapter, we shift our focus from studying unfairness in recommendation decisions to examining the fundamental disparity in the underlying population. We conduct a theoretical analysis of the long-term dynamic of inequality in the fitting between users and items in recommender systems. The results of our study are valuable for anticipating, avoid exacerbating, and reducing inequality in the future. This work also demonstrates that understanding the inequality in recommender systems requires a close examination of the user dynamics as well as the relationship between different item categories.

5.1 Introduction

Previously, we discussed that disparity in the training data (e.g., different true average rating and sparsity) can cause a rating predictor to be biased against disadvantaged subgroups, and we proposed mitigation methods to make model outputs fairer. However, the disparity in training data is often a reflection of the inequality in the real world. Specifically, in the context of recommender systems, the inequality is considered to be the disparity in the level of fit between users and items. The fit includes compatibility in all aspects, such as interest fit, qualification fit and culture fit [102, 103]. For example, a current inequality is that female students may have a lower level of fit with STEM courses, with fewer of them being interested, competitive, or feeling welcomed in these courses [104].

There are two main reasons why inequalities are undesirable. First, if the inequalities in the real world persist, the training data, which is drawn from the real world, will also remain biased. For example, if group A has a lower ratio of users who like an item than group B, then we may observe a lower average rating from group A for this item because users are likely to give higher ratings when they like an item. We may also observe a higher sparsity in ratings for group A since users are less likely to try an item that they do not like and subsequently rate it [105]. Second, when those inequalities exist, a fairness intervention will

always incur a loss in utility since the output of fair models is different from the biased ground truth. In the same example where we have user group A and B, we may build a classifier to make the binary recommendation decisions for all users (to recommend the item or not). Suppose the classifier is subject to the demographic parity (DP) constraint so that the positive rate in classification is required to the same for A and B. Since group A indeed has a lower ratio of users who like the item, then in group A, the DP-fair classifier must have recommended the item to more users than it should have; while in group B, the classifier must have missed some users who it should have recommended to. Therefore, in some cases, it may be desired that such inequality be reduced.

In this work, we conduct a theoretical analysis on the long-term dynamics of inequality across user groups. These dynamics are not yet understood despite active efforts to promote algorithmic fairness. We call the level of fit between a user group and an item category the *fit rate*, measured as the ratio of users from the user group being a fit with the item category. We model the fit rates as changeable [106, 107, 108]. We follow psychological research that has shown human preferences can be changed by the environment. One such phenomenon is the mere exposure effect [91, 109, 110], which refers to the increase in interest and preference after repeated exposure. Since recommendations change user exposure to different items, we focus on the dynamics of user-item fit directly driven by recommendation decisions. Note that we assume those changeable interest and preference are not innate [111]. Thus when we seek to reduce inequality in user preference, we only correct the skewed preferences that are misshaped due to biased environment.

Characterizing the dynamic of fit between users and items is not trivial. First, whether a user fits with an item category can be changed depending on whether this very item category is recommended to him or her. For example, a female student may become a fit with a computer science course after being recommended the course, because of increased exposure or accumulated knowledge after accepting the recommendation and taking the course [112, 113, 114, 115]. Also, on a group level, female students receive encouragement from more role models as more female students participate in STEM [116, 117, 118]. We call the preference transition in an item category due to recommendations in this very item category *self impact*. Second, different item categories may have shared or dissimilar properties [119, 120], so recommendation decisions in one item category may also affect fit rates in another item category. For example, being exposed to math courses may also make a student more interested in computer science since these two disciplines are closely related [121, 122]. We refer to this transition as *cross impact*. Moreover, the changes in the fit happen gradually through repeated interactions between recommender systems and users, therefore, the study requires zooming out from one-step analysis and examine long-term dynamics.

In summary, we conduct a theoretical analysis of the long-term dynamics of inequality in the fit between users and items in recommender systems. Our main contributions are:

1. We pose the task of recommendation as solving a set of binary classification problems through threshold policies, and we mathematically formulate self impact, cross impact,

and the dynamics of fit rates.

2. We prove that equilibrium always exists in a system with the formulated dynamics. We also provide sufficient conditions that guarantee the equilibrium to be unique.

3. We show how recommendations in different item categories shape the equality in the fit rates in each other at equilibrium. We further discuss how fit rates in one item category change due to fairness intervention (demographic parity and equal opportunity) in its co-existing item categories.

4. We validate our theoretical analysis through simulation on synthetic users and items.

5.2 Background

Zhang et al. [123] studied the long-term dynamics of qualification rates in a loan approval classification problem. The classifier adopts a threshold policy. In our setting, the recommendations decisions within each item category are made by solving such classification task and also using threshold policies. This means that if we do not consider the correlation between item categories, we can directly apply the analysis provided by Zhang et al. [123] to study the long-term dynamics of fit rates in each item category. Therefore, we first review the classification problem and threshold policy in this section.

Classification Task Suppose there are two user groups $\mathcal{G}_A$ and $\mathcal{G}_B$, split by a sensitive attribute $S = s \in \{a, b\}$ (e.g., gender), each group has a fraction $p_s := P(S = s)$ of the population. At time t, an individual with attribute s has a feature $X_t^s = x \in \mathbb{R}$ determined by a hidden state $Y_t^s = y \in \{0, 1\}$. Specifically, the features are generated by distribution $G_y^s(x) := P(X_t|Y_t = y, S = s)$. Let $\alpha_t^s := P(Y_t = 1|S = s)$. The convex combination $P(X_t = x|S = s) = \alpha_t^s G_1^s(x) + (1 - \alpha_t^s)G_0^s(x)$ is the composite distribution of Group $\mathcal{G}_s$ at time t. For individuals from group $\mathbb{G}_s$, a classifier makes decisions $D^s = d \in \{0, 1\}$ using a policy π_t where $\pi_t^s(x) := P(D_t = 1|X_t = x, S = s)$ to maximize an total utility $R_t(D_t, Y_t)$, possibly subject to certain constraints.

Threshold policy An unconstrained policy π_t at time t maximizes the instantaneous expected utility $U(D_t, Y_t) = E[R_t(D_t, Y_t)]$, where

$$R_t(D_t, Y_t) := \begin{cases} u_+, & \text{if } Y_t = 1 \text{ and } D_t = 1. \\ -u_-, & \text{if } Y_t = 0 \text{ and } D_t = 1. \\ 0, & \text{if } D_t = 0 \end{cases} \tag{5.1}$$

A fair policy maximizes the total utility defined in Equation (5.1) subject to a fairness constraint $\mathcal{C}$. As shown in [123], a common strategy is to find (π_t^a, π_t^b) that solves the

following constrained optimization

$$\begin{aligned} \max_{\pi_t^a, \pi_t^b} \ & \mathcal{U}(D_t, Y_t) = p_a \mathbb{E}[R_t(D_t, Y_t)|S = a] + p_b \mathbb{E}[R_t(D_t, Y_t)|S = b] \\ s.t. \ & \mathbb{E}_{X_t \sim P_{\mathcal{C}}^a}[\pi^a(X_t)] = \mathbb{E}_{X_t \sim P_{\mathcal{C}}^b}[\pi^a(X_t)] \end{aligned} \tag{5.2}$$

for demographic parity, $P_{DP}^s(x) = (1 - \alpha_t^s)G_0^s(x) + \alpha_t^s G_1^s(x)$; for equality of opportunity, $P_{EqOpt}^s = G_1^s(x)$. Under two mild assumptions, the optimal (fair) policies are in the form of threshold policies, i.e., $\pi_t^s(x) = \mathbb{1}(x \geq \theta_t^s)$. First, $\frac{G_1^s(x)}{G_0^s(x)}$ is strictly increasing in $\mathbb{R}$. Second, $P_{\mathcal{C}^s(x)}$ is continuous and $\frac{P(X=x|S=s)}{P_{\mathcal{C}^s(x)}}$ is non-decreasing.

According to *Lemma 1* by Zhang et al. [123], the optimal unconstrained threshold pair $(\theta_{UN}^{a^*}, \theta_{UN}^{b^*})$ are computed by solving $\gamma^a(\theta_{UN}^{a^*}) = \gamma^b(\theta_{UN}^{b^*}) = \frac{u_-}{u_+ + u_-}$, where

$$\gamma_t^s = \mathcal{P}(Y_t = 1|X_t - x, S = s) = \frac{1}{\frac{G_0^s(x)}{G_1^s(x)}(\frac{1}{\alpha^s} - 1) + 1}.$$

The optimal fair threshold pair under EqOpt fairness $(\theta_{EqOpt}^{a^*}, \theta_{EqOpt}^{b^*})$ is to solve $\frac{p_a}{\gamma^a(\theta_{DP}^{a^*})} + \frac{p_b}{\gamma^b(\theta_{DP}^{b^*})} = \frac{u_-}{u_+ + u_-}$. The optimal fair threshold pair under DP fairness $(\theta_{DP}^{a^*}, \theta_{DP}^{b^*})$ is to solve $p_a\gamma^a(\theta_{DP}^{a^*}) + p_b\gamma^b(\theta_{DP}^{b^*}) = \frac{u_-}{u_+ + u_-}$.

Transitions After receiving the recommendation decisions, an individual's fit state Y and fit score X may change and this is modeled by a set of transitions probabilities T^s where $T_{y,d}^s = \mathbb{P}(Y_{t+1} = 1|Y_t = y, d_t = d, S = s)$. Note that instead of modeling individuals' strategic responses [124, 125], the probabilities characterize the overall effect and encapsulate all individual or social factors that drive the transitions.

5.3 Problem Formulation

In this section, we formally describe the recommendation task. We also introduce additional components and assumptions that is necessary for analyzing the long-term dynamics of fit rates in a recommender system, compared to the classification problem studied by Zhang et al [123].

Recommendation Task We consider a recommendation system with two item categories $\mathcal{G}_J$ and $\mathcal{G}_K$ distinguished by an attribute $i \in \{j, k\}$. In each item category, the recommendation decisions are made by solving the aforementioned classification task. Specifically, at time t, for each item category i, an individual with attribute s has a fit score $X_t^{s,i} = x \in \mathbb{R}$ that represents the match between the individual and $\mathbb{G}_i$. The fit score is determined by

a hidden fit state $Y_t^{s,i} = y \in \{0,1\}$ as $G_y^{s,i}(x) := P(X_t^{s,i}|Y_t^{s,i} = y, S = s)$. The quantity $\alpha_t^{s,i} := P(Y_t^{s,i} = 1|S = s)$ is the fit rate of user group $\mathcal{G}_s$ in item category $\mathcal{G}_i$. A recommender only has access to the fit scores (e.g., output of a rating predictor) and makes recommendation decisions $D^{s,i} = d \in \{0,1\}$ (to recommend or not) using threshold policies $\pi_t^{s,i}$ at time t, i.e., $D^{s,i} = \pi_t^{s,i}(x) = \mathbb{1}(x \geq \theta_t^{s,i})$.

Distribution of Fit States In addition to $\alpha^{s,i}$, we also define the joint probability $P(Y^j = y^j, Y^k = y^k|S = s)$ as $\beta^s_{y^j,y^k}$. Then $\forall s \in \{a,b\}$, for any given $\beta^s_{1,1}$, we have

$$\beta^s_{1,0} = P(Y^j = 1, Y^k = 0|S = s) = P(Y^j = 1|S = s) - P(Y^j = 1, Y^k = 1|S = s) = \alpha^{s,j} - \beta^s_{1,1},$$

$$\beta^s_{0,1} = P(Y^j = 0, Y^k = 1|S = s) = P(Y^k = 1|S = s) - P(Y^j = 1, Y^k = 1|S = s) = \alpha^{s,k} - \beta^s_{1,1},$$

$$\beta^s_{0,0} = P(Y^j = 1, Y^k = 1|S = s) = 1 - \beta^s_{1,1} - \beta^s_{1,0} - \beta^s_{0,1} = 1 - \alpha^{s,j} - \alpha^{s,k} + \beta^s_{1,1}.$$

Let $-i = \{j,k\}\backslash i$. We also denote the conditional probability $\mathbb{P}(Y^{-i} = 1|Y^i = y^i, S = s)$ as $\lambda^{s,i}_{y^i}$, and it is computed as $\lambda^{s,i}_{y^i} = \frac{P(Y^i = y^i, Y^{-i} = 1|S = s)}{P(Y^i = y^i|S = s)}$. Therefore, $\lambda^{s,i}_1 = \frac{\beta^s_{1,1}}{\alpha^{s,i}}$, $\lambda^{s,i}_0 = \frac{\alpha^{s,-i} - \beta^s_{1,1}}{1-\alpha^{s,i}}$. For simplicity, we denote $\beta^s_{1,1}$ as β^s and call it the *joint fit rate*.

Transition of preference states. We expand the transitions from T^s to $T^{s,i,i'}$, where $T^{s,i,i'}_{y^i,d^{i'}} := P(Y^i_{t+1}|Y^i_t = y, D^{i'}_t = d^{i'}, S = s), i, i' \in \{j,k\}$. When $i' = i$, $T^{s,i,i'}$ reduces to T^s and refers to the probabilities a user becoming a fit with an item category after being recommended or not recommended this very item category. We call these transitions *self impact*. When $i' = -i$, $T^{s,i,i'}$ refers to the probabilities a user becoming a fit with an item category after being recommended or not recommended another item category. We call these transitions *cross impact*.

Self impact and cross impact are aggregated into compound transition probabilities $T^{s,i}$ where $T^{s,i}_{y^i,d^i,d^{-i}} := P(Y^i_{t+1}|Y^i_t = y, D^i_t = d^i, D^{-i}_t = d^{-i}, S = s)$ through a function F. $T^{s,i}_{y^i,d^i,d^{-i}} \in [0,1], \forall s \in \{a,b\}, i \in \{j,k\}, y^i, d^i, d^{-i} \in \{0,1\}$. One example of F is $T^{s,i}_{y^i,d^i,d^{-i}} = F(T^{s,i,i}_{y^i,d^i}, T^{s,i,-i}_{y^i,d^{-i}}) = \frac{1}{2}\big((T^{s,i,i}_{y^i,d^i}T^{s,i,-i}_{1,d^{-i}} + (1 - T^{s,i,i}_{y^i,d^i})T^{s,i,-i}_{0,d^{-i}}) + (T^{s,i,-i}_{y^i,d^{-i}}T^{s,i,i}_{1,d^i} + (1 - T^{s,i,-i}_{y^i,d^{-i}})T^{s,i,i}_{0,d^i})\big)$, which means the two impact take place sequentially and independently.

Assumptions. Throughout the study, we make the following assumptions about the feature distributions and transition probabilities.

Assumption 1 (Strictly increasing compound impact)**.** $\forall s \in \{a,b\}, i \in \{j,k\}$, $F(T^{s,i,i}_{y^i,d^i}, T^{s,i,-i}_{y^i,d^{-i}})$ *is continuous and strictly increasing in* $T^{s,i,i}_{y^i,d^i}$ *and* $T^{s,i,-i}_{y^i,d^{-i}}$.

Assumption 1 limits the forms of the compounding function F to be continuous and strictly increasing in $T^{s,i,i}_{y^i,d^i}$ and $T^{s,i,-i}_{y^i,d^{-i}}$. This means increasing either self impact probabilities or cross impact probabilities will always increase the compound transition probabilities.

Assumption 2. $T^{s,i,i}_{0,d^i} < T^{s,i,i}_{1,d^i}$, *and* $T^{s,i,-i}_{0,d^{-i}} < T^{s,i,-i}_{1,d^{-i}}, \forall s \in \{a,b\}, i \in \{j,k\}, d^i \in \{0,1\}$.

Assumption 2 says that for any item category and after any recommendation decisions, those who are already a fit always have a higher probability to maintain being a fit in the next time step than those who were not a fit in the previous time step to become one.

Assumption 3.a. $T^{s,i,i}_{0,1} \geq T^{s,i,i}_{0,0}$ *and* $T^{s,i,i}_{1,1} \geq T^{s,i,i}_{1,0}$.

Assumption 3.b (Group-invariant distribution)**.** $G^{a,i}_y = G^{b,i}_y, \forall y \in \{0,1\}, i \in \{j,k\}$.

Assumption 3.a specifies that under self impact, receiving recommendations always increases the probability of becoming a fit, due to factors such as stimulated interest or increased qualification when the user consumes the recommended item. This further aligns with the mere exposure effect in psychology, which refers to the phenomenon that people tend to develop a preference for things because of familiarity [91]. Note that this is also Condition $1(b)$ in [123]. Assumption 3.b means the fit score distributions given the fit states are demographic invariant. However, the overall fit score distribution is demographic variant because of different fit rates. These two assumptions combined guarantee that both EqOpt and DP constraints in the recommendation policy of one item category will always reduce inequality in this item category.

Problem Statement With these distributions, transitions and assumptions, we are interested in studying how $\alpha^{s,i}_t$ change under different transition probabilities and recommendation policies.

5.4 Equilibrium Analysis

In this section, we discuss how different components of the system change after one step of recommendation. We first characterize the dynamics of fit rates. At each time step, the users that fit with an item category are (i) those who were not a fit in the previous step that become a fit, as well as (ii) those who were already a fit and remain a fit. Mathematically, $\forall s \in \{a,b\}, i \in \{j,k\}$ under policy $\pi^{s,i}_t$ and $\pi^{s,-i}_t$, the dynamic of $\alpha^{s,i}_t$ is as follows:

$$\alpha^{s,i}_{t+1} = g^{0,s,i}(\alpha^{a,j}_t, \alpha^{b,j}_t, \alpha^{a,k}_t, \alpha^{b,k}_t, \beta^s_t)(1-\alpha^{s,i}_t) + g^{1,s,i}(\alpha^{a,j}_t, \alpha^{b,j}_t, \alpha^{a,k}_t, \alpha^{b,k}_t, \beta^s_t)\alpha^{s,i}_t \tag{5.3}$$

where $g^{y,s,i}(\alpha^{a,j}_t, \alpha^{b,j}_t, \alpha^{a,k}_t, \alpha^{b,k}_t, \beta^s_{1,1}) = \mathbb{P}(y^{s,i}_{t+1} = 1 | y^{s,i}_t = y)$ and is computed as

$$\begin{aligned} g^{y,s,i}(\alpha^{a,j}_t, \alpha^{b,j}_t, \alpha^{a,k}_t, \alpha^{b,k}_t, \beta^s_t) := \; & E_{X^i_t, X^{-i}_t | Y_t = y, S = s}\big[\big(1-\pi^{s,i}_t(X^i_t)\big)\big(1-\pi^{s,-i}_t(X^{-i}_t)\big)T^{s,i}_{y^i,0,0} \\ & + (\pi^{s,i}_t(X^i_t))(1-\pi^{s,-i}_t(X^{-i}_t))T^{s,i}_{y^i,1,0} \\ & + (1-\pi^{s,i}_t(X^i_t))(\pi^{s,-i}_t(X^{-i}_t))T^{s,i}_{y^i,0,1} \\ & + \pi^{s,i}_t(X^i_t)\pi^{s,-i}_t(X^{-i}_t)T^{s,i}_{y^i,1,1}\big]. \end{aligned} \tag{5.4}$$

For simplicity, we denote $g^{y,s,i}(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, \beta_t^s)$ as $g^{y,s,i}$. With a threshold policy, i.e., $\pi_t^{s,i}(x) = \mathbb{1}(x \geq \theta_t^{s,i})$, this means

$$\begin{aligned}
g^{y,s,i} :=\ & T^{s,i}_{y^i,0,0} \int_{-\infty}^{\theta_t^{s,i}} G_y^i(x^i)dx \Big((1-\lambda^{s,i}_{y^i,t}) \int_{-\infty}^{\theta_t^{s,-i}} G_0^{-i}(x^{-i})dx + \lambda^{s,i}_{y^i,t} \int_{-\infty}^{\theta_t^{s,-i}} G_1^{-i}(x^{-i})dx\Big) \\
& + T^{s,i}_{y^i,1,0} \int_{\theta_t^{s,i}}^{\infty} G_y^i(x^i)dx \Big((1-\lambda^{s,i}_{y^i,t}) \int_{-\infty}^{\theta_t^{s,-i}} G_0^{-i}(x^{-i})dx + \lambda^{s,i}_{y^i,t} \int_{-\infty}^{\theta_t^{s,-i}} G_1^{-i}(x^{-i})dx\Big) \\
& + T^{s,i}_{y^i,0,1} \int_{-\infty}^{\theta_t^{s,i}} G_y^i(x^i)dx \Big((1-\lambda^{s,i}_{y^i,t}) \int_{\theta_t^{s,-i}}^{\infty} G_0^{-i}(x^{-i})dx + \lambda^{s,i}_{y^i,t} \int_{\theta_t^{s,-i}}^{\infty} G_1^{-i}(x^{-i})dx\Big) \\
& + T^{s,i}_{y^i,1,1} \int_{\theta_t^{s,i}}^{\infty} G_y^i(x^i)dx \Big((1-\lambda^{s,i}_{y^i,t}) \int_{\theta_t^{s,-i}}^{\infty} G_0^{-i}(x^{-i})dx + \lambda^{s,i}_{y^i,t} \int_{\theta_t^{s,-i}}^{\infty} G_1^{-i}(x^{-i})dx\Big) \\
=\ & \Big(T^{s,i}_{y^i,0,0}\mathbb{G}_y^i(\theta_t^{s,i}) + T^{s,i}_{y^i,1,0}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\Big)\tilde{\mathbb{G}}_{y^i}^{-i}(\lambda^{s,i}_{y^i,t}, \theta_t^{s,-i}) \\
& + \Big(T^{s,i}_{y^i,0,1}\mathbb{G}_y^i(\theta_t^{s,i}) + T^{s,i}_{y^i,1,1}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\Big)\big(1-\tilde{\mathbb{G}}_{y^i}^{-i}(\lambda^{s,i}_{y^i,t}, \theta_t^{s,-i})\big),
\end{aligned} \tag{5.5}$$

where $\mathbb{G}_y^i(\theta_t^{s,i}) = \int_{-\infty}^{\theta_t^{s,i}} G_y^i(x^i)dx = \mathbb{P}(D^i = 0|S = s, Y^i = y^i)$. The quantity $\tilde{\mathbb{G}}_{y^i}^{-i}(\lambda^{s,i}_{y^i,t}, \theta_t^{s,-i}) = (1-\lambda^{s,i}_{y^i,t})\mathbb{G}_0^{-i}(\theta_t^{s,-i}) + \lambda^{s,i}_{y^i,t}\mathbb{G}_1^{-i}(\theta_t^{s,-i})$refers to $\mathbb{P}(D^{-i} = 0|S = s, Y^i = y^i)$. The value of $\mathbb{G}_y^i(\theta_t^{s,i})$ is dependent on $\alpha^{s,i}$ and $\alpha^{-s,i}$ through $\theta^{s,i}$. The value of $\tilde{\mathbb{G}}_{y^i}^{-i}(\lambda^{s,i}_{y^i,t}, \theta_t^{s,-i})$ is dependent on $\alpha^{s,-i}$ and $\alpha^{-s,-i}$ through $\theta^{s,-i}$, as well as on β^s, $\alpha^{s,i}$ and $\alpha^{s,-i}$ through $\lambda^{s,i}_{y^i,1}$.

Next we specify the dynamics of joint probabilities $\beta^{s,i}_{t+1}$. At each time step, the users who fit with both item categories consist of users with four different joint fit states in the previous time step: users who were previously not a fit with either $\mathcal{G}_J$ or $\mathcal{G}_K$; users who were a fit with $\mathcal{G}_J$ but not with $\mathcal{G}_K$; users who were a fit with $\mathcal{G}_K$ but not with $\mathcal{G}_J$; and users who were already a fit with both item categories. Mathematically, $\forall s \in \{a, b\}, i \in \{j, k\}$, this is represented as follows:

$$\begin{aligned}
\beta^s_{1,1,t+1} =& f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, 0, 0)\beta^s_{0,0,t} + f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, 0, 1)\beta^s_{0,1,t} \\
& + f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, 1, 0)\beta^s_{1,0,t} + f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, 1, 1)\beta^s_{1,1,t},
\end{aligned} \tag{5.6}$$

where $f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, y^j, y^k)$ represents $\mathbb{P}(y^j_{t+1} = 1, y^k_{t+1} = 1|y^j_t = y^j, y^k_t = y^k)$. When

$\pi_t^{s,i}(x) = \mathbb{1}(x \geq \theta_t^{s,i})$, it is computed as

$$\begin{aligned} f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, y^j, y^k) :=& T^{s,j}_{y^j,0,0} T^{s,k}_{y^k,0,0} \int_{-\infty}^{\theta_t^{s,j}} G^j_{y^j}(x^j)dx \int_{-\infty}^{\theta_t^{s,k}} G^k_{y^k}(x^k)dx \\ &+ T^{s,j}_{y^j,1,0} T^{s,k}_{y^k,0,1} \int_{\theta_t^{s,j}}^{\infty} G^j_{y^j}(x^j)dx \int_{-\infty}^{\theta_t^{s,k}} G^k_{y^k}(x^k)dx \\ &+ T^{s,j}_{y^j,0,1} T^{s,k}_{y^k,1,0} \int_{-\infty}^{\theta_t^{s,j}} G^j_{y^j}(x^j)dx \int_{\theta_t^{s,k}}^{\infty} G^k_{y^k}(x^k)dx \\ &+ T^{s,j}_{y^j,1,1} T^{s,k}_{y^k,1,1} \int_{\theta_t^{s,j}}^{\infty} G^j_{y^j}(x^j)dx \int_{\theta_t^{s,k}}^{\infty} G^k_{y^k}(x^k)dx. \end{aligned} \tag{5.7}$$

The value of $f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, y^j, y^k)$ is dependent on $\alpha_t^{a,j}$ and $\alpha_t^{b,j}$ through $\theta_t^{s,j}$, and dependent on $\alpha_t^{a,k}$ and $\alpha_t^{b,k}$ through $\theta_t^{s,k}$.

5.4.1 Existence and Uniqueness of equilibrium

Suppose a system is updated repeatedly according to the dynamics formulated above, do fit rates evolve toward a particular direction over time? Do they converge to and maintain a certain value? To answer these questions, we first show that a system with threshold policy and the dynamics as in Equation (5.3) and Equation (5.6) will always converge over time to at least one equilibrium point, i.e., $\alpha_{t+1}^{s,i} = \alpha^t, \beta_{t+1}^s = \beta_t^s, \forall s \in \{a,b\}, i \in \{j,k\}$.

Theorem 1 (Existence)**.** *Given the dynamics as in Equation* (5.3) *and Equation* (5.6) *with a threshold policy* $\theta(\alpha^{a,j}, \alpha^{b,j}, \alpha^{a,k}, \alpha^{b,k})$ *that is continuous in* $\alpha^{a,j}, \alpha^{b,j}, \alpha^{a,k}, \alpha^{b,k}$. $\forall T^{a,j}_{y^i,d^i,d^{-k}} \in (0,1)$*, there exists at least one equilibrium* $(\hat{\alpha}^{a,j}, \hat{\alpha}^{b,j}, \hat{\alpha}^{a,k}, \hat{\alpha}^{b,k}, \hat{\beta}^a, \hat{\beta}^b)$.

Next we discuss the conditions for a system to have only one unique equilibrium. We can then focus on scenarios where the uniqueness conditions are satisfied, and conveniently compare different settings by examining the fit rates at their unique equilibrium.

Theorem 2 (Uniqueness)**.** *Consider a system with dynamics as in Equation* (5.3) *and Equation* (5.6)*, and a recommender with either unconstrained or fair optimal threshold policy in both item categories. Denote the quadruplet* $\mathbb{A} = (\alpha^{a,j}, \alpha^{b,j}, \alpha^{a,k}, \alpha^{b,k})$ *and the pair* $\mathbb{B} = (\beta^a, \beta^b)$*. Let* $\phi^s(\mathbb{A}, \beta^s) = \frac{f^s(0,0)(1-\alpha^{s,j}-\alpha^{s,k})+f^s(0,1)\alpha^{s,k}+f^s(1,0)\alpha^{s,j}}{1+f^s(0,1)+f^s(1,0)-f^s(0,0)-f^s(1,1)}$ *and* $h^{s,i}(\mathbb{A}, \mathbb{B}) = \frac{1-g^{1,s,i}(\alpha^{s,i}, \mathbb{A}^{s,i}, \mathbb{B})}{g^{0,s,i}(\mathbb{A}, \mathbb{B})}$*. A sufficient condition for the system to have a unique equilibrium is that* $\forall s \in \{a,b\}, i \in \{j,k\}$,

$$\left|\frac{\partial \phi^s(\mathbb{A}, \beta^{-s})}{\partial \alpha^{s,i}}\right| < 1, \left|\frac{\partial \phi^s(\mathbb{A}, \beta^{-s})}{\partial \alpha^{-s,i}}\right| < 1, \left|\frac{\partial \phi^s(\mathbb{A}, \beta^{-s})}{\partial \alpha^{s,-i}}\right| < 1, \left|\frac{\partial \phi^s(\mathbb{A}, \beta^{-s})}{\partial \alpha^{-s,-i}}\right| < 1,$$

$$\frac{\partial h^{s,i}(\mathbb{A}, \mathbb{B})}{\partial \alpha^{s,i}} \leq 0, \left|\frac{\partial h^{s,i}(\mathbb{A}, \mathbb{B})}{\partial \alpha^{-s,i}}\right| < 1, \left|\frac{\partial h^{s,i}(\mathbb{A}, \mathbb{B})}{\partial \alpha^{s,-i}}\right| < 1, \left|\frac{\partial h^{s,i}(\mathbb{A}, \mathbb{B})}{\partial \alpha^{-s,-i}}\right| < 1, \left|\frac{\partial h^{s,i}(\mathbb{A}, \mathbb{B})}{\partial \beta^s}\right| < 1.$$

The conditions in Theorem 2 imply that for all fit rate or joint fit rate values, a small perturbation to any one of them will only lead to changes in other values by a smaller magnitude. Also note that these conditions can only guarantee the uniqueness of equilibrium but not its stability, so it is possible that the fit rates will oscillate and never converge.

5.5 Long-Term Dynamics

As we have proved the existence of equilibria and provided sufficient conditions for a unique equilibrium, we now focus on cases with a unique equilibrium and study how the long-term fit rates are affected by different cross impact as well as different recommendation policies (unconstrained or fair) in the co-existing item categories.

5.5.1 Influence of Cross Impact

We first discuss how fit rates change as we increase or decrease cross impact probabilities. This will help us understand whether and how we can shape the fit rates in one item category by adjusting cross impact.

Theorem 3. *Given any policy π^j, π^k, let $\hat{\alpha}^{s,i}$ be the fit rate for user group $\mathcal{G}_s$ in item group $\mathcal{G}_i$ at equilibrium. Then $\hat{\alpha}^{s,i}$ is strictly increasing in $T^{s,i,i}_{y^i,d^i,d^i}$ and $T^{s,i,-i}_{y^i,d^i,d^{-i}}$, $y^i \in \{0,1\}, i \in \{j,k\}, d^i, d^{-i} \in \{0,1\}$.*

Theorem 3 says that for any user group $\mathcal{G}_s$ and item category $\mathcal{G}_i$, increasing self impact $T^{s,i,i}_{y^i,d^i,d^i}$ or cross impact $T^{s,i,-i}_{y^i,d^i,d^{-i}}$ will always increase the group's fit rate in the item category at equilibrium $\hat{\alpha}^{s,i}$. Taking university classes as an example, if we want to make a subgroup of students more interested in a class A, we can take two strategies. The first strategy is to make class A more appealing to this subgroup of students. Then once these students take the class, they will have a higher probability to actually like it. The second strategy is that we can modify another class B to promote class A. This means when these students take class B, they are more likely to have a increased interest in class A.

Now that we discussed that fit rates can be increased or decreased by adjusting cross impact for any user group and any item categories. Next we wonder whether inequality can be completely eliminated by adjusting the transition probabilities. We call the equilibrium state where $\hat{\alpha}^{a,i} = \hat{\alpha}^{b,i}$ an equitable equilibrium for item i. Proposition 1 says that it is always feasible to reach an equitable equilibrium if we properly adjust the self impact and cross impact probabilities.

Proposition 1. *For any user group $\mathcal{G}_s$ and item category $\mathcal{G}_i$, given $T^{s,i,i}_{y^i,d^i}$ and $T^{s,i,-i}_{y^i,d^{-i}}$, there always exist $T^{-s,i,i}_{y^i,d^i}$ and $T^{-s,i,-i}_{y^i,d^{-i}}$ that lead to equitable equilibrium, i.e., $\hat{\alpha}^{s,i} = \hat{\alpha}^{-s,i}$.*

5.5.2 Influence of Fairness Intervention

Now we discuss the how fit rates in one item category are influenced by recommendation policies in its co-existing item categories. For example, suppose there are two university classes in which inequalities are of concern. Ideally, we want to mitigate the inequality in both classes. A natural question that arises is, when we apply fair recommendation policy in one class, which supposedly reduces the inequality in this very class, how does it influence the inequality in another class?

Since the direct outcome of fairness intervention for threshold policies is to change the threshold values, we first consider the influence of $\theta^{s,-i}$ on $\hat{\alpha^{s,i}}$. Specifically, we study its influence under three conditions (directions) of cross impact.

Condition 1 (Positive cross impact)**.** $T_{0,0}^{s,i,-i} < T_{0,1}^{s,i,-i}$ *and* $T_{1,0}^{s,i,-i} < T_{1,1}^{s,i,-i}$.

Condition 2 (Negative cross impact)**.** $T_{0,0}^{s,i,-i} > T_{0,1}^{s,i,-i}$ *and* $T_{1,0}^{s,i,-i} > T_{1,1}^{s,i,-i}$.

Condition 3 (Zero cross impact)**.** $T_{0,0}^{s,i,-i} = T_{0,1}^{s,i,-i}$ *and* $T_{1,0}^{s,i,-i} = T_{1,1}^{s,i,-i}$.

Lemma 1. *Given any* $T^{s,i,-i}$*, under Condition 1,* $\hat{\alpha}^{s,i}$ *is strictly decreasing in* $\theta^{s,-i}$*. Under Condition 2,* $\hat{\alpha}^{s,i}$ *is strictly increasing in* $\theta^{s,-i}$*. Under Condition 3,* $\hat{\alpha}^{s,i}$ *is constant in* $\theta^{s,-i}$*.*

Lemma 1 says that the influence of $\theta^{s,-i}$ on $\hat{\alpha^{s,i}}$ is dependent on the direction of cross impact. For example, if the interest in class J can be promoted by exposure in class K, then lowering the recommendation threshold in class K increases the exposure to class K and subsequently increases interest in class J. On the other hand, if the interest in class J will be demoted by exposure in class K, then increasing the recommendation threshold in class K reduces the exposure to class K and subsequently increases interest in class J. If class J and class K are unrelated, then interest in class J is not affected by the recommendation threshold in class K.

Next we discuss how fairness intervention are related. Specifically, $\forall i \in \{j,k\}$, we compare how $\hat{\alpha}^{s,i}$ change when policy in $\mathcal{G}_{-i}$ is unconstrained or subject to fairness constraints. $\forall s \in \{a,b\}, i \in \{j,k\}$, given unconstrained or fair policies $\pi^i, \pi^{-i} \in \{\text{UN}, \text{DP}, \text{EqOpt}\}$, let $\hat{\alpha}^{s,i}_{\pi^i,\pi^{-i}}$ be the fit rate of user group $\mathcal{G}_s$ in item category $\mathcal{G}_i$ at equilibrium. We first limit ourselves to the scenario where $\pi^i = \text{UN}$.

Theorem 4. $\forall s \in \{a,b\}, i \in \{j,k\}$*, given any* $T^{a,j}$*,* $T^{b,j}$*,* $T^{a,k}$ *and* $T^{b,k}$ *as well as an unconstrained policy* π^i_{UN}*, let* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}}$ *be the fit rate of user group* $\mathcal{G}_s$ *in item category* $\mathcal{G}_i$ *at equilibrium when* π^{-i} *is subject to fairness constraints* $\mathcal{C}$*, we have*

- *When* $\hat{\alpha}^{s,-i}_{UN,UN} < \hat{\alpha}^{-s,-i}_{UN,UN}$*, if* $T^{s,i,-i}$ *satisfies Condition 1,* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} > \hat{\alpha}^{s,i}_{UN,UN}$*; if* $T^{s,i,-i}$ *satisfies Condition 2,* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} < \hat{\alpha}^{s,i}_{UN,UN}$*; if* $T^{s,i,-i}$ *satisfies Condition 3,* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} = \hat{\alpha}^{s,i}_{UN,UN}$*.*
- *When* $\hat{\alpha}^{s,-i}_{UN,UN} > \hat{\alpha}^{-s,-i}_{UN,UN}$*, if* $T^{s,i,-i}$ *satisfies Condition 1,* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} < \hat{\alpha}^{s,i}_{UN,UN}$*; if* $T^{s,i,-i}$ *satisfies Condition 2,* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} > \hat{\alpha}^{s,i}_{UN,UN}$*; if* $T^{s,i,-i}$ *satisfies Condition 3,* $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} = \hat{\alpha}^{s,i}_{UN,UN}$*.*

- *When* $\hat{\alpha}^{s,-i}_{UN,UN} = \hat{\alpha}^{-s,-i}_{UN,UN}$, $\hat{\alpha}^{s,i}_{UN,\mathcal{C}} = \hat{\alpha}^{s,i}_{UN,UN}$.

Theorem 4 says that, when $\pi^i = \text{UN}$, how fit rates are affected by cross policy is dependent on two factors. The first factor is the direction of inequality when cross policy is unconstrained, which means which user group has a lower fit rate at equilibrium. The second factor is the direction of cross impact, i.e., the relation between two item categories such that recommendations in one item category promote, demote or do not affect fit rates in another item category. Note that fairness intervention in one item category always reduces the inequality in itself (given Assumption 3.b and Assumption 3.a and Theorem 4 in [123]). Therefore, if such intervention increases the disparity in the fit rates in its neighboring item categories, it suggests that a trade-off needs to be made since equality cannot be promoted at the same time for both item categories through this fairness intervention.

The patterns are more complicated for scenarios where $\pi^i = \text{EqOpt}$ or $\pi^i = \text{DP}$. Besides the two factors mentioned above (the direction of inequality when cross policy is unconstrained and the direction of cross impact), the analysis of $\hat{\alpha}^{s,i}$ further requires considering a third factor, that is the direction of change in $\hat{\alpha}^{-s,i}$. This is because when π^i is a fair policy, the recommendation decisions for $\mathcal{G}_s$ is dependent on the fit rate of $\mathcal{G}_{-s}$. Therefore, the value of $\hat{\alpha}^{s,i}$ is also affected by the value of $\hat{\alpha}^{-s,i}$. We show the empirical results in Appendix for these more complicated scenarios.

5.6 Experiments

In this section, we validate the results of our theoretical analysis through simulation on synthetic users and items.

5.6.1 Synthetic Data

We create two groups of users and let $p_a = p_b = 0.5$. We generate the fit scores in each item group as $G^i_y = \mathcal{N}(\mu^i_y, (\sigma^i_y)^2), \forall y \in \{0, 1\}$. Specifically, $[\mu^j_0, \mu^j_1, \mu^k_0, \mu^k_1] = [-5, 5, -3, 3]$, $[\sigma^j_0, \sigma^j_1, \sigma^k_0, \sigma^k_1] = [5, 5, 5, 5]$. The compound transition probabilities are calculated as $T^{s,i}_{y^i,d^i,d^{-i}} = \frac{1}{2}\big((T^{s,i,i}_{y^i,d^i} T^{s,i,-i}_{1,d^{-i}} + (1 - T^{s,i,i}_{y^i,d^i}) T^{s,i,-i}_{0,d^{-i}}) + (T^{s,i,-i}_{y^i,d^{-i}} T^{s,i,i}_{1,d^i} + (1 - T^{s,i,-i}_{y^i,d^{-i}}) T^{s,i,i}_{0,d^i})\big), \forall s \in \{a, b\}, i \in \{j, k\}$. The self impact probabilities are $[T^{a,i,i}_{0,0}, T^{a,i,i}_{0,1}, T^{a,i,i}_{1,0}, T^{a,i,i}_{1,1}] = [0.1, 0.5, 0.5, 0.7]$, $[T^{b,i,i}_{0,0}, T^{b,i,i}_{0,1}, T^{b,i,i}_{1,0}, T^{b,i,i}_{1,1}] = [0.4, 0.5, 0.5, 0.9], \forall i \in \{j, k\}$. We will examine different settings of cross impact $T^{a,j,k}$ and $T^{b,j,k}$ below.

The cross impact probabilities are set as follows. First, since zero cross impact means the fit rate in one item category is not affected by its neighboring item categories, we set zero cross impact (i.e., Condition 3) to be $T^{s,i,-i}_{1,d^{-i}} = 1$, $T^{s,i,-i}_{0,d^{-i}} = 0, \forall s \in \{a, b\}$. These values ensure that $T^{s,i}_{y^i,d^i,0} = T^{s,i}_{y^i,d^i,1} = T^{s,i,i}_{y^i,d^i}$, $\forall i \in \{j, k\}, d^i, y^i \in \{0, 1\}$. Then we denote this zero cross

impact as $\tilde{T}^s$ and use it as a base to set positive and negative cross impact probabilities. Let $C^{s,i}$ be a cross impact parameter such that its sign (positive or negative) represents the direction of cross impact while its magnitude $|C^{s,i}|$ represents how strong the cross impact is. Specifically, for positive cross impact (i.e., Condition 1, where receiving recommendations in $\mathcal{G}_J$ and $\mathcal{G}_K$ increases the level of fit in each other), we set $C^{s,i} \in (0, 1]$ and $T^{s,i,-i}_{y^i,1} = \min(1, \max(0, \tilde{T}^s_{y^i,d^i,1} + C^{s,i})), \forall y^i \in \{0, 1\}$; for negative cross impact (i.e., Condition 2, which means receiving recommendations in $\mathcal{G}_J$ and $\mathcal{G}_K$ decrease the level of fitting in each other), we set $C^{s,i} \in [-1, 0)$ and $T^{s,i,-i}_{y^i,1} = \min(1, \max(0, \tilde{T}^s_{y^i,d^i,1} + C^{s,i})), \forall y^i \in \{0, 1\}$; the min and max functions are used to keep the probabilities between 0 and 1. Furthermore, for simplicity, we let $T^{s,j,k} = T^{s,k,j}, \forall s \in \{a, b\}$, which means the cross impact is reciprocal and recommendations in $\mathcal{G}_J$ and $\mathcal{G}_K$ have the same cross impact on each other.

We try all combinations of cross impact $C^{a,j}, C^{b,j} \in [-1.0, -0.75, -0.5, -0.25, 0.0, 0.25, 0.5, 0.75, 1.0]$. This means the cross impact can be group variant, for example, it is possible that more exposure to computer science classes may make male students more interested in math while less so for female students. Then we apply different policies $\pi^j, \pi^k \in \{\text{UN}, \text{DP}, \text{EqOpt}\}$. At each time step, we update the fit rates and joint fit rates as indicated in Equation (5.3) and Equation (5.6). We run simulations till convergence and record the values of fit rates and joint fit rates ($\alpha^{a,j}$, $\alpha^{b,j}$, $\alpha^{a,k}$, $\alpha^{b,k}$, β^a, and β^b) at equilibriums.

5.6.2 Influence of Cross Impact

We start with showing how the fit rates and joint fit rates change with different cross impact probabilities. Figure 5.1 shows the results when $\pi^j = \pi^k = \text{UN}$. We can see that $\hat{\alpha}^{a,j}$ and increases as $C^{a,j}$ increases, $\hat{\alpha}^{b,j}$ increases as $C^{b,j}$ increases. This behavior aligns with the conclusion in Theorem 3. Similar patterns are observed when we switch recommendation policies from unconstrained (π_{UN}) to fair (π_{EqOpt} or π_{DP}) (in appendix). Note that in this particular setting, because unconstrained policies are applied, $\hat{\alpha}^{a,j}$ is independent of $C^{b,i}$ and $\hat{\alpha}^{b,j}$ is independent of $C^{a,i}$. Further, Figure 5.2 shows the values of $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$, which stands for inequality in $\mathcal{G}_J$. For this given self impact in our simulation, as $C^{a,j}$ increases or $C^{b,j}$ decreases, the values of $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ change from positive to negative, crossing the boundary that equitable equilibriums sit at, echoing the conclusion in Proposition 1.

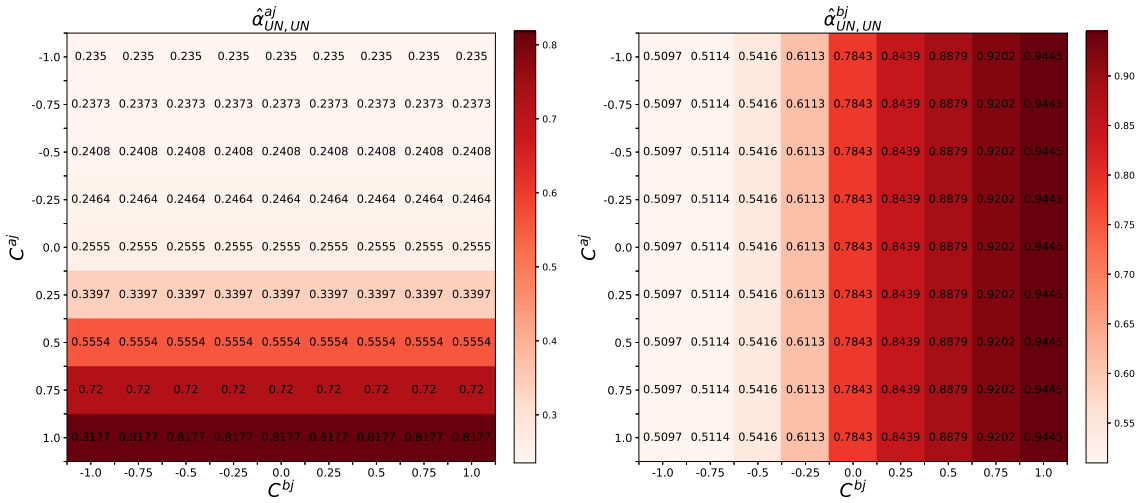

Figure 5.1: Values of $\alpha^{a,j}$ and $\alpha^{b,j}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0). The policies in both item categories are unconstrained.

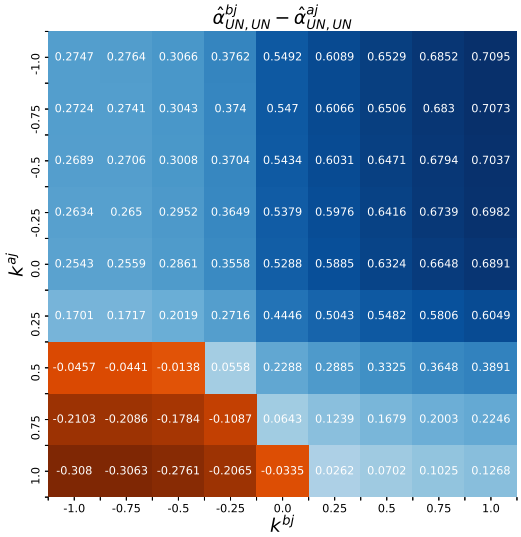

Figure 5.2: Values of $\alpha^{b,j} - \alpha^{a,j}$ with different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0). The policies in both item categories are unconstrained. Cells with positive values are colored in blue and cells with negative values are in red.

Also, for each cross impact pair, we run 5 trials with different initialization. We show two sets of sample trajectories in Figure 5.3. These two sets of trajectories are generated with unconstrained policy in both item categories. As for cross impact probabilities, in the first row, $C^{a,j} = C^{b,j} = 0.0$, which means zero cross impact as in Condition 3; in the second row, $C_{a,j} = 0.25$, $C_{b,j} = -0.25$, which means recommendations in $\mathcal{G}_K$ has a positive cross impact

on $\alpha^{a,j}$ and has a negative cross impact on $\alpha^{b,j}$. The plots in the first column are $\alpha^{a,j}$ and $\alpha^{b,j}$; the plots in the second column are $\alpha^{a,k}$ and $\alpha^{b,k}$; and the plots in the third column are β^a and β^b. Different sets of initialization are annotated with different colors. We can see that the trajectories within each plot all converge to the same point, this validates the conclusions in Theorem 1 and Theorem 2.

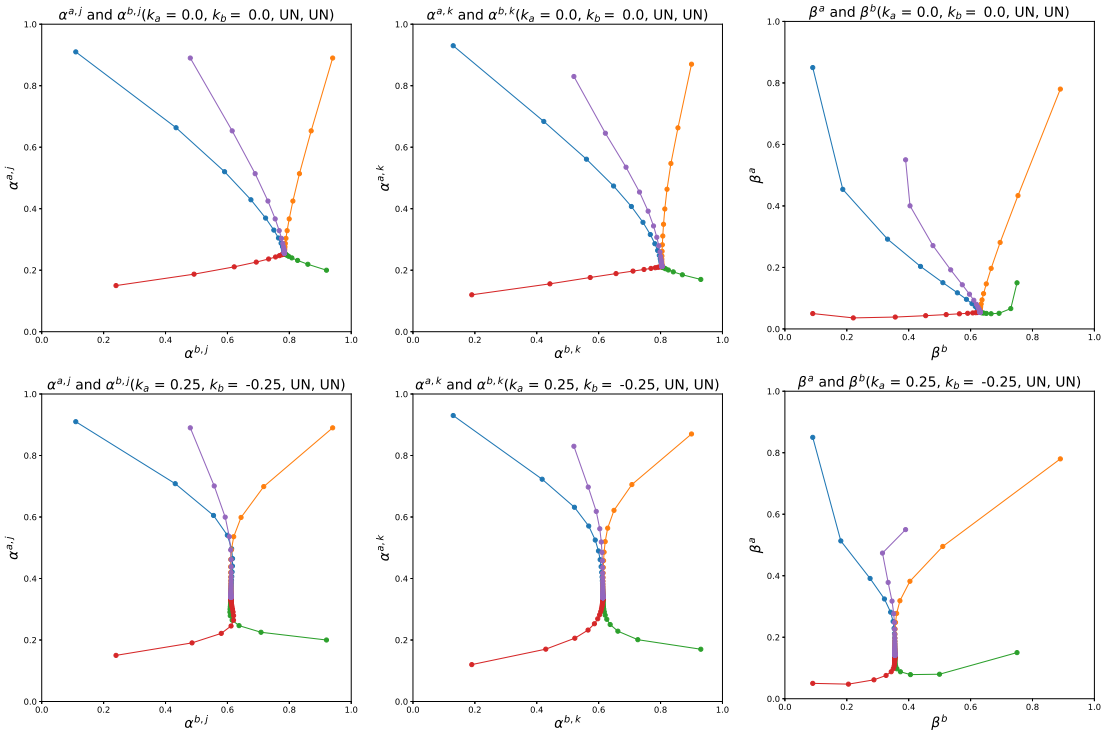

Figure 5.3: Trajectories of $\alpha^{a,j}$ vs. $\alpha^{b,j}$ (left), $\alpha^{a,k}$ vs. $\alpha^{b,k}$ (middle) and β^a vs. β^b (right). In the first row, $C^{a,j} = C^{b,j} = 0.0$. In the second row, $C^{a,j} = 0.25$, $C^{b,j} = -0.25$. Different sets of initialization are annotated with different colors.

5.6.3 Influence of Recommendation Policy

Last, we show how $\hat{\alpha}^{s,i}$ changes as we switch π^{-i} from unconstrained policy to fair policies. For example, Figure 5.4 shows the value of $\hat{\alpha}^{s,j}_{\text{UN,DP}} - \hat{\alpha}^{s,j}_{\text{UN,UN}}$, which means the changes in $\hat{\alpha}^{s,j}$ when $\pi^j = \text{UN}$ and π^k changes from unconstrained to DP-constrained. If $\hat{\alpha}^{s,j}_{\text{UN,DP}} - \hat{\alpha}^{s,j}_{\text{UN,UN}}$ is negative, then $\hat{\alpha}^{s,j}$ decreases; if $\hat{\alpha}^{s,j}_{\text{UN,DP}} - \hat{\alpha}^{s,j}_{\text{UN,UN}}$ is positive, then $\hat{\alpha}^{s,j}$ increases. The two scatter plots are distinguished by different directions of inequality when $\pi^k = \text{UN}$. In the plot on the left, $\hat{\alpha}^{s,j}_{\text{UN,UN}} < \hat{\alpha}^{-s,j}_{\text{UN,UN}}$. In the plot on the right, $\hat{\alpha}^{s,j}_{\text{UN,UN}} > \hat{\alpha}^{-s,j}_{\text{UN,UN}}$. The x-axis indicates the value of cross impact $C^{s,j}$ ranging from -1.0 to 1.0. We observe that when $\hat{\alpha}^{s,j}_{\text{UN,UN}} < \hat{\alpha}^{-s,j}_{\text{UN,UN}}$ (the plot on the left), if the cross impact $C^{s,j}$ is negative, $\hat{\alpha}^{s,j}$ decreases; if $C^{s,j}$ is positive, $\hat{\alpha}^{s,j}$ increases. This means if a user group is the disadvantaged group (lower

fit rate at equilibrium) in $\mathcal{G}_K$ with an unconstrained policy, and we know $\mathcal{G}_K$ promotes interest in $\mathcal{G}_J$, then we are expected to see an increase in the fit rate of this user group in $\mathcal{G}_J$ when π^K is switched to a fair policy; on the other hand, if we know $\mathcal{G}_K$ demotes interest in $\mathcal{G}_J$, then we are expected to see an decrease in the fit rate of this user group in $\mathcal{G}_J$ when π^K is switched to a fair policy. The plot on the right shows the opposite pattern. When $\hat{\alpha}^{s,j}_{\text{UN,UN}} > \hat{\alpha}^{-s,j}_{\text{UN,UN}}$, if the cross impact $C^{s,j}$ is negative, $\hat{\alpha}^{s,j}$ increases; if $C^{s,j}$ is positive, $\hat{\alpha}^{s,j}$ decreases. These results align with the conclusion in Theorem 4. To summarize, if we aim to increase the fit rate of a user group $\mathcal{G}_s$ in an item category $\mathcal{G}_i$, one option is to enforce fairness constraints on another item category $\mathcal{G}_{i'}$ that a) is currently with an unconstrained policy and $\mathcal{G}_s$ is the disadvantaged group, and b) $\mathcal{G}_{i'}$ promotes interest in $\mathcal{G}_i$. A second option is to enforce fairness constraints on another item category $\mathcal{G}_{i'}$ that a) is currently with an unconstrained policy and $\mathcal{G}_s$ is the advantaged group, and b) $\mathcal{G}_{i'}$ demotes interest in $\mathcal{G}_i$. The points in each vertical line correspond to different values of $C^{-s,j}$, which does not affect the direction of change in $\hat{\alpha}^{s,j}$.

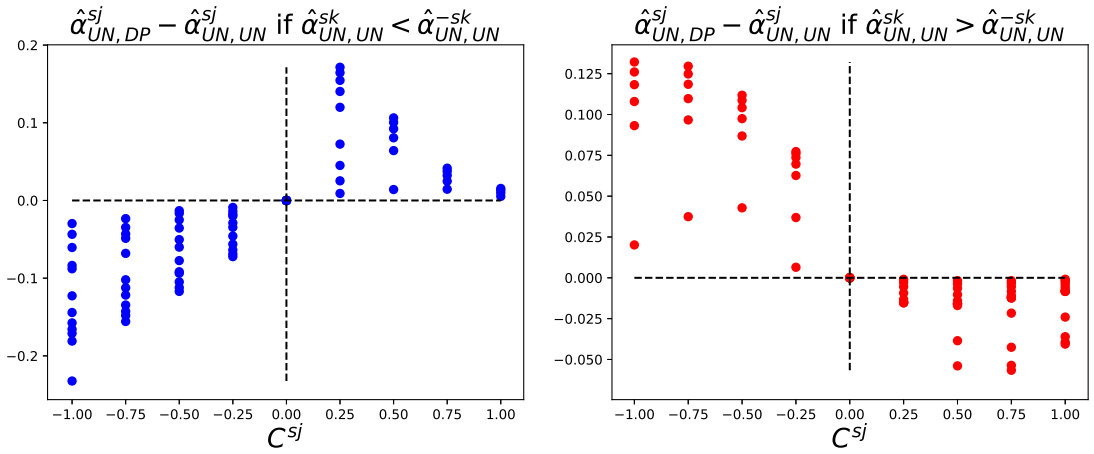

Figure 5.4: The difference in $\hat{\alpha}^{s,j}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to DP-constrained, under different cross impact. The plot on the left shows the cases when $\hat{\alpha}^{s,k}_{\text{UN,UN}} < \hat{\alpha}^{-s,k}_{\text{UN,UN}}$, the plot on the right shows the cases when $\hat{\alpha}^{s,k}_{\text{UN,UN}} > \hat{\alpha}^{s,k}_{\text{UN,UN}}$.

5.7 Discussion

To summarize, in this work, we theoretically study the long-term dynamic of the fit between users and items in recommender systems. Specifically, we characterize the changes in the fit stimulated by recommender decision as conditional transition probabilities, then we examine how fit rates change over time. We prove there always exists at least one equilibrium of the system, and the equilibrium is unique when the provided conditions are satisfied. We further discuss the influence of different cross impact and fairness intervention. Last, we run thorough experiments on synthetic data to validate our analysis. The results of our

study demonstrate that understanding the inequality in recommender systems requires close examination of the user dynamics as well as the relation between different item categories. For future research directions, since our analysis is limited to cases with a unique equilibrium, it is worthwhile to explore scenarios with multiple equilibrium of the system. We also do not discuss the convergence rate, which is important if we consider the scenarios where recommendation policy is constantly changing before a system could converge.

5.8 Appendix

Table of Notation

Table 5.1: Table of notation in Chapter 5.

$\mathcal{G}_s$	user groups, $s \in \{a, b\}$
$\mathcal{G}_i$	item categories, $i \in \{j, k\}$
p_s	user group ratio
$Y_t^{s,i}$	fit state of $\mathcal{G}_s$ in $\mathcal{G}_i$ at t, $Y_t^{s,i} \in \{0, 1\}$
$X_t^{s,i}$	fit score of $\mathcal{G}_s$ in $\mathcal{G}_i$ at t
$D_t^{s,i}$	decision for an individual from $\mathcal{G}_s$ in $\mathcal{G}_i$ at t, $D_t^{s,i} \in \{0, 1\}$
$\pi^{s,i}$	policy for $\mathcal{G}_s$ in $\mathcal{G}_i$ at t
π^i	policy in $\mathcal{G}_i$ at t
$\theta_t^{s,i}$	threshold of $\mathcal{G}_s$ in $\mathcal{G}_i$ at t
G_y^i	fit score distribution in $\mathcal{G}_i$ when $Y = y$
$\mathbb{G}_y^i$	CDF of G_y^i, $\mathbb{G}_y^i = \int_{-\infty}^{\theta_t^{s,i}} G_y^i(x^i)dx$
$\widetilde{\mathbb{G}}_{y^i}^{-i}$	$\mathbb{P}(D^{-i} = 0 \mid S = s, Y^i = y^i)$, $\widetilde{\mathbb{G}}_{y^i}^{-i}(\lambda_{y^i,t}^{s,i}, \theta_t^{s,-i}) = (1 - \lambda_{y^i,t}^{s,i})\mathbb{G}_0^{-i}(\theta_t^{s,-i}) + \lambda_{y^i,t}^{s,i}\mathbb{G}_1^{-i}(\theta_t^{s,-i})$
$\alpha_t^{s,i}$	fit rate of $\mathcal{G}_s$ in $\mathcal{G}_i$ at t
β_t^s	joint fit rate of $\mathcal{G}_s$ at t
λ_t^s	conditional fit rate of $\mathcal{G}_s$ at t
$T^{s,i,i}$	self impact, $\mathcal{P}(Y_{t+1}^i \mid Y_t^i = y, D_t^i = d^i, S = s)$
$T^{s,i,-i}$	cross impact, $\mathcal{P}(Y_{t+1}^i \mid Y_t^i = y, D_t^{-i} = d^{-i}, S = s)$
$T^{s,i}$	compound transition probabilities, $\mathcal{P}(Y_{t+1}^i \mid Y_t^i = y, D_t^i = d^i, D_t^{-i} = d^{-i}, S = s)$
γ_t^s	$\mathcal{P}(Y_t = 1 \mid X_t - x, S = s)$
$\mathbb{A}$	the quadruplet $(\alpha^{a,j}, \alpha^{b,j}, \alpha^{a,k}, \alpha^{b,k})$
$\mathbb{A}^{s,i}$	triplet $(\alpha^{-s,i}, \alpha^{-s,-i}, \alpha^{s,-i})$
$\mathbb{B}$	the pair (β^a, β^b)
$\hat{\alpha^{s,i}}$	fit rate of $\mathcal{G}_s$ in $\mathcal{G}_i$ at equilibrium
$\hat{\beta}^s$	joint fit rate of $\mathcal{G}_s$at equilibrium

Derivations

Aggregated Transition Dynamic Given y.

$$
\begin{aligned}
g^{y,s,i} :=\ & T^{s,i}_{y^i,0,0}\int_{-\infty}^{\theta_t^{s,i}} G_y^{s,i}(x^i)dx\Big((1-\lambda^{s,i})\int_{-\infty}^{\theta_t^{s,-i}} G_0^{s,-i}(x^{-i})dx+\lambda^{s,i}\int_{-\infty}^{\theta_t^{s,-i}} G_1^{s,-i}(x^{-i})dx\Big)\\
&+T^{s,i}_{y^i,1,0}\int_{\theta_t^{s,i}}^{\infty} G_y^{s,i}(x^i)dx\Big((1-\lambda^{s,i})\int_{-\infty}^{\theta_t^{s,-i}} G_0^{s,-i}(x^{-i})dx+\lambda^{s,i}\int_{-\infty}^{\theta_t^{s,-i}} G_1^{s,-i}(x^{-i})dx\Big)\\
&+T^{s,i}_{y^i,0,1}\int_{-\infty}^{\theta_t^{s,i}} G_y^{s,i}(x^i)dx\Big((1-\lambda^{s,i})\int_{\theta_t^{s,-i}}^{\infty} G_0^{s,-i}(x^{-i})dx+\lambda^{s,i}\int_{\theta_t^{s,-i}}^{\infty} G_1^{s,-i}(x^{-i})dx\Big)\\
&+T^{s,i}_{y^i,1,1}\int_{\theta_t^{s,i}}^{\infty} G_y^{s,i}(x^i)dx\Big((1-\lambda^{s,i})\int_{\theta_t^{s,-i}}^{\infty} G_0^{s,-i}(x^{-i})dx+\lambda^{s,i}\int_{\theta_t^{s,-i}}^{\infty} G_1^{s,-i}(x^{-i})dx\Big)\\
=\ & T^{s,i}_{y^i,0,0}\mathbb{G}_y^i(\theta_t^{s,i})\big((1-\lambda^{s,i})\mathbb{G}_0^{-i}(\theta_t^{s,-i})+\lambda^{s,i}\mathbb{G}_1^{-i}(\theta_t^{s,-i})\big)\\
&+T^{s,i}_{y^i,1,0}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\big((1-\lambda^{s,i})\mathbb{G}_0^{-i}(\theta_t^{s,-i})+\lambda^{s,i}\mathbb{G}_1^{-i}(\theta_t^{s,-i})\big)\\
&+T^{s,i}_{y^i,0,1}\mathbb{G}_y^i(\theta_t^{s,i})\big((1-\lambda^{s,i})(1-\mathbb{G}_0^{-i}(\theta_t^{s,-i}))+\lambda^{s,i}(1-\mathbb{G}_1^{-i}(\theta_t^{s,-i}))\big)\\
&+T^{s,i}_{y^i,1,1}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\big((1-\lambda^{s,i})(1-\mathbb{G}_0^{-i}(\theta_t^{s,-i}))+\lambda^{s,i}(1-\mathbb{G}_1^{-i}(\theta_t^{s,-i}))\big)\\
=\ & \Big(T^{s,i}_{y^i,0,0}\mathbb{G}_y^i(\theta_t^{s,i})+T^{s,i}_{y^i,1,0}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\Big)\Big((1-\lambda^{s,i})\mathbb{G}_0^{-i}(\theta_t^{s,-i})+\lambda^{s,i}\mathbb{G}_1^{-i}(\theta_t^{s,-i})\Big)\\
&+\Big(T^{s,i}_{y^i,0,1}\mathbb{G}_y^i(\theta_t^{s,i})+T^{s,i}_{y^i,1,1}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\Big)\Big((1-\lambda^{s,i})(1-\mathbb{G}_0^{-i}(\theta_t^{s,-i}))+\lambda^{s,i}(1-\mathbb{G}_1^{-i}(\theta_t^{s,-i}))\Big)\\
=\ & \Big(T^{s,i}_{y^i,0,0}\mathbb{G}_y^i(\theta_t^{s,i})+T^{s,i}_{y^i,1,0}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\Big)\tilde{\mathbb{G}}_y^{-i}(\theta_t^{s,-i})\\
&+\Big(T^{s,i}_{y^i,0,1}\mathbb{G}_y^i(\theta_t^{s,i})+T^{s,i}_{y^i,1,1}\big(1-\mathbb{G}_y^i(\theta_t^{s,i})\big)\Big)\big(1-\tilde{\mathbb{G}}_y^{-i}(\theta_t^{s,-i})\big)
\end{aligned}
\tag{5.8}
$$

Proofs

Proof of Theorem 1

Proof. We denote the quadruplet $\mathbb{A}=(\alpha^{a,j},\alpha^{b,j},\alpha^{a,k},\alpha^{b,k})$, and the pair $\mathbb{B}=(\beta^a,\beta^b)$. At equilibrium, $\forall s\in\{a,b\}, i\in\{j,k\}$, we have

$$\alpha^{s,i}_{t+1}=\alpha^{s,i}_t=g^{0,s,i}(\mathbb{A},\mathbb{B})(1-\alpha^{s,i}_t)+g^{1,s,i}(\mathbb{A},\mathbb{B})\alpha^{s,i}_t, \forall s\in a,b, i\in\{j,k\}. \tag{5.9}$$

We denote triplet $(\alpha^{-s,i},\alpha^{-s,-i},\alpha^{s,-i})$ as $\mathbb{A}^{s,i}$ where $-s=\{a,b\}\backslash s$ and $-i=\{j,k\}\backslash i$. We also define function $l(\alpha^{s,i}):=\frac{1}{\alpha^{s,i}}-1$ and $h^{s,i}(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B}):=\frac{1-g^{1,s,i}(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B})}{g^{0,s,i}(\mathbb{A},\mathbb{B})}$. Given $\mathbb{A}^{s,i}$ and $\mathbb{B}$, Equation (5.9) is satisfied for $\alpha^{s,i}_t$ when $l(\alpha^{s,i})=h^{s,i}(\mathbb{A},\mathbb{B})$.

First we prove that, $\forall s \in \{a,b\}, i = \{j,k\}$, given a fixed $\mathbb{A}^{s,i}$ and $\mathbb{B}$, there must exist at least one $\bar{\alpha}^{s,i} \in [0,1]$ such that $l^s(\bar{\alpha}^{s,i}) = h^{s,i}(\bar{\alpha}^{s,i}, \mathbb{A}^{s,i}, \mathbb{B})$.

Since $h^{s,i}(\mathbb{A},\mathbb{B})$ is continuous in $\mathbb{G}_y^i$, $\mathbb{G}_y^i$ is continuous in $\theta_t^{s,i}$, $\theta_t^{s,i}$ is continuous in $\alpha^{s,i}$ and $\alpha^{-s,i}$, we know $h^{s,i}(\mathbb{A},\mathbb{B})$ is continuous in $\alpha^{s,i}$ and $\alpha^{-s,i}$. Similarly, since $h^{s,i}(\mathbb{A},\mathbb{B})$ is continuous in $\tilde{\mathbb{G}}_y^{-i}$, $\tilde{\mathbb{G}}_y^{-i}$ is continuous in $\theta_t^{s,-i}$, $\theta_t^{s,-i}$ is continuous in $\alpha^{s,-i}$ and $\alpha^{-s,-i}$, we know $h^{s,i}(\mathbb{A},\mathbb{B})$ is continuous in $\alpha^{s,-i}$ and $\alpha^{-s,-i}$. By definition Equation (5.5), $\tilde{\mathbb{G}}_y^{-i}$ is continuous in β^s and constant in β^{-s}, so $h^{s,i}(\mathbb{A},\mathbb{B})$ is also continuous in β^s and β^{-s}. Therefore, $h^{s,i}(\mathbb{A},\mathbb{B})$ is continuous in $\mathbb{A}$ and $\mathbb{B}$.

Further, because $g^{y,s,i}(\mathbb{A},\mathbb{B})$ is a convex combination of $T^{s,i}_{y^i,0,0}, T^{s,i}_{y^i,0,1}, T^{s,i}_{y^i,1,0}$, and $T^{s,i}_{y^i,1,1}$, $\forall s \in \{a,b\}, \forall i \in \{j,k\}, \alpha^{s,i} \in [0,1]$ we have

$$\min\{T^{s,i}_{y^i,0,0}, T^{s,i}_{y^i,0,1}, T^{s,i}_{y^i,1,0}, T^{s,i}_{y^i,1,1}\} \leq g^{y,s,i}(\mathbb{A},\mathbb{B}) \leq \max\{T^{s,i}_{y^i,0,0}, T^{s,i}_{y^i,0,1}, T^{s,i}_{y^i,1,0}, T^{s,i}_{y^i,1,1}\},$$

so we know

$$0 < \frac{1-\max\{T^{s,i}_{1,0,0}, T^{s,i}_{1,0,1}, T^{s,i}_{1,1,0}, T^{s,i}_{1,1,1}\}}{\min\{T^{s,i}_{0,0,0}, T^{s,i}_{0,0,1}, T^{s,i}_{0,1,0}, T^{s,i}_{0,1,1}\}} \leq h^{s,i}(\mathbb{A},\mathbb{B}) \leq \frac{1-\min\{T^{s,i}_{1,0,0}, T^{s,i}_{1,0,1}, T^{s,i}_{1,1,0}, T^{s,i}_{1,1,1}\}}{\max\{T^{s,i}_{0,0,0}, T^{s,i}_{0,0,1}, T^{s,i}_{0,1,0}, T^{s,i}_{0,1,1}\}} < +\infty$$

Since $l(\alpha^{s,i})$ is continuous and strictly decreasing in $\alpha^{s,i}$ with values from $+\infty$ to 0, for any given $\mathbb{A}^{s,i}$ and $\mathbb{B}$, there must exist at least one $\bar{\alpha}^{s,i}$ such that $l(\bar{\alpha}^{s,i}) = h^{s,i}(\bar{\alpha}^{s,i}, \mathbb{A}^{s,i}, \mathbb{B})$.

Next we prove that for any given $\mathbb{A}$, there always exist a valid $\bar{\beta}^a$ and $\bar{\beta}^b$. From Equation (5.6) we know that β^s is at equilibrium if $\beta^s_{t+1} = \beta^s_t$. Here for brevity, we omit $\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}$, and $\alpha_t^{b,k}$, and denote $f^s(\alpha_t^{a,j}, \alpha_t^{b,j}, \alpha_t^{a,k}, \alpha_t^{b,k}, y^j, y^k)$ as $f^s(y^j, y^k)$. At equilibrium, we need to satisfy $\beta_t^s = f^s(0,0)(1-\alpha^{s,j}-\alpha^{s,k}+\beta_t^s)+f^s(0,1)(\alpha^{s,k}-\beta_t^s)+f^s(1,0)(\alpha^{s,j}-\beta_t^s)+f^s(1,1)\beta_t^s$, therefore

$$\begin{aligned}\beta_t^s = \phi^s(\mathbb{A}) &= \frac{f^s(0,0)(1-\alpha^{s,j}-\alpha^{s,k})+f^s(0,1)\alpha^{s,k}+f^s(1,0)\alpha^{s,j}}{1+f^s(0,1)+f^s(1,0)-f^s(0,0)-f^s(1,1)} \\ &= \frac{f^s(0,0)+\alpha^{s,j}(f^s(1,0)-f^s(0,0))+\alpha^{s,k}(f^s(0,1)-f^s(0,0))}{f^s(0,0)+(f^s(1,0)-f^s(0,0))+(f^s(0,1)-f^s(0,0))+1-f^s(1,1)}.\end{aligned} \tag{5.10}$$

Given Assumption 2, we know $f(1,0) > f(0,0)$ and $f(0,1) > f(0,0)$, so $\alpha^{s,j}(f^s(1,0) - f^s(0,0))+\alpha^{s,k}(f^s(0,1)-f^s(0,0)) < (f^s(1,0)-f^s(0,0))+(f^s(0,1)-f^s(0,0))$, thus $0 < \beta^s < 1$, which means β^s is always a valid probability value.

Finally, we have the six-dimensional space $\{(\mathbb{A},\mathbb{B})\}$ and six non-empty sets of six-dimensional hyper-planes, $C_1 = \{(\bar{\alpha}^{a,j}, \mathbb{A}^{a,j}, \mathbb{B}) : (\mathbb{A}^{a,j}, \mathbb{B}) \in [0,1]^5\}$, $C_2 = \{(\bar{\alpha}^{b,j}, \mathbb{A}^{b,j}, \mathbb{B}) : (\mathbb{A}^{b,j}, \mathbb{B}) \in [0,1]^5\}$, $C_3 = \{(\bar{\alpha}^{a,k}, \mathbb{A}^{a,k}, \mathbb{B}) : (\mathbb{A}^{a,k}, \mathbb{B}) \in [0,1]^5\}$, $C_4 = \{(\bar{\alpha}^{b,k}, \mathbb{A}^{b,k}, \mathbb{B}) : (\mathbb{A}^{b,k}, \mathbb{B}) \in [0,1]^5\}$, $C_5 = \{(\bar{\beta}^a, \mathbb{A}, \beta^b : (\mathbb{A}, \beta^b) \in [0,1]^5\}$, $C_6 = \{(\bar{\beta}^b, \mathbb{A}, \beta^a) : (\mathbb{A}, \beta^a) \in [0,1]^5\}$. Geometrically, any six hyper-planes $(c_1, c_2, c_3, c_4, c_5, c_6)$, $\forall c_i \in C_i, i \in \{1,2,3,4,5,6\}$ must have at least one intersection. This intersection $(\hat{\alpha}^{a,j}, \hat{\alpha}^{b,j}, \hat{\alpha}^{a,k}, \hat{\alpha}^{b,k}, \hat{\beta}^a, \hat{\beta}^b)$ is where the equilibrium condition is satisfied for the entire system. □

Proof of Theorem 2

Proof. The inequality $\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \alpha^{s,i}} \leq 0$ means $h^{s,i}(\mathbb{A},\mathbb{B})$ is non-increasing in $\alpha^{s,i}$. Therefore, given any $\mathbb{A}^{s,i}$ and $\mathbb{B}$, strictly decreasing function $l(\alpha^{s,i})$ and non-increasing function $h^{s,i}(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B})$ must have exactly one intersection, i.e, $\forall \mathbb{A}^{s,i}, \forall \mathbb{B}$, the set $\Psi(\mathbb{A}^{s,i}) = \{\bar{\alpha}^{s,i} : l(\alpha^{s,i})\} = h(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B})$ has only one element, and they constitute function $\bar{\alpha}^{s,i} = \psi(\mathbb{A}^{s,i},\mathbb{B})$. Previously we have already proved that $\forall \mathbb{A}$, there is one unique valid $\bar{\beta}^s = \phi^s(\mathbb{A},\beta^{-s})$.

In the six-dimensional space, $\{(\alpha^{a,j},\alpha^{b,j},\alpha^{a,k},\alpha^{b,k},\beta^a,\beta^b), \alpha^{a,j} \in [0,1], \alpha^{b,j} \in [0,1], \alpha^{a,k} \in [0,1], \alpha^{b,k} \in [0,1], \beta^a \in [0,1], \beta^b \in [0,1]\}$, given six six-dimensional hyper-planes, $c_1 = \{(\bar{\alpha}^{a,j},\mathbb{A}^{a,j},\mathbb{B})) : \bar{\alpha}^{a,j} = \psi(\mathbb{A}^{a,j},\mathbb{B})\}$, $c_2 = \{(\bar{\alpha}^{b,j},\mathbb{A}^{b,j},\mathbb{B}) : \bar{\alpha}^{b,j} = \psi(\mathbb{A}^{b,j},\mathbb{B})\}$, $c_3 = \{(\bar{\alpha}^{a,k},\mathbb{A}^{a,k},\mathbb{B})) : \bar{\alpha}^{a,k} = \psi(\mathbb{A}^{a,k},\mathbb{B}))\}$, $c_4 = \{(\bar{\alpha}^{b,k},\mathbb{A}^{b,k},\mathbb{B}) : \bar{\alpha}^{b,k} = \psi(\mathbb{A}^{b,k},\mathbb{B})\}$, $c_5 = \{(\bar{\beta}^a,\mathbb{A},\beta^b) : \bar{\beta}^a = \phi^a(\mathbb{A},\beta^b)\}$, $c_6 = \{(\bar{\beta}^b,\mathbb{A},\beta^a) : \bar{\alpha}^{b,k} = \phi^b(\mathbb{A},\beta^a)\}$, one sufficient condition to guarantee c_1, c_2, c_3, c_4, c_5, and c_6 to have exact one intersection is that, $\forall s \in \{a,b\}, \forall i \in \{j,k\}$,

$$\Big|\frac{\partial \phi^s(\mathbb{A},\beta^{-s})}{\partial \alpha^{s,i}}\Big| < 1, \Big|\frac{\partial \phi^s(\mathbb{A},\beta^{-s})}{\partial \alpha^{-s,i}}\Big| < 1, \Big|\frac{\partial \phi^s(\mathbb{A},\beta^{-s})}{\partial \alpha^{s,-i}}\Big| < 1, \Big|\frac{\partial \phi^s(\mathbb{A},\beta^{-s})}{\partial \alpha^{-s,-i}}\Big| < 1, \tag{5.11}$$

$$\Big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{-s,i}}\Big| < 1, \Big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{s,-i}}\Big| < 1, \Big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{-s,-i}}\Big| < 1, \Big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \beta^{s}}\Big| < 1. \tag{5.12}$$

Note that by definition, $\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \beta^{-s}} = 0$ and $\frac{\partial \phi^s(\mathbb{A})}{\partial \beta^{-s}} = 0$. We then prove that the four conditions in Equation (5.12) will hold if $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \alpha^{-s,i}}\big| < 1$, $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \alpha^{s,-i}}\big| < 1$, $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \alpha^{-s,-i}}\big| < 1$, and $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \beta^{s}}\big| < 1$, respectively. Denote $v := h^{s,i}(\psi(\mathbb{A}^{s,i},\mathbb{B}),\mathbb{A}^{s,i},\mathbb{B})$, because $l(\psi(\mathbb{A}^{s,i},\mathbb{B})) = h^{s,i}(\psi(\mathbb{A}^{s,i},\mathbb{B}),\mathbb{A}^{s,i},\mathbb{B})$, $\forall \alpha^{-s,i}$,

$$\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{-s,i}} = \frac{\partial l^{-1}(v)}{\partial \alpha^{-s,i}} = \frac{dl^{-1}(v)}{dv}\frac{\partial v}{\partial \alpha^{-s,i}} = \frac{1}{l'\big(l^{-1}(v)\big)}\frac{\partial v}{\partial \alpha^{-s,i}} = -\big(l^{-1}(v)\big)^2\frac{\partial v}{\partial \alpha^{-s,i}}$$

Because $l^{-1}(v) = \psi(\mathbb{A}^{s,i},\mathbb{B}) \in [0,1], -\big(l^{-1}(v)\big)^2 \in [-1,0]$. When $\big|\frac{\partial h^{s,i(\mathbb{A}),\mathbb{B}}}{\partial \alpha^{-s,i}}\big| < 1$, we have $\big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{-s,i}}\big| < 1$. Similarly, we can prove that $\big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{s,-i}}\big| < 1$ holds $\forall \alpha^{s,-i}$ if $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \alpha^{s,-i}}\big| < 1$, $\big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \alpha^{-s,-i}}\big| < 1$ holds $\forall \alpha^{-s,-i}$ if $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \alpha^{-s,-i}}\big| < 1$, $\big|\frac{\partial \psi(\mathbb{A}^{s,i},\mathbb{B})}{\partial \beta^{s}}\big| < 1$ holds $\forall \beta^{s}$ if $\big|\frac{\partial h^{s,i}(\mathbb{A},\mathbb{B})}{\partial \beta^{s}}\big| < 1$. □

Proof of Theorem 3

Proof. According to the definition $h^{s,i}(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B}) := \frac{1-g^{1,s,i}(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B})}{g^{0,s,i}(\mathbb{A},\mathbb{B})}$, $h^{s,i}$ is strictly decreasing in $g^{0,s,i}$ and $g^{1,s,i}$. From Equation (5.3), we know $g^{y,s,i}$ is strictly increasing in any $T^{s,i}_{y^i,d^i,d^{-i}}$, which is further strictly increasing in $T^{s,i,i}_{y^i,d^i}$ and $T^{s,i,-i}_{y^i,d^{-i}}$, according to Assumption 1. Therefore, $h^{s,i}$ is strictly decreasing in any $T^{s,i}_{y^i,d^i,d^{-i}}$. When $h^{s,i}(\alpha^{s,i},\mathbb{A}^{s,i},\mathbb{B})$ decreases, its intersection with $l(\alpha^{s,i})$ increases, therefore, $\hat{\alpha}^{s,i}$ is strictly increasing in $T^{s,i,i}_{y^i,d^i}$ and $T^{s,i,-i}_{y^i,d^{-i}}$. □

Proof of Proposition 1

Proof. When $T^{-s,i}_{y^i,d^i,d^{-i}} \to 1$, $g^{0,-s,i} \approx g^{1,-s,i} \to \Big(\mathbb{G}^i_y(\theta^{-s,i}_t)+\big(1-\mathbb{G}^i_y(\theta^{-s,i}_t)\big)\Big)\tilde{\mathbb{G}}^{-i}_{y^i}(\lambda^{-s,i}_{y^i,0},\theta^{-s,-i}_t)+\Big(\mathbb{G}^i_y(\theta^{-s,i}_t)+\big(1-\mathbb{G}^i_y(\theta^{-s,i}_t)\big)\Big)\big(1-\tilde{\mathbb{G}}^{-i}_{y^i}(\lambda^{-s,i}_{y^i,0},\theta^{-s,-i}_t)\big)=1$, then $h^{s,i}\to 0$, $\hat{\alpha}^{-s,i}\to 1$. On the other hand, when $T\to 0$, $g^{0,s,i}\approx g^{1,-s,i}\to 0$, then $h^{s,i}\to\infty$, $\hat{\alpha}^{-s,i}\to 0$. Given Assumption 1 and Theorem 3, now we know $\hat{\alpha}^{-s,i}$ is continuous and strictly increasing over $T^{-s,i,i}_{y^i,d^i}\in[0,1]$ and $T^{-s,i,-i}_{y^i,d^{-i}}\in[0,1]$ with value varying between 0 and 1. Therefore, given any $T^{s,i,i}_{y^i,d^i}$ and $T^{s,i,-i}_{y^i,d^{-i}}$ which lead to a particular $\hat{\alpha}^{s,i}$, we can always find $T^{-s,i,i}_{y^i,d^i}$ and $T^{-s,i,-i}_{y^i,d^{-i}}$ that lead to $\hat{\alpha}^{a,i}=\hat{\alpha}^{b,i}$. □

Proof of Lemma 1

Proof. We already discussed that $\hat{\alpha}^{s,i}$ is strictly decreasing in $h^{s,i}$, and $h^{s,i}$ is strictly decreasing in $g^{0,s,i}$ and $g^{1,s,i}$. According to Equation (5.5), $\tilde{\mathbb{G}}^{-i}_{y^i}(\lambda^{s,i}_{y^i,1},\alpha^{s,i},\theta^{s,-i}_t)=(1-\lambda^{s,i}_{y^i,1})\mathbb{G}^{-i}_0(\theta^{s,-i}_t)+\lambda^{s,i}_{y^i,1}\mathbb{G}^{-i}_1(\theta^{s,-i}_t)$. As $\theta^{s,-i}$ increases, $\mathbb{G}^{-i}_0(\theta^{s,-i}_t)$ and $\mathbb{G}^{-i}_1(\theta^{s,-i}_t)$ increase, thus $\tilde{\mathbb{G}}^{-i}_{y^i}(\theta^{s,-i}_t)$ increases. Under Condition 1, we know $T^{s,i}_{y^i,d^i,0}<T^{s,i}_{y^i,d^i,1},\forall d^i\in\{0,1\}$ given Assumption 1, so $g^{y,s,i}$ is strictly decreasing in $\tilde{\mathbb{G}}^{-i}_{y^i}(\theta^{s,-i}_t)$. Therefore, $\hat{\alpha}^{s,i}$ is strictly decreasing in $\theta^{s,-i}$. Under Condition 2, we know $T^{s,i}_{y^i,d^i,0}>T^{s,i}_{y^i,d^i,1},\forall d^i\in\{0,1\}$ given Assumption 1, so $g^{y,s,i}$ is strictly increasing in $\tilde{\mathbb{G}}^{-i}_{y^i}(\theta^{s,-i}_t)$. Therefore, $\hat{\alpha}^{s,i}$ is strictly increasing in $\theta^{s,-i}$. Under Condition 3, $g^{y,s,i}$ is constant in $\tilde{\mathbb{G}}^{-i}_{y^i}(\theta^{s,-i}_t)$. Therefore, $\hat{\alpha}^{s,i}$ is constant in $\theta^{s,-i}$. □

Proof of theorem 4

Proof. Let $\theta^{s,-i}_{UN}$ be the threshold for user group $\mathcal{G}_s$ in item category $\mathcal{G}_{-i}$ with an unconstrained policy, and $\theta^{s,-i}_{\mathcal{C}}$ be the threshold for $\mathcal{G}_s$ in $\mathcal{G}_{-i}$ with a fair policy with constraint $\mathcal{C}$. Given a fixed π^i, when $\hat{\alpha}^{s,-i}_{\pi^i,\text{UN}}<\hat{\alpha}^{-s,-i}_{\pi^i,\text{UN}}$, $\theta^{s,-i}_{\mathcal{C}}<\theta^{s,-i}_{\text{UN}}$, $\theta^{s,-i}$ decreases. Given Lemma 1, we know that if $T^{s,i,-i}$ satisfies Condition 1, $\hat{\alpha}^{s,i}$ increases, so $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}>\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$; if $T^{s,i,-i}$ satisfies Condition 2, $\hat{\alpha}^{s,i}$ decreases, so $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}<\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$; if $T^{s,i,-i}$ satisfies Condition 3, $\hat{\alpha}^{s,i}$ remain unchanged, so $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}=\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$.

Similarly, when $\hat{\alpha}^{s,-i}_{\pi^i,\text{UN}}>\hat{\alpha}^{-s,-i}_{\pi^i,\text{UN}}$, $\theta^{s,-i}_{\mathcal{C}}>\theta^{s,-i}_{\text{UN}}$, $\theta^{s,-i}$ increases. Given Lemma 1, we know that if $T^{s,i,-i}$ satisfies Condition 1, $\hat{\alpha}^{s,i}$ decreases, so $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}<\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$; if $T^{s,i,-i}$ satisfies Condition 2, $\hat{\alpha}^{s,i}$ increases, so $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}>\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$; if $T^{s,i,-i}$ satisfies Condition 3, $\hat{\alpha}^{s,i}$ remain unchanged, so $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}=\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$.

Last, when $\hat{\alpha}^{s,-i}=\hat{\alpha}^{-s,-i}$, $\theta^{s,-i}_{\text{UN}}=\theta^{s,-i}_{\mathcal{C}}$, $\theta^{-s,-i}_{\text{UN}}=\theta^{s,-i}_{\mathcal{C}}$. Therefore, we always have $\hat{\alpha}^{s,i}_{\pi^i,\mathcal{C}}=\hat{\alpha}^{s,i}_{\pi^i,\text{UN}}$. □

Full Experiment Results

Influence of Cross Impact. First we show the influence of cross impact under all policy combinations.

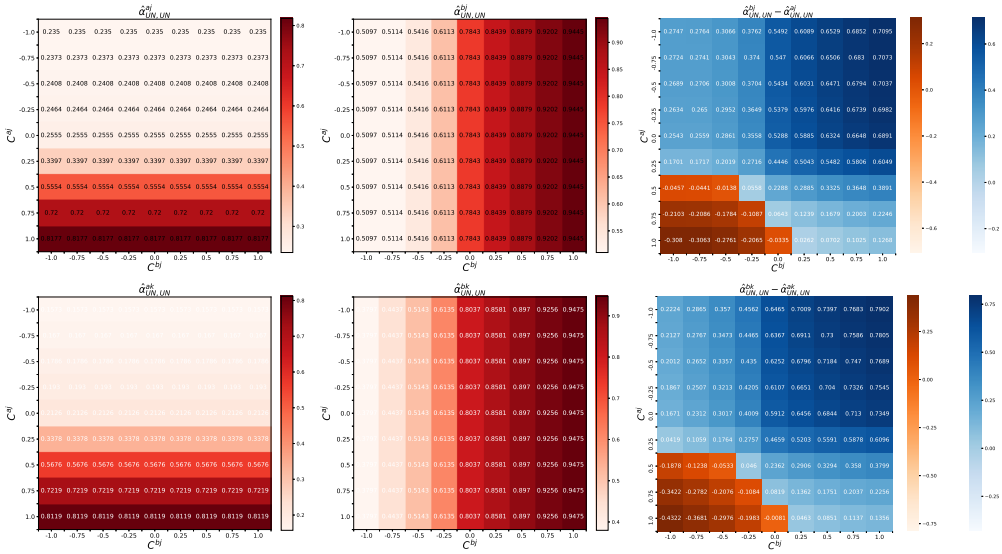

Figure 5.5: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), both π^j and π^k are unconstrained.

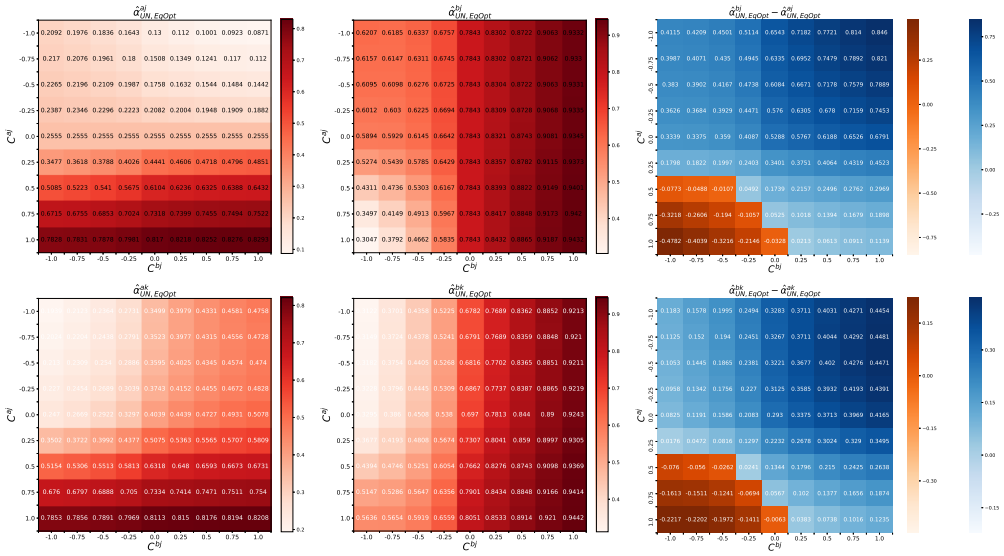

Figure 5.6: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), π^j is unconstrained and π^k is EqOpt-constrained.

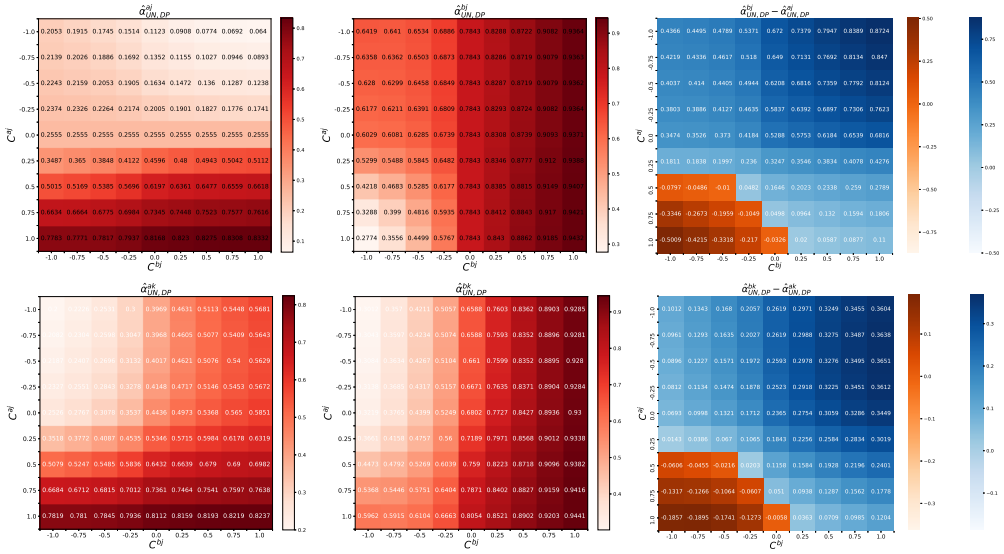

Figure 5.7: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), π^j is unconstrained and π^k is DP-constrained.

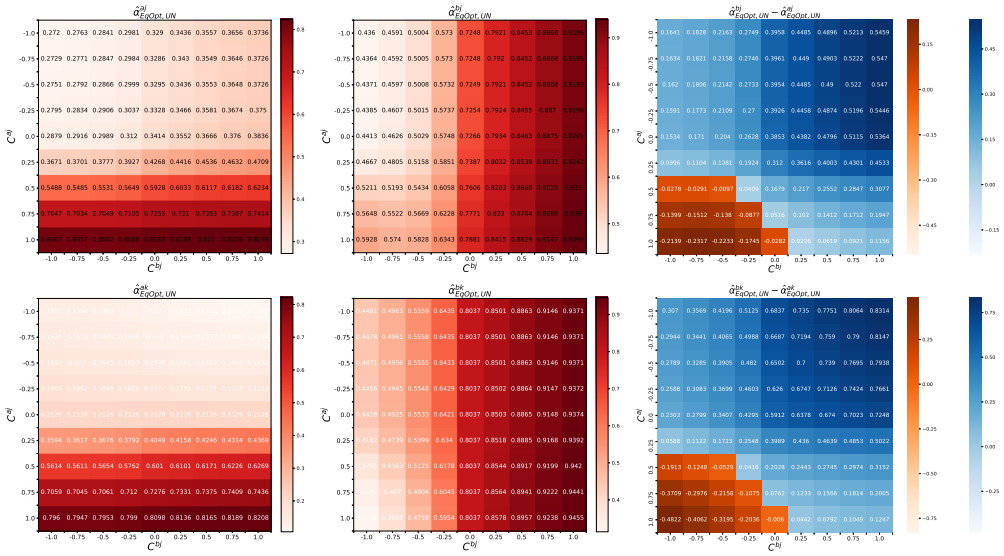

Figure 5.8: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), π^j is EqOpt-constrained and π^k is unconstrained.

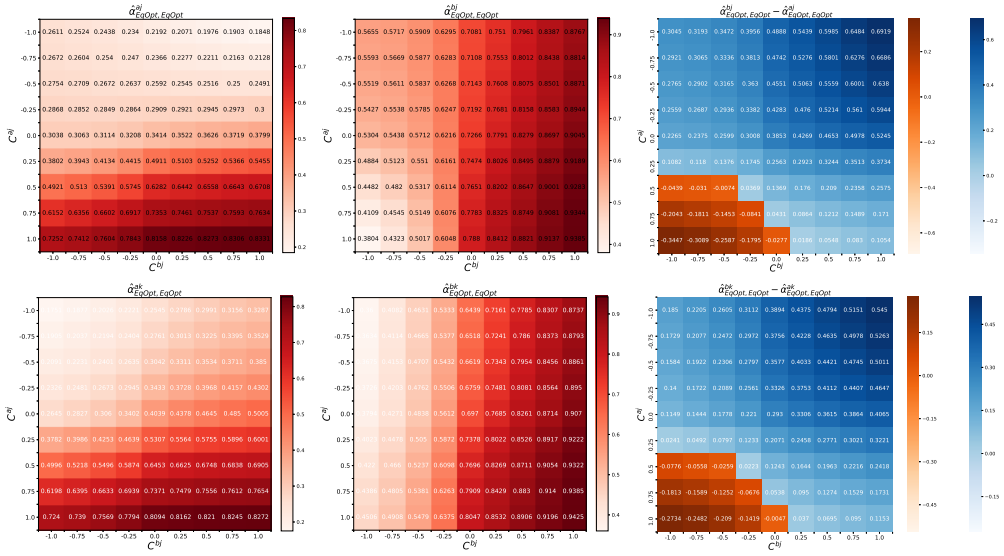

Figure 5.9: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), both π^j and π^k are EqOpt-constrained.

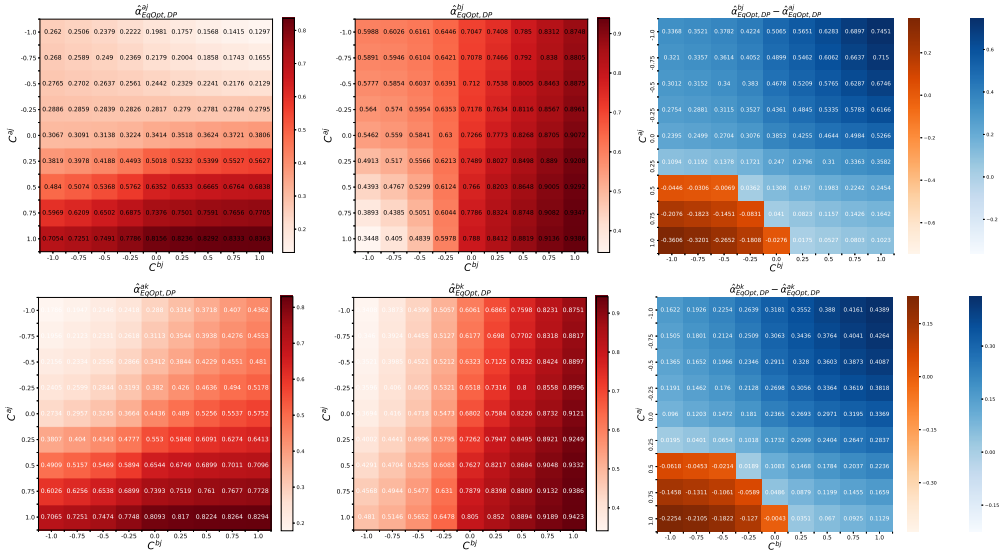

Figure 5.10: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), π^j is EqOpt-constrained and π^k is DP-constrained.

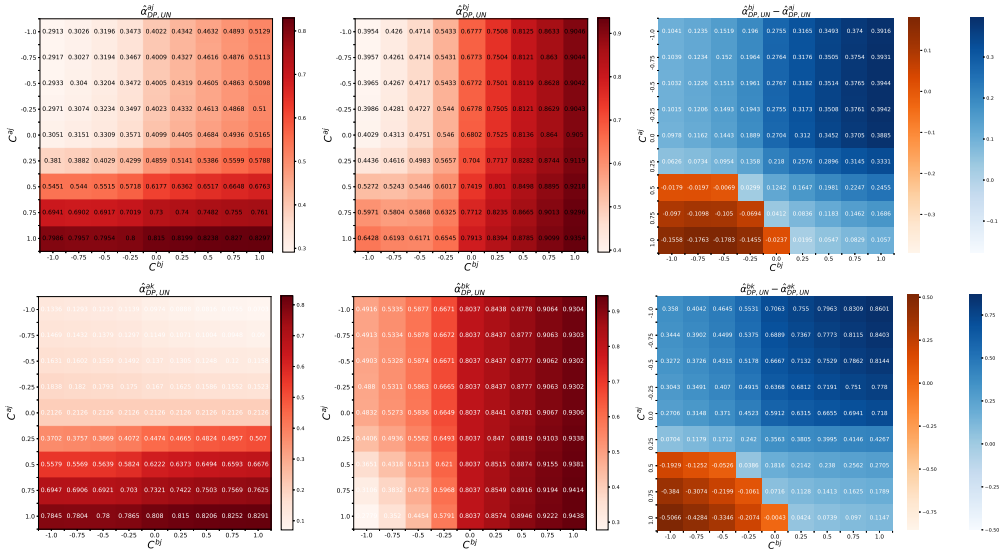

Figure 5.11: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), π^j is DP-constrained and π^k is unconstrained.

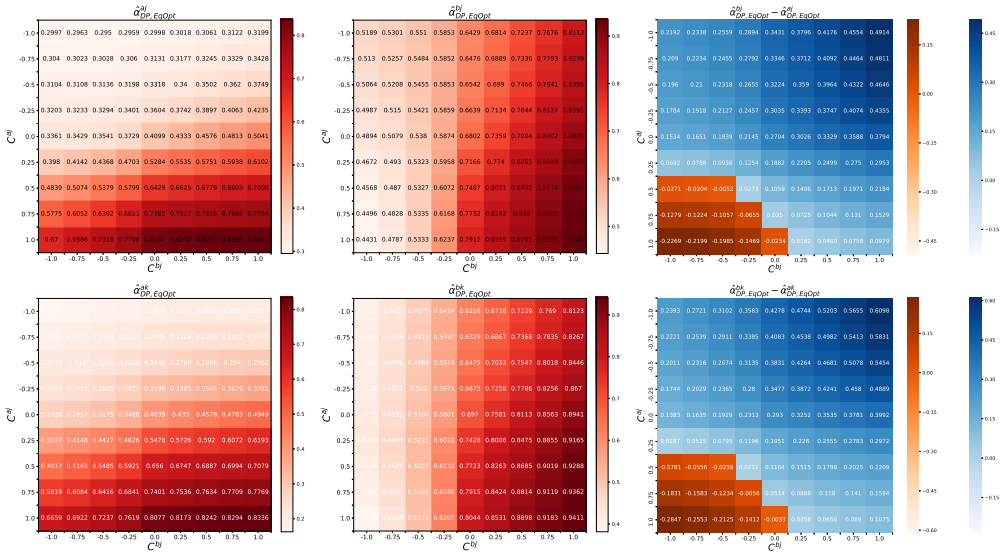

Figure 5.12: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), π^j is DP-constrained and π^k is EqOpt-constrained.

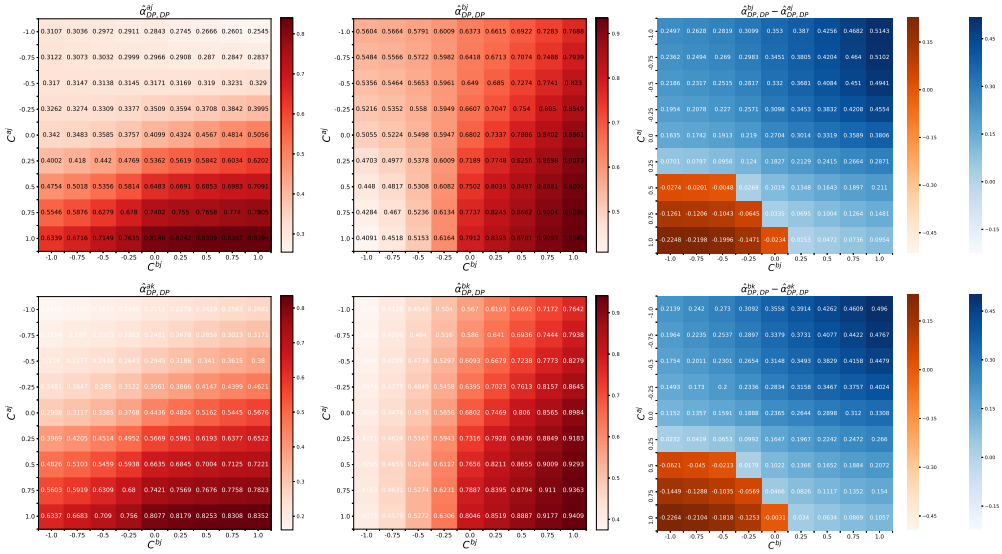

Figure 5.13: Values of fit rates $\hat{\alpha}^{a,j}$, $\hat{\alpha}^{b,j}$, $\hat{\alpha}^{a,k}$, $\hat{\alpha}^{b,k}$ as well as inequalities $\hat{\alpha}^{b,j} - \hat{\alpha}^{a,j}$ and $\hat{\alpha}^{b,k} - \hat{\alpha}^{a,k}$ under different cross impact ($C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0), both π^j and π^k are DP-constrained.

Influence of Fairness Intervention. Next we show the influence of recommendation policy in neighboring item categories.

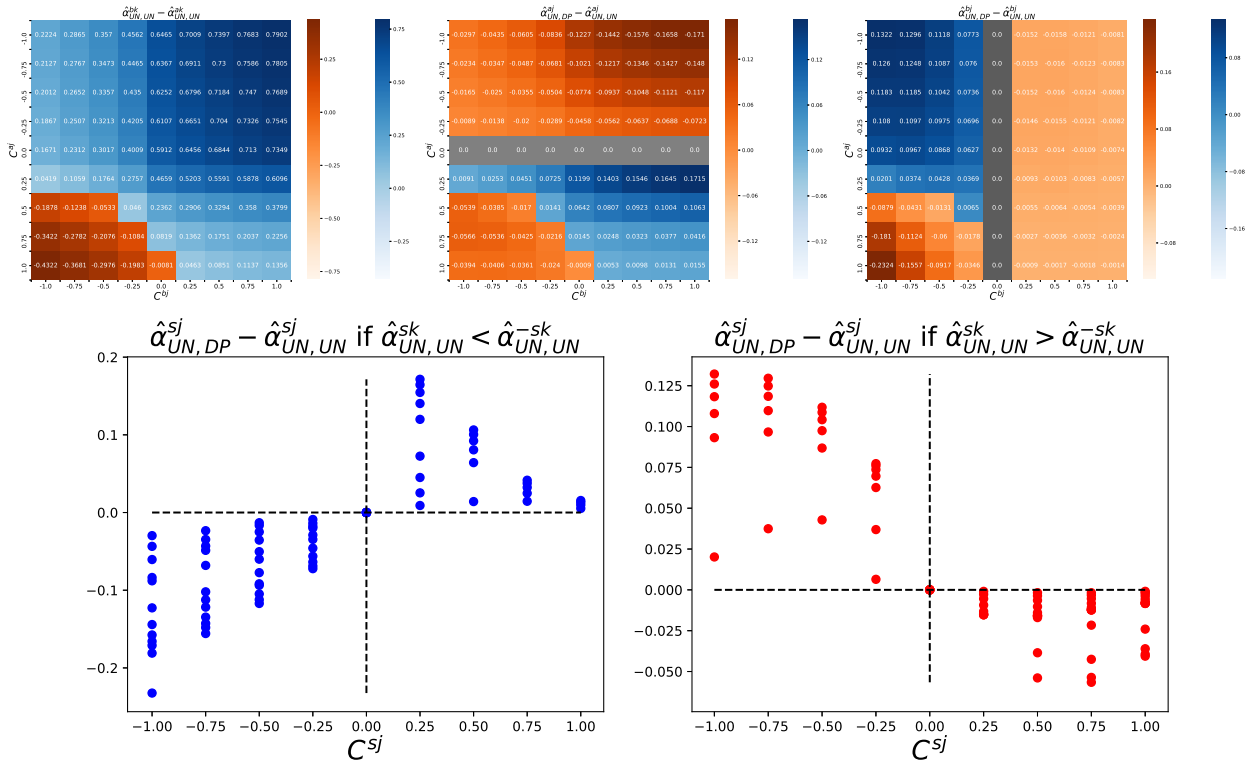

Figure 5.14: The values of $\hat{\alpha}^{a,j}_{UN,DP} - \hat{\alpha}^{a,j}_{UN,UN}$, $\hat{\alpha}^{b,j}_{UN,DP} - \hat{\alpha}^{b,j}_{UN,UN}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to DP-constrained. π^j is unconstrained. $C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0.

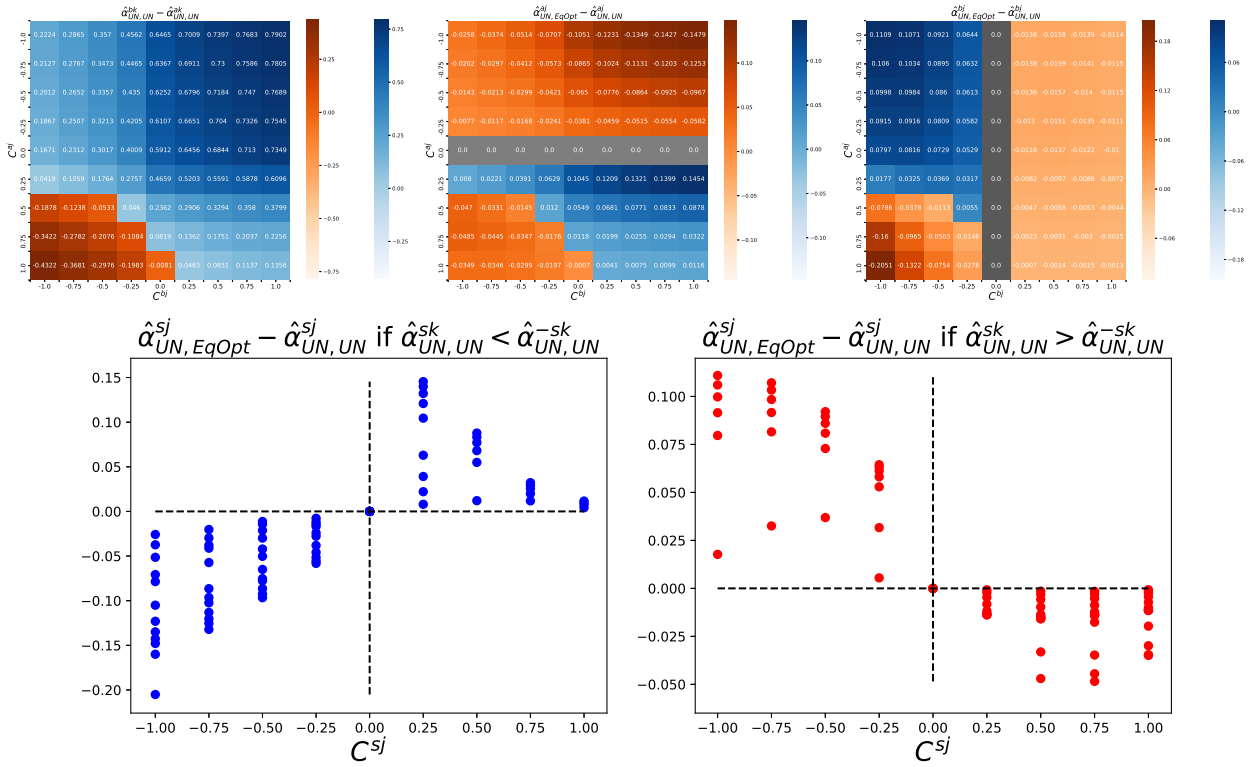

Figure 5.15: The values of $\hat{\alpha}^{a,j}_{UN,EqOpt} - \hat{\alpha}^{a,j}_{UN,UN}$, $\hat{\alpha}^{b,j}_{UN,EqOpt} - \hat{\alpha}^{b,j}_{UN,UN}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to EqOpt-constrained. π^j is unconstrained. $C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0.

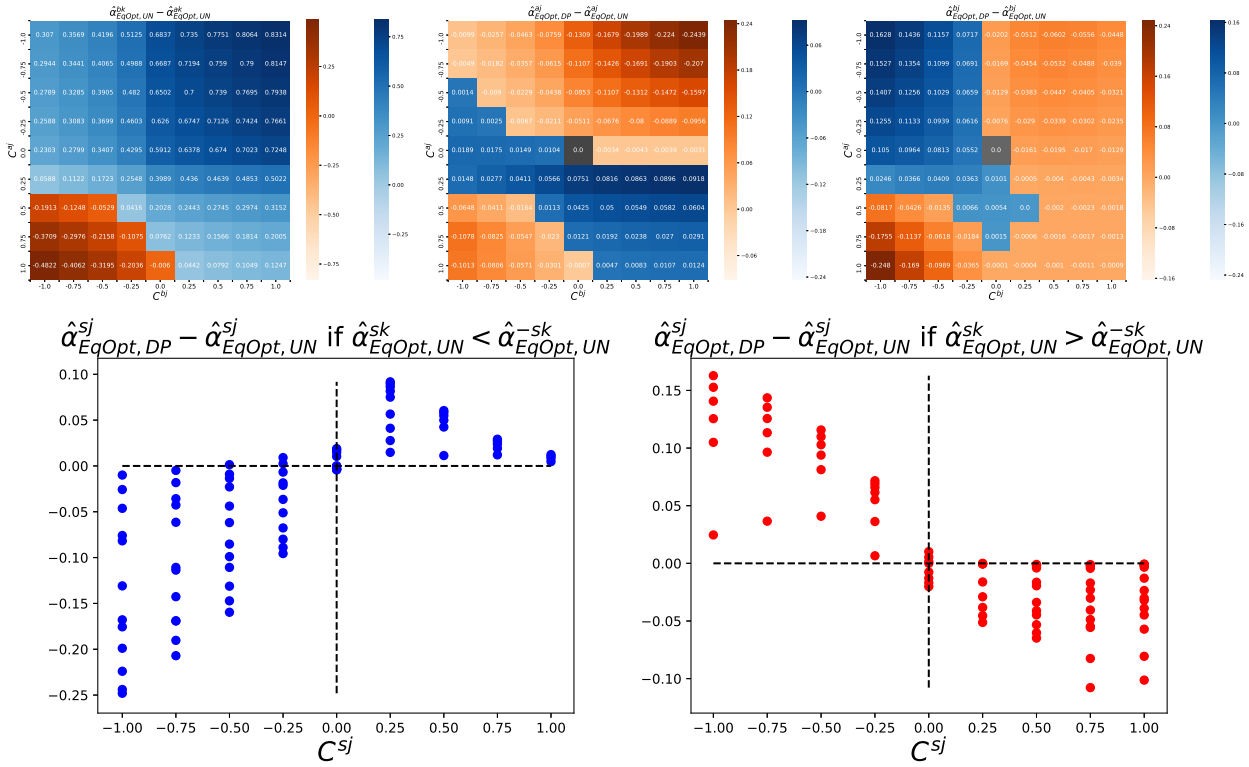

Figure 5.16: The values of $\hat{\alpha}^{a,j}_{EqOpt,DP} - \hat{\alpha}^{a,j}_{EqOpt,UN}$, $\hat{\alpha}^{b,j}_{EqOpt,DP} - \hat{\alpha}^{b,j}_{EqOpt,UN}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to DP-constrained. π^j is EqOpt-constrained. $C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0.

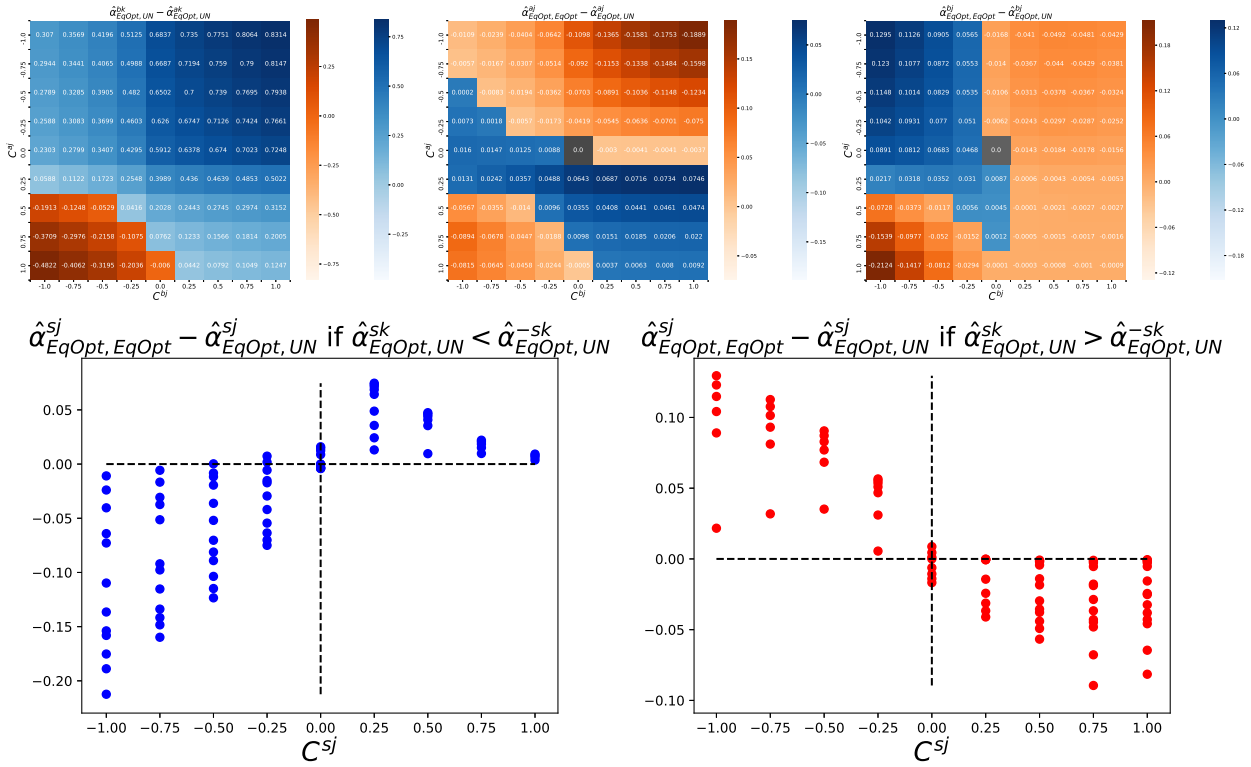

Figure 5.17: The values of $\hat{\alpha}^{a,j}_{UN,EqOpt} - \hat{\alpha}^{a,j}_{UN,UN}$, $\hat{\alpha}^{b,j}_{UN,EqOpt} - \hat{\alpha}^{b,j}_{UN,UN}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to EqOpt-constrained. π^j is EqOpt-constrained. $C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0.

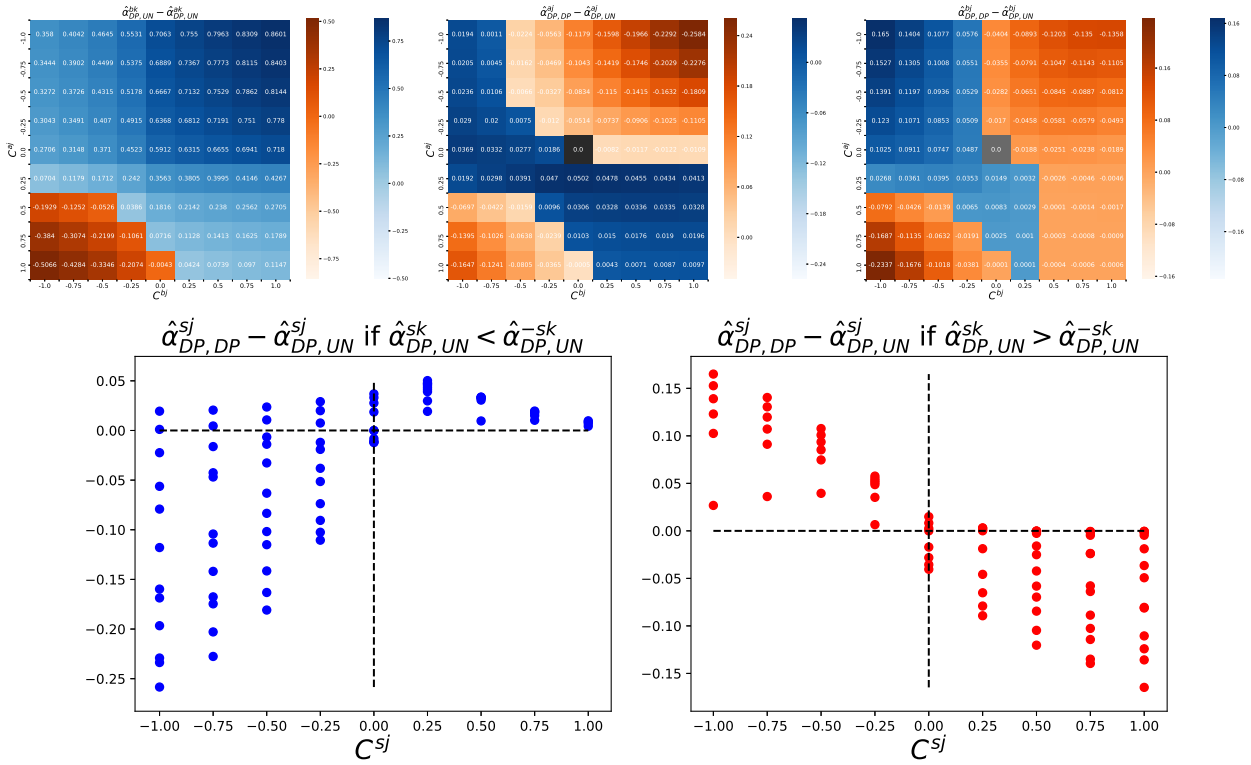

Figure 5.18: The values of $\hat{\alpha}^{a,j}_{DP,DP} - \hat{\alpha}^{a,j}_{DP,UN}$, $\hat{\alpha}^{b,j}_{DP,DP} - \hat{\alpha}^{b,j}_{DP,UN}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to DP-constrained. π^j is DP-constrained. $C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0.

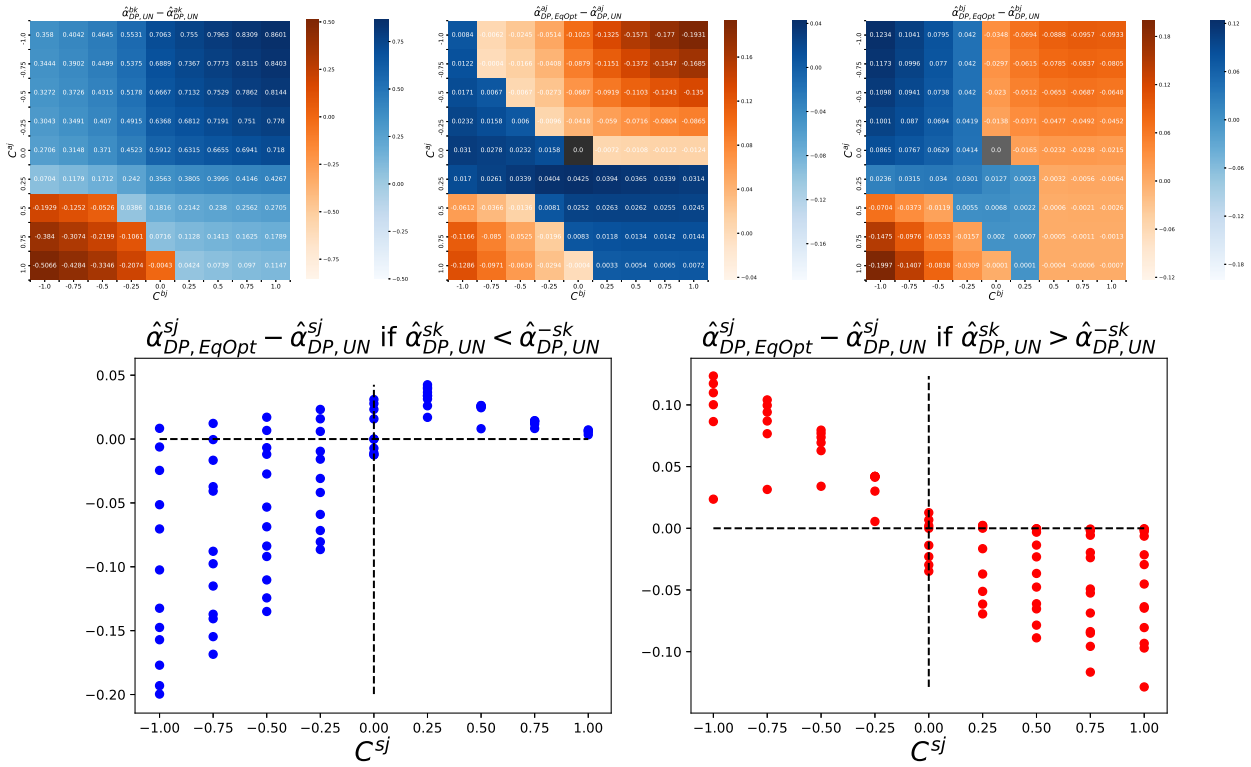

Figure 5.19: The values of $\hat{\alpha}^{a,j}_{DP,EqOpt} - \hat{\alpha}^{a,j}_{DP,UN}$, $\hat{\alpha}^{b,j}_{DP,EqOpt} - \hat{\alpha}^{b,j}_{DP,UN}$ as recommendation policy in $\mathcal{G}_K$ change from unconstrained to EqOpt-constrained. π^j is DP-constrained. $C^{a,j}$ and $C^{b,j}$ vary from -1.0 to 1.0.

Chapter 6

Summary and Conclusion

Recommender systems are machine learning systems that model the relationship between users and items. In this chapter, we summarize our contributions on evaluating, understanding, and mitigating unfairness in recommender systems. We also discuss the limitations of our work and potential directions for future research.

6.1 Contributions

We start with unfairness evaluation and focused on the tasks of rating prediction. We first proposed a set of error-based unfairness metrics that are appropriate for scenarios where the target variable is justifiably dependent on demographic features. Specifically, we measure unfairness as the discrepancy in how much prediction deviates from the true ratings. We further proposed to mitigate these error-based unfairness in matrix factorization models by explicitly adding unfairness penalty terms to the learning objective, so that accuracy and fairness are optimized simultaneously. Experiments on synthetic and real datasets show that this optimization approach is effective in reducing error-based unfairness in matrix factorization models.

Next we focused on the unfairness in the form of disparate prediction error, which means a model has higher prediction error on one group of users than the others. We identified four types of biases in the training data that cause higher subpopulation error in matrix factorization models. Specifically, the biases refer to the differences in the sparsity, rank, level of noise in per-user data, and subpopulation size of user groups. To address these biases, we offered Personalized Regularization Learning (PRL), which learns a set of personalized regularization parameters to replace the global regularization parameter in a vanilla matrix factorization model. The personalized regularization parameters are learned efficiently by back-propagating through the closed-form solutions of alternating least squares. We ran experiments on the benchmark Movielens 100K dataset and split the users based on five

different user attributes. Experiment results showed that PRL was more effective in reducing the highest subpopulation error compared with five baseline models. We also interpreted the learned parameters to understand how PRL handles the data biases.

Last, we turned to the inequality in the fit between users and items. We conducted a theoretical analysis on the long-term dynamics of fit rates in recommender systems. Fit rates measure the level of user-item fit and are computed as the ratio of users from a user group being a fit with an item category. We first mathematically formulate the one-step update rule of fit rates. Then we prove that there always exists at least one system equilibrium if the formulated one-step dynamics repeat. We further provide sufficient conditions for one unique equilibrium. We then focus on the scenarios with a unique equilibrium. We found that the dynamics of fit rates are dependent on the relationship between item categories and the recommendation policies. We validate our theoretical analysis using simulation on synthetic users and items. The results provide valuable insight for anticipating and mitigating the future inequality in the underlying population.

Overall, we unfold the complexity of recommendation unfairness in four dimensions. First, we consider different notions of unfairness for different scenarios depending on whether the disparities in user preferences are justifiable. Second, we study unfairness at different stages of recommendation, from rating prediction (matrix factorization) to downstream classification. Third, we analyze the dependency between a)disparities in the underlying population and b)unfairness in recommender decisions in both directions. Fourth, we zoom out from analyzing one-step decisions to the long-term dynamic of interactions between recommender system and users.

6.2 Limitations and Future Prospect

Now we discuss the limitations of our work. First, in terms of unfairness evaluation, the error-based unfairness metrics have two main limitations. One limitation is that we measure the overall unfairness as the average per-item unfairness. However, due to the long-tail problem in recommendation data where most items (especially the newly-added items) have very few ratings, the per-item fairness estimations are likely to be inaccurate because of limited sampling. Another limitation is that the error-based unfairness metrics are formulated based on a fundamental assumption, that is the average observed ratings is accurate enough to reflect the average true preference of a user group. However, this assumption may be violated since ratings can also be influenced by various environmental factors. For example, in education, a student's rating for a course also depends on whether the course has an inclusive and welcoming learning environment.

Second, we proposed two mitigation methods and they each have their own limitations. We just discussed that the error-based unfairness metrics are vulnerable to sampling bias because they are measured based on per-item data. Since the unfairness regularization method uses

the error-based unfairness metrics as penalty terms, it also runs the risk of optimizing an inaccurate objective when data is sparse. As for PRL, which learns a set of personalized parameters that minimize the generalization error on a validation set. When data is sparse and the number of users is large, the learned personalized regularization parameters are likely to overfit the validation data, leading to poor generalization to unseen data. Furthermore, the two mitigation approaches we proposed both directly interfere the training process. This means they require access to the raw data and the training pipeline, which may not be feasible in some cases. For future research, it is worthwhile to explore pre-processing strategies that directly reduce data biases so that the trained model will be more fair. We can also explore post-processing strategies that modify the output of a trained model with minimum harm to accuracy.

Third, in our analysis on the long-term dynamics of fit between users and items, we summarized the transitions in user-item fit as a set of time-invariant probabilities. However, the transitions in the real world can be much more complicated due to the complexity of human behaviors and environmental factors. For example, the transitions may change over time as the items evolve. Also, our analysis focus only on scenarios where the equilibrium is unique. Therefore, it is important for future research to look deeper into the cases with more complicated transition dynamics and scenarios with multiple equilibria.

Last, throughout this research, we assume that demographic features are available. However, this may not be satisfied in practice because demographic features are usually sensitive information that users are not willing to disclose. A key question to answer in future research is how we still evaluate and mitigate unfairness in recommender systems without these demographic information.

Bibliography

[1] Xavier Amatriain and Justin Basilico. Netflix recommendations: Beyond the 5 stars (part 1). *Netflix Tech Blog*, 6, 2012.

[2] Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 2015.

[3] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[4] Kurt Jacobson, Vidhya Murali, Edward Newett, Brian Whitman, and Romain Yon. Music personalization at spotify. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 373–373, 2016.

[5] Saman Forouzandeh and Atae Rezaei Aghdam. Health recommender system in social networks: A case of facebook. *Webology*, 16(1), 2019.

[6] Luis Fernandez-Luque, Randi Karlsen, and Lars Kristian Vognild. Challenges and opportunities of using recommender systems for personalized health education. In *MIE*, pages 903–907, 2009.

[7] Hendrik Drachsler, Katrien Verbert, Olga C Santos, and Nikos Manouselis. Panorama of recommender systems to support learning. In *Recommender systems handbook*, pages 421–451. Springer, 2015.

[8] Shaha T Al-Otaibi and Mourad Ykhlef. A survey of job recommender systems. *International Journal of the Physical Sciences*, 7(29):5127–5142, 2012.

[9] Zheng Siting, Hong Wenxing, Zhang Ning, and Yang Fan. Job recommender systems: a survey. In *2012 7th International Conference on Computer Science & Education (ICCSE)*, pages 920–924. IEEE, 2012.

[10] Mamadou Diaby, Emmanuel Viennet, and Tristan Launay. Toward the next generation of recruitment tools: an online social network-based job recommender system. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 821–828. IEEE, 2013.

[11] Muhammad Aljukhadar, Sylvain Senecal, and Charles-Etienne Daoust. Using recommendation agents to cope with information overload. *International Journal of Electronic Commerce*, 17(2):41–70, 2012.

[12] Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM conference on recommender systems*, pages 7–10, 2016.

[13] Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4):1–23, 2019.

[14] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NeurIPS Tutorial*, 2017.

[15] Ayman Farahat and Michael C Bailey. How effective is targeted advertising? In *Proceedings of the 21st international conference on World Wide Web*, pages 111–120, 2012.

[16] Robin Burke. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*, 2017.

[17] Bashir Rastegarpanah, Krishna P Gummadi, and Mark Crovella. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 231–239, 2019.

[18] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2212–2220, 2019.

[19] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. Personalized fairness-aware re-ranking for microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 467–471, 2019.

[20] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 429–438, 2020.

[21] Julia, Alexander. LGBTQ YouTubers Are Suing YouTube Over Alleged Discrimination. "`https://www.theverge.com/2019/8/14/20805283/lgbtq-youtuber-lawsuit-discrimination-alleged-video-recommendations-demonetization`'

[22] Alex Beutel, Ed H Chi, Zhiyuan Cheng, Hubert Pham, and John Anderson. Beyond globally optimal: Focused learning for improved recommendations. In *Proceedings of the 26th International Conference on World Wide Web*, pages 203–212, 2017.

[23] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on Fairness, Accountability and Transparency*, pages 202–214, 2018.

[24] Gary Stanley Becker, Walter Block, C.-B.) Fraser Institute (Vancouver, Thomas Sowell, and Kurt Vonnegut. *Discrimination, affirmative action, and equal opportunity*. Citeseer, 1982.

[25] John Hasnas. Equal opportunity, affirmative action, and the anti-discrimination principle: The philosophical basis for the legal prohibition of discrimination. *Fordham L. Rev.*, 71:423, 2002.

[26] Nijole V Benokraitis. *Affirmative action and equal opportunity: Action, inaction, reaction*. Routledge, 2019.

[27] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232, 2018.

[28] Liang Xiang, Quan Yuan, Shiwan Zhao, Li Chen, Xiatian Zhang, Qing Yang, and Jimeng Sun. Temporal recommendation on graphs via long-and short-term preference fusion. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 723–732, 2010.

[29] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

[30] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.

[31] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[32] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[33] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[34] Andriy Mnih and Russ R Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2008.

[35] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[36] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5(Nov):1457–1469, 2004.

[37] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2005.

[38] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[39] H Martin Bücker, George Corliss, Paul Hovland, Uwe Naumann, and Boyana Norris. *Automatic differentiation: applications, theory, and implementations*, volume 50. Springer Science & Business Media, 2006.

[40] Gábor Takács and Domonkos Tikk. Alternating least squares for personalized ranking. In *Proc. of the ACM Conference on Recommender Systems*, pages 83–90, 2012.

[41] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. Ieee, 2008.

[42] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer, 2011.

[43] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[44] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333, 2013.

[45] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Adv. in Neural Information Processing Systems*, pages 3315–3323, 2016.

[46] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[47] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[48] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[49] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[50] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[51] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[52] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[53] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

[54] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

[55] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[56] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*, pages 10976–10987, 2018.

[57] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[58] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Efficiency improvement of neutrality-enhanced recommendation. In *Decisions@ RecSys*, pages 1–8, 2013.

[59] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation independence. In *Conference on Fairness, Accountability and Transparency*, pages 187–201, 2018.

[60] Ziwei Zhu, Xia Hu, and James Caverlee. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1153–1162, 2018.

[61] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 405–414, 2018.

[62] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840*, 2017.

[63] Himan Abdollahpouri and Robin Burke. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158*, 2019.

[64] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[65] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-Based Systems*, 123:154–162, 2017.

[66] Shameem A Puthiya Parambath, Nicolas Usunier, and Yves Grandvalet. A coverage-based approach to recommendation diversity on similarity graph. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 15–22, 2016.

[67] Keith Bradley and Barry Smyth. Improving recommendation diversity. In *Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland*, pages 85–94. Citeseer, 2001.

[68] Yong Zheng. Multi-stakeholder recommendation: Applications and challenges. *arXiv preprint arXiv:1707.08913*, 2017.

[69] Benjamin Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. Collaborative filtering and the missing at random assumption. *arXiv preprint arXiv:1206.5267*, 2012.

[70] Benjamin M Marlin and Richard S Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, pages 5–12. ACM, 2009.

[71] Shaghayegh Sahebi and Peter Brusilovsky. It takes two to tango: An exploration of domain pairs for cross-domain collaborative filtering. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 131–138. ACM, 2015.

[72] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352*, 2016.

[73] Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, and Anupam Datta. Feature-wise bias amplification. *arXiv preprint arXiv:1812.08999*, 2018.

[74] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

[75] Xueru Zhang, Mohammadmahdi Khaliligarekani, Cem Tekin, et al. Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, pages 15243–15252, 2019.

[76] Masoud Mansoury, Himan Abdollahpouri, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. Investigating potential factors associated with gender discrimination in collaborative recommender systems. *arXiv preprint arXiv:2002.07786*, 2020.

[77] Marc Claesen and Bart De Moor. Hyperparameter search in machine learning. *arXiv preprint arXiv:1502.02127*, 2015.

[78] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

[79] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb):281–305, 2012.

[80] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

[81] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122, 2015.

[82] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

[83] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellström, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *arXiv preprint arXiv:2010.11300*, 2020.

[84] Stephen Coate and Glenn C Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

[85] Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368, 2019.

[86] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[87] Xiaoyu Shi, Ming-Sheng Shang, Xin Luo, Abbas Khushnood, and Jian Li. Long-term effects of user preference-oriented recommendation method on the evolution of online system. *Physica A: Statistical Mechanics and its Applications*, 467:490–498, 2017.

[88] Dietmar Jannach, Lukas Lerche, and Michael Jugovac. Adaptation and evaluation of recommendations for short-term shopping goals. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 211–218, 2015.

[89] Robin Devooght and Hugues Bersini. Long and short-term recommendations with recurrent neural networks. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 13–21, 2017.

[90] Lei Li, Li Zheng, and Tao Li. Logo: a long-short user interest integration in personalized news recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 317–320, 2011.

[91] Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1, 1968.

[92] Tobias Schnabel, Paul N Bennett, Susan T Dumais, and Thorsten Joachims. Short-term satisfaction and long-term coverage: Understanding how users tolerate algorithmic exploration. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 513–521, 2018.

[93] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. Crank up the volume: preference bias amplification in collaborative recommendation. *arXiv preprint arXiv:1909.06362*, 2019.

[94] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*, 2018.

[95] Steven Broad and Meredith McGee. Recruiting women into computer science and information systems. *Proceedings of the Association Supporting Computer Users in Education Annual Conference*, pages 29–40, 2014.

[96] Paul W. Holland and Samuel Leinhardt. Local structure in social networks. *Sociological Methodology*, 7:1–45, 1976.

[97] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, pages 73–101, 1964.

[98] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

[99] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930, 2017.

[100] Andreas Griewank and Andrea Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, 2008.

[101] Hung-Hsuan Chen and Pu Chen. Differentiating regularization weights–a simple mechanism to alleviate cold start in recommender systems. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):1–22, 2019.

[102] Winfred Arthur Jr, Suzanne T Bell, Anton J Villado, and Dennis Doverspike. The use of person-organization fit in employment decision making: an assessment of its criterion-related validity. *Journal of applied psychology*, 91(4):786, 2006.

[103] Laurie T O'Brien, Alison Blodorn, Glenn Adams, Donna M Garcia, and Elliott Hammer. Ethnic variation in gender-stem stereotypes and stem participation: An intersectional approach. *Cultural Diversity and Ethnic Minority Psychology*, 21(2):169, 2015.

[104] Sadan Kulturel-Konak, Mary Lou D'Allegro, Sarah Dickinson, et al. Review of gender differences in learning styles: Suggestions for stem education. *Contemporary Issues in Education Research (CIER)*, 4(3):9–18, 2011.

[105] Wei Ma and George H Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. *arXiv preprint arXiv:1910.12774*, 2019.

[106] James N Druckman and Arthur Lupia. Preference change in competitive political environments. *Annual Review of Political Science*, 19:13–31, 2016.

[107] Jerome I Rotgans and Henk G Schmidt. Interest development: Arousing situational interest affects the growth trajectory of individual interest. *Contemporary Educational Psychology*, 49:175–184, 2017.

[108] Ofem U Arikpo and Grace Domike. Pupils learning preferences and interest development in learning. *Journal of Education and Practice*, 6(21):31–38, 2015.

[109] Patricia Pliner. The effects of mere exposure on liking for edible substances. *Appetite*, 3(3):283–290, 1982.

[110] Robert F Bornstein and Paul R D'agostino. Stimulus recognition and the mere exposure effect. *Journal of personality and social psychology*, 63(4):545, 1992.

[111] Nancy Tuana. Re-fusing nature/nurture. In *Women's Studies International Forum*, volume 6, pages 621–632. Pergamon, 1983.

[112] Allison Master, Sapna Cheryan, Adriana Moscatelli, and Andrew N Meltzoff. Programming experience promotes higher stem motivation among first-grade girls. *Journal of experimental child psychology*, 160:92–106, 2017.

[113] Kelly Miller, Gerhard Sonnert, and Philip Sadler. The influence of students' participation in stem competitions on their interest in stem careers. *International Journal of Science Education, Part B*, 8(2):95–114, 2018.

[114] Hosun Kang, Angela Calabrese Barton, Edna Tan, Sandra D Simpkins, Hyang-yon Rhee, and Chandler Turner. How do middle school girls of color develop stem identities? middle school girls' participation in science activities and identification with stem careers. *Science Education*, 103(2):418–439, 2019.

[115] Jennifer D Adams, Preeti Gupta, and Alix Cotumaccio. Long-term participants: A museum program enhances girls' stem interest, motivation, and persistence. *Afterschool Matters*, 20:13–20, 2014.

[116] Laura R Ramsey, Diana E Betz, and Denise Sekaquaptewa. The effects of an academic environment intervention on science identification among women in stem. *Social Psychology of Education*, 16(3):377–397, 2013.

[117] Jiyun Elizabeth L Shin, Sheri R Levy, and Bonita London. Effects of role model exposure on stem and non-stem student engagement. *Journal of Applied Social Psychology*, 46(7):410–427, 2016.

[118] Sarah D Herrmann, Robert Mark Adelman, Jessica E Bodford, Oliver Graudejus, Morris A Okun, and Virginia SY Kwan. The effects of a female role model on academic performance and persistence of women in stem courses. *Basic and Applied Social Psychology*, 38(5):258–268, 2016.

[119] GH Hardy. The shared appeal of science and music: Aesthetic commonalities in two ostensibly dissimilar disciplines.

[120] Yvonne J Vermetten, Hans G Lodewijks, and Jan D Vermunt. Consistency and variability of learning strategies in different university courses. *Higher Education*, 37(1):1–21, 1999.

[121] Eileen Goold. *The role of mathematics in engineering practice and in the formation of engineers*. PhD thesis, National University of Ireland Maynooth, 2012.

[122] Brenda Cantwell Wilson and Sharon Shrock. Contributing to success in an introductory computer science course: a study of twelve factors. *Acm sigcse bulletin*, 33(1):184–188, 2001.

[123] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18457–18469. Curran Associates, Inc., 2020.

[124] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 259–268, 2019.

[125] Behzad Tabibian, Stratis Tsirtsis, Moein Khajehnejad, Adish Singla, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*, 2019.