5-2023

Achieving Causal Fairness in Recommendation

Wen Huang
University of Arkansas-Fayetteville

Achieving Causal Fairness in Recommendation

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Engineering with a concentration in Computer Science

by

Wen Huang
Southeast University
Bachelor of Science in Mathematics, 2017
University of Wisconsin-Madison
Master of Science in Statistics, 2018

May 2023
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Xintao Wu, Ph.D.
Dissertation Director

Susan Gauch, Ph.D.
Committee member

Xiao Liu, Ph.D.
Committee member

Lu Zhang, Ph.D.
Committee member

ABSTRACT

Recommender systems provide personalized services for users seeking information and play an increasingly important role in online applications. While most research papers focus on inventing machine learning algorithms to fit user behavior data and maximizing predictive performance in recommendation, it is also very important to develop fairness-aware machine learning algorithms such that the decisions made by them are not only accurate but also meet desired fairness requirements. In personalized recommendation, although there are many works focusing on fairness and discrimination, how to achieve user-side fairness in bandit recommendation from a causal perspective still remains a challenging task. Besides, the deployed systems utilize user-item interaction data to train models and then generate new data by online recommendation. This feedback loop in recommendation often results in various biases in observational data.

The goal of this dissertation is to address challenging issues in achieving causal fairness in recommender systems: achieving user-side fairness and counterfactual fairness in bandit-based recommendation, mitigating confounding and sample selection bias simultaneously in recommendation and robustly improving bandit learning process with biased offline data. In this dissertation, we developed the following algorithms and frameworks for research problems related to causal fairness in recommendation.

- We developed a contextual bandit algorithm to achieve group level user-side fairness and two UCB-based causal bandit algorithms to achieve counterfactual individual fairness for personalized recommendation;
- We derived sufficient and necessary graphical conditions for identifying and estimating three causal quantities under the presence of confounding and sample selection biases

and proposed a framework for leveraging the causal bound derived from the confounded and selection biased offline data to robustly improve online bandit learning process;

- We developed a framework for discrimination analysis with the benefit of multiple causes of the outcome variable to deal with hidden confounding;

- We proposed a new causal-based fairness notion and developed algorithms for determining whether an individual or a group of individuals is discriminated in terms of equality of effort.

ACKNOWLEDGEMENTS

During my Ph.D. study, I have been fortunate to receive plenty of help and support from many ends. I hereby express my sincere gratitude to all of them.

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Xintao Wu, for all the time, advice, support, and patience he gave me throughout my Ph.D study. He set me an example by his own conduct on how to become an aspiring and outstanding researcher when I was new to the field. His excellency and extraordinary passion for research will continue to motivate me in my future career.

My gratitude also goes to Dr. Lu Zhang and Dr. Yongkai Wu with whom I have been truly fortunate to collaborate throughout my Ph.D. journey. Their advice, experience, and ideas are vital in our research projects. I have benefited tremendously from the collaborations and discussions with both of them.

I would like to thank my dissertation committee members, Dr. Susan Gauch, Dr. Xiao Liu, and Dr. Lu Zhang. Their constructive comments significantly improve the quality of this dissertation.

I also thank my colleagues in the Social Awareness and Intelligent Learning Lab at the University of Arkansas: Hao Van, Aneesh Komanduri, Karuna Bhaila, Huy Mai, Alycia Carey and Vinay Madanbhavi Shashidhar, who give me consistent assist. Thanks also go to the other friends and former members in SAIL lab: Dr. Panpan Zheng, Dr. Depeng Xu, Dr. Shuhan Yuan, Dr. Wei Du, and Dr. Kevin Labille. I have been truly honored to work with these excellent researchers.

Finally, I would express my deepest gratitude to my family for their unconditional love and support. This dissertation is dedicated to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

This chapter introduces the motivation and provides an overview of this dissertation, and then summarizes the contributions of this research.

1.1 Motivation

Machine learning has been widely used to solve challenging problems in both academia and industry communities due to its powerful and effective abilities. Numerous machine learning algorithms have been designed and deployed to make decisions in a variety of real-world applications. The rapid development of these techniques has benefited many AI-related tasks, including computer vision, natural language processing, and recommender systems, etc.

Among those rapidly developing tasks, recommender systems serve as important and valuable tools for many Web-based services such as online advertising, social software and digital media systems. The research literature in this field has grown tremendously in recent years. However, most of the papers focus on inventing machine learning models to maximize predictive performance, e.g., estimating click through ratio based on the historical training data. Since user behavior data is observational rather than experimental, blindly fitting the data without considering the inherent biases will result in many serious issues, e.g., discrimination and unfairness caused by these machine learning algorithms may have serious consequences for minority groups, as well as perpetuate and exacerbate existing prejudices and social inequalities.

Fairness-aware machine learning is receiving an increasing attention in machine learn-

ing fields. Discrimination is unfair treatment towards individuals based on the group to which they are perceived to belong. The first endeavor of the research community to achieve fairness is developing correlation or association-based measures, including demographic disparity (e.g., risk difference), mistreatment disparity, calibration, etc. [1, 2, 3, 4, 5], which mainly focus on discovering the disparity of certain statistical metrics between two groups of individuals. However, as paid increasing attention recently [6, 7, 8, 9, 10, 11, 12, 13, 14], unlawful discrimination is a causal connection between the challenged decision and a protected characteristic, which cannot be captured by simple correlation or association concepts. To address this limitation, causal-based fairness measures have been proposed, including total effect [15], direct and indirect discrimination [6, 15, 16], counterfactual fairness [17, 18, 9], and path-specific counterfactual fairness [19]. Besides, how to strike a balance between accurate predictions and fairness is receiving increasing attention in the machine learning field. Causal modeling based fair learning models [17, 18, 6, 19, 20, 9, 16], which are based on Pearl's (probabilistic) causal model [21], have been developed to capture and quantify different fairness measures through counterfactual inference along specific paths in causal graphs.

Recently researchers have started taking fairness and discrimination into consideration in the design of recommendation algorithms. It is known that many existing recommendation algorithms are designed solely based on learning correlative patterns from observational data, and could incur biases, even discrimination that can influence recommendation performance and ethical treatment of customers with different profile attributes. To transform the large volume of research models into practical improvements, it is highly urgent to explore the impacts of various biases and conduct debiasing procedure when necessary. It is also imperative to develop explainable, trustworthy, and fairness-aware algorithm frameworks in recommendation with the assistance of causal inference techniques, such that the deci-

sions made by those algorithms are able to achieve fairness and high predictive performance simultaneously.

1.2 Overview

The goal of this dissertation is to address challenging issues in achieving causal fairness in recommendation.

First, we focus on user-side fairness in bandit-based recommendation. Personalized recommendation based on multi-arm bandit (MAB) algorithms has become a popular topic of research and shown to lead to high utility and efficiency [22] as it dynamically adapts the recommendation strategy based on feedback. However, it is also known that such personalization could incur biases or even discrimination that can influence decisions and opinions [23, 24]. Recently researchers have started taking fairness and discrimination into consideration in the design of MAB based personalized recommendation algorithms [25, 26, 27]. However, they focused on the fairness of the recommended items (e.g., services provided by small or large companies) instead of the customers who received those items. For example, [26] focused on individual fairness, i.e., "treating similar individuals similarly," and considered the individual as an arm with the aim of ensuring the probability of selecting an arm is equal to the probability with which the arm has the best quality realization. [25] aimed to achieve group fairness over items by ensuring the probability distribution from which items are sampled satisfies certain fairness constraints at all time steps. In this dissertation, we aim to develop novel algorithms to ensure fair and ethical treatment of customers with different profile attributes (e.g., gender, race, education, disability, and economic conditions) in a contextual bandit based personalized recommendation.

Second, we focus on how to achieve counterfactual fairness in causal bandit. Partic-

ularly, we focus on online recommendation, e.g., customers are being recommended items, and consider the setting where customers arrive in a sequential and stochastic manner from an underlying distribution and the online decision model recommends a chosen item for each arriving individual based on some strategy. The challenge here is how to choose the arm at each step to maximize the expected reward while achieving user-side fairness for customers, i.e., customers who share similar profiles will receive similar rewards regardless of their sensitive attributes and items being recommended. By incorporating causal inference into bandits and adopting soft intervention to model the arm selection strategy, we first propose the d-separation based UCB algorithm (D-UCB) to explore the utilization of the d-separation set in reducing the amount of exploration needed to achieve low cumulative regret. Based on that, we then propose the fair causal bandit (F-UCB) for achieving the counterfactual individual fairness.

Third, we focus on dealing with compound biases in recommender systems. Recommender systems provide personalized services for users seeking information and play an increasingly important role in online applications. However, the user-item interaction data, which are used to train recommender systems and then generated by the deployed systems, often have both selection and confounding biases. The confounding bias arises when hidden variables determine user/item features and an outcome variable simultaneously. For example, popularity bias is one classic instance of confounding bias. It occurs when items are over-displayed and therefore have more chances to be seen as well as clicked by users. Under popularity bias, the click through rate (CTR) of the users does not accurately reflect the users' true preference on an over-exposed item. Additionally, the selection mechanism, e.g., choosing users based on a certain time or location, can lead to sample selection bias. Several attempts have been made to alleviate such biases from both causal and counterfactual

inference perspectives [28, 29, 30]. However, previous work has focused on dealing with one specific source of bias rather than handling multiple sources simultaneously. Neglecting the presence of both confounding and sample selection biases leads to poor recommendation performance. In this dissertation, we formulate both confounding and selection biases and show that they can be separately mitigated by conditioning on a bias adjustment set that satisfies certain criteria. We further investigate how to robustly improve online bandit algorithms with confounded and selection biased offline data. We derive a unified framework that improves the arm-picking strategies of bandit algorithms and achieve lower regret with the help of prior causal bounds extracted from the biased observational data.

Last but not least, we move to fairness-aware machine learning in general recommendation settings and include two extensions. One is to derive a multi-cause discrimination analysis framework under the presence of multiple protected and redlining attributes. The other aims to achieve fairness regarding to equality of effort.

The remainder of this dissertation is organized as follows. In Chapter 2, we discuss related work in a wide scope of fairness-aware machine learning, bandit-based recommendation, as well as causal-based debiasing methods in recommendation. Then in Chapter 3 we clarify some notations that are used through all the proposed research and present background knowledge for causal inference and bandit-based recommendation. The main body of this dissertation is in Chapters 4 - 9. Finally, we conclude this dissertation and discuss future work in Chapter 10.

1.3 Summary of Contributions

In Chapter 4, we propose a fair contextual bandit algorithm for personalized recommendation. While current research in fair recommendation mainly focus on how to achieve

fairness on the items that are being recommended, our work differs by focusing on fairness on the individuals whom are being recommended an item. Specifically, we aim to recommend items to users while insuring that both the protected group and privileged group improve their learning performance equally. Our developed Fair-LinUCB improves upon the state-of-the-art LinUCB algorithm by automatically detecting unfairness, and adjusting its arm-picking strategy such that it maximizes the fairness outcome. We further provide a regret analysis of our fair contextual bandit algorithm and demonstrate that the regret bound is only worse than LinUCB up to an additive constant. Finally, we evaluate the performances of our Fair-LinUCB against that of LinUCB by comparing both their effectiveness and degree of fairness. Experimental evaluations show that our Fair-LinUCB achieves competitive effectiveness while outperforming LinUCB in terms of fairness. We further show that our algorithm is robust against numerous factors that would otherwise induce or increase discrimination in the traditional LinUCB algorithm.

In Chapter 5, we study how to learn optimal interventions sequentially by incorporating causal inference in bandits. We develop D-UCB and F-UCB algorithms which leverage the d-separation set identified from the underlying causal graph and adopt soft intervention to model the arm selection strategy. Our F-UCB further achieves counterfactual individual fairness in each round of exploration by choosing arms from a subset of arms satisfying counterfactual fairness constraint. Our theoretical analysis and empirical evaluation show the effectiveness of our algorithms against baselines.

In Chapter 6, we study both confounding and sample selection biases in recommendation systems and develop a causal based debiased recommendation algorithm that simultaneously controls for confounding and selection biases via some auxiliary external data. We present sufficient and necessary graphical conditions for conditional causal effects, path-

specific effects, and counterfactual effects. We also derive a procedure to estimate an adjustment under confounding and selection biases based on the inverse probability weighting technique. Our empirical evaluation shows the effectiveness of our approach.

In Chapter 7, we aim to extract the causal bound for each arm that is robust towards compound biases from biased observational data. The derived bounds contain the ground truth mean reward and can effectively guide the bandit agent to learn a nearly-optimal decision policy. We also conduct regret analysis in both contextual and non-contextual bandit settings and showed that prior causal bounds could help consistently reduce the asymptotic regret.

In Chapter 8, we develop one approach based on the potential outcome framework to analyze the discrimination effects of protected and redlining attributes on the decision. The developed approach is based on the potential outcome framework and combines the deconfounder and inverse probability of treatment weighting. It can better handle the presence of hidden confounders and can lead to a more robust estimate of causal effects. We empirically compare our approach with the structural causal modeling based approach and experimental results demonstrate the advantages of the proposed approach.

In Chapter 9, we propose a new causality-based fairness notion called the equality of effort. Although previous notions can be used to judge discrimination from various perspectives (e.g., demographic parity, equal opportunity), they cannot quantify the (difference in) efforts that individuals need to make in order to achieve certain outcome levels. Our proposed notion, on the other hand, can help answer counterfactual questions like "how much credit score an applicant should improve such that the probability of her loan application approval is above a threshold", and judge discrimination from the equal-effort perspective. To quantify the average effort discrepancy, we develop a general method under certain assumptions and

specific methods based on three common causal inference techniques. When equality of effort is not achieved in a dataset, we develop an optimization method to remove discrimination. In the experiments, we show that the Adult dataset does contain effort discrepancy at system, group, and also individual levels, and our removal method can ensure the newly generated dataset satisfies equality of effort.

2 Related Work

2.1 Fairness-aware Machine Learning

Fairness in machine learning has been a research subject with rapid growth and attention recently. In machine learning, training data may have historically biased decisions against the protected group; models learned from such data may make discriminatory predictions against the protected group. The fair learning research community has developed extensive fair machine learning algorithms based on a variety of fairness metrics, e.g., equality of opportunity and equalized odds [31, 32], direct and indirect discrimination [6, 15, 16], counterfactual fairness [17, 18, 9], and path-specific counterfactual fairness [19]. There are survey papers that comprehensively and systematically studied various categories of statistical fairness [33] and causality-based fairness [34] metrics.

Among those publications, related but different from our work include long term fairness (e.g., [35]), which concerns for how decisions affect the long-term well-being of disadvantaged groups measured in terms of a temporal variable of interest, fair pipeline or multi-stage learning (e.g., [36, 37, 38, 39]), which primarily consider the combination of multiple non-adaptive sequential decisions and evaluate fairness at the end of the pipeline, and fair sequential learning (e.g., [40]), which sequentially considers each individual and makes decision for them. In [35], the authors proposed the study of delayed impact of fair machine learning and introduced a one-step feedback model of decision-making to quantify the long-term impact of classification on different groups in the population.

2.2 Causal Inference and Bias in Recommendation

Recently, causal inference has emerged as powerful tool for dealing with biases in recommendation. [41] presented a systematic survey on biases in recommendation systems and categorize them into selection, conformity, exposure, position, inductive, popularity, and unfairness bias. Existing methods on causal inference based recommendation debiasing usually focus on addressing only one particular bias, such as position bias [42] or popularity bias [28]. The majority of existing work [43, 28, 29] employs extracted user/item feature embeddings and focus on predefined abstract causal structures. In these approaches, recommender models estimate the conditional probability of clicks given user/item representations that are derived from logged user-item interaction data. An abstract causal graph is constructed to analyze the causal relations among user representations, item representations, and prediction scores. To address hidden confounders, e.g., item popularity, that affect both the user representation and the prediction score, they developed approximations of backdoor adjustment to eliminate the impact of confounders. Additionally, [28] developed a popularity-bias deconfounding and adjusting method via causal intervention. [29] examined the causal effect of user representation on the prediction scores and develop a deconfounded recommender system (DecRS) to prevent bias amplification.

When considering both confounding and sample selection biases in recommendation systems, [44] was the first to study the use of adjustment for simultaneously dealing with both confounding and selection biases based on the SCM. They introduced the selection-backdoor criterion as a sufficient condition for recovering causal effects from a biased distribution and externally unbiased data. Correa et al. [45] developed a set of complete conditions for two cases: when none of the covariates are measured externally, and when all of them are mea-

sured without selection bias. [46] further studied a general case when only a subset of the covariates require external measurement. They introduced the notion of an adjustment pair, present graphical conditions for identifying causal effects by adjustment, design an algorithm for finding all admissible adjustment pairs, and develop an estimation procedure. The developed adjustment technique combines the partial unbiased data with the biased data to produce an estimand of the causal effect in the overall population. Different from these works that focus on conventional causal effects, our work derives an adjustment criterion and procedure for conditional causal effects, which is needed in personalized causal recommendation. Moreover, we derive results for path-specific causal effects and counterfactual effects, which are important in recommendation analysis.

2.3 Bandit-based Recommendation

Personalized recommendation based on multi-arm bandit (MAB) algorithms has become a popular topic of research and shown to lead to high utility and efficiency [47] as it dynamically adapts the recommendation strategy based on feedback. Contextual bandit [48] is an extension of the classic multi-armed bandit (MAB) algorithm [49]. The MAB chooses an action from a fixed set of choices to maximize the expected gain where each choice's properties are only partially known at the time of choice and the gain of a choice will be observed only after the action is taken. In other words, the MAB simultaneously attempts to acquire new information (exploration) and optimize decisions based on existing knowledge (exploitation). Compared to the traditional content-based recommendation approaches, the MAB is able to fit dynamically changing user preferences over time and address the cold-start problem by balancing the exploration and exploitation trade-off in the recommendation system. However, the MAB does not use any information about the state of the environment.

The contextual bandit model extends the MAB model by making the recommendation conditional on the state of the environment. Other variations include stochastic [50], Bayesian [51], adversarial [52], and non-stationary [53] bandits. The contextual information is the customer's features and the features of the items under exploration, and the reward is derived from purchase record or customer feedback.

Recently researchers have also started taking fairness and discrimination into consideration in the design of MAB based personalized recommendation algorithms [40, 54, 55, 25, 26, 27, 56, 57, 58]. Among them, [40] was the first paper of studying fairness in classic and contextual bandits. It defined fairness with respect to one-step rewards introduced a notion of meritocratic fairness, i.e., the algorithm should never place higher selection probability on a less qualified arm (e.g., job applicant) than on a more qualified arm. This was inspired by equal treatment, i.e., similar people should be treated similarly. [59] developed a metric-free individual fairness and a cooperative contextual bandits (CCB) algorithm. The CCB algorithm utilizes fairness as a reward and attempts to maximize it. It tries to achieve individual fairness unlimited to problem-specific similarity metrics using multiple gradient contextual bandits. The following works along this direction include [54] for infinite and contextual bandits, [55] for reinforcement learning, [26] for the simple stochastic bandit setting with calibration based fairness. In [60], the authors studied the problem of learning fair stochastic multi-armed bandit where each arm is required to be pulled for at least a given fraction of the total available rounds. In [61], the authors studied fairness in the setting that multiple arms can be simultaneously played and an arm could sometimes be sleeping. [62] used an unknown Mahalanobis similarity metric from some weak feedback that identifies fairness violations through an oracle rather than adopting a quantitative fairness metric over individuals. The fairness constraint requires that the difference between the probabili-

ties that any two actions are taken is bounded by the distance between their contexts. All the above papers require some fairness constraint on arms at every round of the learning process, which is different from our user-side fairness setting. How to achieve fairness in other related contexts have also been studied, e.g., sequential decision making [63], online stochastic classification [64], offline contextual bandits [65], and collaborative filtering based recommendation systems [66, 67].

There are also several state-of-the-art research works that focus on confounding issue in bandit setting [68, 69]. It is shown in [68] that in MAB problems, neglecting unobserved confounders will lead to a sub-optimal arm selection strategy. They also demonstrated that one cannot simulate the optimal arm-picking strategy by a single data collection procedure, such as pure offline or online evaluation. To this end, another line of research works considers combining offline causal inference techniques and online bandit learning to approximate a nearly-optimal policy. [69] studied a linear bandit problem where the agent is provided with partially observed offline data. [70, 71] derived causal bounds based on structural causal model and used them to guide arm selection in online bandit algorithms. [72] further leveraged the information provided by the lower bound of the mean reward to reduce the cumulative regret. Nevertheless, none of the bounds derived by these methods are based on a feature-level causal graph extracted from the offline data. [73, 74] proposed another direction to unify offline causal inference and online bandit learning by extracting appropriate logged data and feed it to online learning phase. Their VirUCB-based framework mitigates the cold start problem and can thus boost the learning speed for a bandit algorithm without any cost on the regret.

3 Preliminaries

In this chapter, we present the essential notations and fundamental background for the whole dissertation. We start with the notations of describing data, variables and distributions. Then we continue with the necessary background knowledge on causal inference and bandit-based Recommendation.

3.1 Notations

Throughout the dissertation, an uppercase denotes a variable, e.g., S; a bold uppercase denotes a set of variables, e.g., $\mathbf{X}$; a lowercase denotes a value or a set of values of the variables, e.g., s and $\mathbf{x}$; and a lowercase with superscript denotes a particular value, e.g., s^+ and x^-. We use $||\mathbf{x}||_2$ to define the L-2 norm of a vector $\mathbf{x} \in \mathbb{R}^d$. For a positive definite matrix $A \in \mathbb{R}^{d \times d}$, we define the weighted 2-norm of $\mathbf{x} \in \mathbb{R}^d$ to be $||\mathbf{x}||_A = \sqrt{\mathbf{x}^{\mathrm{T}} A \mathbf{x}}$.

3.2 Causal Inference

3.2.1 Potential Outcomes Framework

The potential outcomes framework, also known as Neyman-Rubin potential outcomes or Rubin causal model, has been widely used in many research areas to perform causal inference. It refers to the outcomes one would see under each treatment option. Let Y be the outcome variable, T be the binary or multiple valued ordinal treatment variable, and $\mathbf{X}$ be the pre-treatment variables (covariates). $Y_i(t)$ represents the potential outcome for individual i given treatment level $T = t$ and $\mathbb{E}[Y_i(t)]$ denotes the individual-level expectation of

outcome variable. The "fundamental problem of causal inference" claims that one can never observe all the potential outcomes for any individual [75] and we need to compare potential outcomes and make inference from observed data. We use $\mathbb{E}[Y(t)]$ to denote population-level expectation of outcome variable and $\mathbb{E}[Y_\diamond(t)]$ to denote the conditional expectation of outcome variable within certain sub-population $\diamond$.

Classic causal inference focuses on estimating the potential outcome and treatment effect given the information of treatment variable and pre-treatment variables [76]. For example, the average treatment effect $ATE = \mathbb{E}[Y(t') - Y(t)]$ answers the question of how, on average, the outcome of interest Y would change if everyone in the population of interest had been assigned to a particular treatment t' relative to if they had received another treatment t. The average treatment effect on the treated, $ATT = \mathbb{E}[Y(t') - Y(t)|T = t]$ is about how the average outcome would change if everyone who received one particular treatment t had instead received another treatment t'.

The potential outcome framework relies on three assumptions: (1) Stable Unit Treatment Value Assumption (SUTVA) which basically requires the potential outcome observation on one unit should be unaffected by the particular assignment of treatments to the other units. (2) Consistency assumption which means that the value of potential outcomes would not change no matter how the treatment is observed or assigned through an intervention. (3) Strong ignorability (unconfoundedness) assumption which is equal to the assumption that there are no unobserved confounders. A confounder is a pre-treatment variable that affects both treatment and outcome variables.

Propensity Score Method

Definition 1 (Propensity Score)**.** For a binary treatment variable, propensity score is the

conditional probability of receiving treatment T given the pre-treatment variables $\mathbf{X}$,

$$e(\mathbf{x}) = Pr(T = 1|\mathbf{X} = \mathbf{x})$$

The estimation of propensity scores requires the model or functional form of $e(\cdot)$ and the variables to include in $\mathbf{X}$. Let $e(i)$ denote the propensity score for individual i, for binary valued groups, the propensity score is estimated by logistic regression:

$$logit(e(i)) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

where $x_1, ..., x_k$ are values of the selected covariates and $\beta_1, ..., \beta_k$ are regression coefficients. If correctly estimated, the reciprocal of propensity score can be used as the weight for each individual such that the distribution of the group under treatment 1 and that under treatment 0 becomes identical. [77] showed that conditional on the propensity score, all observed covariates are independent of treatment assignment, and they will not confound estimated treatment effects.

Hence after weighting procedure, a pseudo-balanced population can be built in which the imbalance caused by measured covariates between the treatment groups has been eliminated. The average potential outcome can thus be estimated by some standard estimators. For example, one unbiased estimator of the population-level ATE can be written as: $\frac{1}{N_1}\sum_{i\in N} \mathbb{1}_{T_i=1}\omega_i y_i - \frac{1}{N_2}\sum_{i\in N} \mathbb{1}_{T_i=0}\omega_i y_i$ where $N_1 = \sum_{i\in N} \mathbb{1}_{T_i=1}$ and $N_2 = \sum_{i\in N} \mathbb{1}_{T_i=0}$.

3.2.2 Structural Causal Model

Definition 2 (Structural Causal Model [21])**.** A structural causal model $\mathcal{M}$ is represented by a quadruple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where

1. $\mathbf{U}$ is a set of exogenous (external) variables that are determined by factors outside the model.

2. $P(\mathbf{U})$ is a joint probability distribution defined over $\mathbf{U}$.

3. $\mathbf{V}$ is a set of endogenous (internal) variables that are determined by variables in $\mathbf{U} \cup \mathbf{V}$.

4. $\mathbf{F}$ is a set of structural equations from $\mathbf{U} \cup \mathbf{V}$ to $\mathbf{V}$. Specifically, for each $V \in \mathbf{V}$, there is a function $f_V \in \mathbf{F}$ mapping from $\mathbf{U} \cup (\mathbf{V} \backslash V)$ to V, i.e., $v = f_V(Pa(V), u_V)$, where $Pa(V)$ is a realization of a set of endogenous variables $Pa(V) \in \mathbf{V} \setminus V$ that directly determines V, and u_V is a realization of a set of exogenous variables that directly determines V.

The causal model $\mathcal{M}$ is associated with a causal graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges. Each node of $\mathcal{V}$ corresponds to a variable of $\mathbf{V}$ in $\mathcal{M}$. Each edge in $\mathcal{E}$, denoted by a directed arrow $\rightarrow$, points from a node $X \in \mathbf{U} \cup \mathbf{V}$ to a different node $Y \in \mathbf{V}$ if f_Y uses values of X as input. A *causal path* from X to Y is a directed path that traces arrows directed from X to Y. For a node X, its parents, ancestors, children, and descendants are denoted by $Pa(X)$, $An(X)$, $Ch(X)$, and $De(X)$, respectively. $\mathcal{G}_{\overline{\mathbf{X}}}$ is the graph resulting from removing all incoming edges to $\mathbf{X}$ in $\mathcal{G}$, and $\mathcal{G}_{\underline{\mathbf{X}}}$ is the graph resulting from removing all outgoing edges from $\mathbf{X}$.

Quantitatively measuring causal effects in a causal model is facilitated with the *do*-operator [21] which forces some variable X to take on a certain value x, which can be formally denoted by $do(X = x)$ or $do(x)$. In a causal model $\mathcal{M}$, the intervention $do(x)$ is defined as the substitution of the structural equation $X = f_X(Pa(X), U_X)$ with $X = x$, which corresponds to a modified causal graph that has removed all edges into X and in turn sets X to x. For an observed variable Y affected by the intervention, its interventional variant is denoted by

Y_x. The distribution of Y_x, also referred to as the post-intervention distribution of Y under $do(x)$, is denoted by $P(Y_x = y)$, or simply $P(y_x)$.

Similarly, the intervention that sets the value of a set of variables $\mathbf{X}$ to $\mathbf{x}$ is denoted by $do(\mathbf{X} = \mathbf{x})$. The post-intervention distribution of all other attributes $\mathbf{Y} = \mathbf{V} \backslash \mathbf{X}$, i.e., $P(\mathbf{Y} = \mathbf{y} | do(\mathbf{X} = \mathbf{x}))$, or simply $P(\mathbf{y}|do(\mathbf{x}))$, can be computed by the truncated factorization formula [21],

$$P(\mathbf{y}|do(\mathbf{x})) = \prod_{Y \in \mathbf{Y}} P(y|Pa(Y))\delta_{\mathbf{X}=\mathbf{x}}, \tag{3.1}$$

where $\delta_{\mathbf{X}=\mathbf{x}}$ means assigning attributes in $\mathbf{X}$ involved in the term ahead with the corresponding values in $\mathbf{x}$.

Each causal model $\mathcal{M}$ is associated with a causal graph $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{V}$ is a set of nodes and $\mathbf{E}$ is a set of directed edges. Each node in $\mathcal{G}$ corresponds to a variable V in $\mathcal{M}$. Each edge, denoted by an arrow $\rightarrow$, points from each member of $Pa(V)$ toward V to represent the direct causal relationship specified by equation $f_V(\cdot)$. The well-known d-separation criterion [78] connects the causal graph with conditional independence.

Definition 3 (d-Separation [78])**.** Consider a causal graph $\mathcal{G}$. $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{W}$ are disjoint sets of attributes. $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{W}$ in $\mathcal{G}$, if and only if $\mathbf{W}$ blocks all paths from every node in $\mathbf{X}$ to every node in $\mathbf{Y}$. A path p is said to be blocked by $\mathbf{W}$ if and only if: 1) p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that the middle node m is in $\mathbf{W}$, or 2) p contains an collider $i \rightarrow m \leftarrow j$ such that the middle node m is not in $\mathbf{W}$ and no descendant of m is in $\mathbf{W}$.

3.2.3 Causal Inference under Confounding and Selection Biases

Confounding Bias occurs when there exist hidden variables that simultaneously determine user/item features and the outcome variable. It is well known that, in the absence of hidden confounders, all causal effects can be estimated consistently from non-experimental data. However, in the presence of hidden confounders, whether the desired causal quantity can be estimated depends on the locations of the unmeasured variables, the intervention set, and the outcome. To adjust for confounding bias, one common approach is to condition on a set of covariates that satisfy the backdoor criterion. [79] further generalize the backdoor criterion to identify causal effect if all non-proper causal paths are blocked.

Definition 4 (Proper Causal Path and Proper Backdoor Graph)**.** A causal path from a node on $\mathbf{I}$ to Y is called proper if it does not intersect $\mathbf{I}$ except at the starting point. The proper backdoor graph, denoted as $\mathcal{G}^{pbd}_{Y\mathbf{I}}$, is obtained from $\mathcal{G}$ by removing the first edge of every proper causal path from $\mathbf{I}$ to Y.

Definition 5 (Generalized Backdoor Criterion)**.** A set of variables $\mathbf{Z}$ satisfies the adjustment criterion relative to $(\mathbf{I}, Y)$ in $\mathcal{G}$ if: (i) no element in $\mathbf{Z}$ is a descendant in $\mathcal{G}_{\overline{\mathbf{I}}}$ of any $W \notin \mathbf{I}$ lying on a proper causal path from $\mathbf{I}$ to Y; (ii) all non-causal paths in $\mathcal{G}$ from $\mathbf{I}$ to Y are blocked by $\mathbf{Z}$.

The causal effect can be computed by controlling for a set of covariates $\mathbf{Z}$.

$$P(Y = y|do(\mathbf{I} = \mathbf{i})) = \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z})P(\mathbf{z}) \tag{3.2}$$

Sample Selection Bias arises with a biased selection mechanism, e.g., choosing users based on a certain time or location. Theorem 1 shows one general criteria to test whether a

conditional probability is recoverable from biased data.

Theorem 1. For any disjoint sets $\mathbf{X}$ and $\mathbf{Y}$, the conditional distribution $P(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x})$ is s-recoverable from $\mathcal{G}_s$ if and only if $(S \perp\!\!\!\perp \mathbf{Y}|\mathbf{X})$ where S is the selection mechanism.

[44] study the use of adjustment for recovering causal effects in the presence of confounding and selection biases. They denote $\mathbf{V}$ to be the set of variables measured under selection bias and $\mathbf{T} \subset \mathbf{V}$ to be the subset of variables that are also measured externally, and unbiasedly, in the whole population. They introduce the selection-backdoor criterion as a sufficient and necessary condition for recovering causal effects from a biased distribution $P(\mathbf{v}|S = 1)$ with externally unbiased data $P(\mathbf{t})$.

Theorem 2 (Generalized Adjustment for Causal Effect)**.** Given a causal diagram $\mathcal{G}$ augmented with selection variable S, disjoint sets of variables $Y, \mathbf{I}, \mathbf{Z}$, a set of externally and unbiasedly measured variables $\mathbf{T}$, and a set $\mathbf{Z}^\intercal \subseteq \mathbf{Z} \cap \mathbf{T}$, for every model compatible with $\mathcal{G}$, we have

$$P(y|do(\mathbf{i})) = \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, S = 1)P(\mathbf{z} \backslash \mathbf{z}^\intercal|\mathbf{z}^\intercal, S = 1)P(\mathbf{z}^\intercal) \tag{3.3}$$

if and only if $(\mathbf{Z}, \mathbf{Z}^\intercal)$ satisfies the following generalized adjustment criterion:

1. No element in $\mathbf{Z}$ is a descendant in $\mathcal{G}_{\overline{\mathbf{I}}}$ of any $W \notin \mathbf{I}$ lying on a proper causal path from $\mathbf{I}$ to $\mathbf{Y}$.

2. All non-causal paths in $\mathcal{G}$ from $\mathbf{I}$ to $\mathbf{Y}$ are blocked by $\mathbf{Z}$ and S.

3. $\mathbf{Z}^\intercal$ d-separates $\mathbf{Y}$ from S in the proper backdoor graph, i.e., $(Y \perp\!\!\!\perp S|\mathbf{Z}^\intercal)_{\mathcal{G}^{pbd}_{Y\mathbf{I}}}$.

$(\mathbf{Z}, \mathbf{Z}^\intercal)$ is said to be an adjustment pair for recovering the causal effect of $\mathbf{I}$ on Y.

Theorem 2 studies a general case $\mathbf{Z}^{\intercal} \subseteq \mathbf{Z} \cap \mathbf{T}$ where only a subset $\mathbf{Z}^{\intercal}$ of the covariates $\mathbf{Z}$ requires external measurement. It naturally covers two extreme cases: when none of the covariates is measured externally ($\mathbf{Z} \cap \mathbf{T} = \emptyset$) and when all of them are measured without selection bias ($\mathbf{Z} \subseteq \mathbf{T}$) [45]. Equation 3.3 reduces to $P(y|do(\mathbf{i})) = \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, S = 1)P(\mathbf{z}|S = 1)$ in the former case and $P(y|do(\mathbf{i})) = \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, S = 1)P(\mathbf{z})$ in the latter case.

Instead of identifying causal effect in presence of selection bias by adjustment, [80] proposed a parallel procedure to justify whether a causal quantity is identifiable and recoverable from selection bias using axiomatical c-components factorization [81]. Basically, c-component factorization first partitions nodes in $\mathcal{G}$ into a set of c-components, then expresses the target intervention in terms of the c-factors corresponding to each c-component. Specifically, a *c-component* $\boldsymbol{C}$ denotes a subset of variables in $\mathcal{G}$ such that any two nodes in $\boldsymbol{C}$ are connected by a path entirely consisting of bi-directed edges. A *c-factor* $Q[\boldsymbol{C}](\mathbf{v})$ is a function that demonstrates the post-intervention distribution of $\boldsymbol{C}$ after conducting interventions on the remaining variables $\mathbf{V} \backslash \boldsymbol{C}$ and is defined as

$$Q[\boldsymbol{C}](\mathbf{v}) = P(\boldsymbol{c}|do(\mathbf{v} \backslash \boldsymbol{c})) = \sum_{\mathbf{U}} \prod_{V \in \boldsymbol{C}} P(v|Pa(v), \mathbf{u}_v)P(\mathbf{u})$$

where $Pa(v)$ and $\mathbf{u}_v$ denote the set of observed and unobserved parents for node V. We explicitly denote $Q[\boldsymbol{C}](\mathbf{v})$ as $Q[\boldsymbol{C}]$ and list the factorization formula.

Theorem 3 (C-component Factorization)**.** Given a causal graph $\mathcal{G}$, the target intervention $P(\mathbf{y}|do(\mathbf{x}))$ could be expressed as a product of c-factors associated with the c-components as follows:

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\boldsymbol{C} \backslash \mathbf{Y}} Q[\boldsymbol{C}] = \sum_{\boldsymbol{C} \backslash \mathbf{Y}} \prod_{i=1}^{l} Q[\boldsymbol{C}_i] \tag{3.4}$$

where $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$ could be arbitrary sets, $\boldsymbol{C} = An(\mathbf{Y})_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{X}}}$, and $\boldsymbol{C}_1, ..., \boldsymbol{C}_l$ are the c-components of $\mathcal{G}_{\boldsymbol{C}}$.

Based on the factorization above, [44] showed that $P(\mathbf{y}|do(\mathbf{x}))$ is recoverable and could be computed by Equation 3.4 if each factor $Q[\boldsymbol{C}_i]$ is recoverable from the observational data. Accordingly, they developed the RC algorithm to determine the recoverability of each c-factor.

3.3 Bandit-based Recommendation

LinUCB Algorithm We use the linear contextual bandit [82] as one baseline model for our personalized recommendation. In the linear contextual bandit, the reward for each action is an unknown linear function of the contexts. Formally, we model the personalized recommendation as a contextual multi-armed bandit problem, where each user u is a "bandit player", each potential item $a \in \mathcal{A}$ is an arm and k is the number of item candidates. At time t, there is a coming user u. For each item $a \in \mathcal{A}$, its contextual feature vector $\mathbf{x}_{t,a} \in \mathbb{R}^d$ represents the concatenation of the user and the item feature vectors. The algorithm takes all contextual feature vectors as input, recommends an item $a_t \in \mathcal{A}$ and observes the reward r_{t,a_t}, and then updates its item recommendation strategy with the new observation $(\mathbf{x}_{t,a_t}, a_t, r_{t,a_t})$. During the learning process, the algorithm does not observe the reward information for unchosen items.

The total reward by round t is defined as $\sum_t r_{t,a_t}$ and the optimal expected reward as $\mathbb{E}[\sum_t r_{t,a^*}]$, where a^* indicates the best item that can achieve the maximum reward at time t. We aim to train an algorithm so that the maximum total reward can be achieved. Equivalently, the algorithm aims to minimize the regret $R(T) = \mathbb{E}[\sum_t r_{t,a^*}] - \mathbb{E}[\sum_t r_{t,a_t}]$. The contextual bandit algorithm balances exploration and exploitation to minimize regret since

there is always uncertainty about the user's reward given the specific item.

Algorithm 1 LinUCB

Input: $\alpha \in \mathbb{R}^+$
for t = 1,2,3,...,T **do**
    Observe contextual features of all arms $a \in \mathcal{A}_t : \mathbf{x}_{t,a} \in \mathbb{R}^d$
    for $a \in \mathcal{A}_t$ **do**
        if a is new **then**
            $A_a \leftarrow \mathbf{I_d}$ (d-Dimension identity matrix)
            $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$ (d-Dimension zero vector)
        end if
        $\hat{\boldsymbol{\theta}}_a \leftarrow A_a^{-1}\mathbf{b}_a$
        $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^{\mathrm{T}}\mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^{\mathrm{T}}A_a^{-1}\mathbf{x}_{t,a}}$
    end for
    Choose arm $a_t = argmax_{a\in\mathcal{A}_t}p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_{t,a_t}
    $A_{a_t} \leftarrow A_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^{\mathrm{T}}$
    $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_{t,a_t}\mathbf{x}_{t,a_t}$
end for

Algorithm 1 shows the LinUCB algorithm as introduced by [83]. It assumes the expected reward is linear in its d-dimensional features $\mathbf{x}_{t,a}$ with some unknown coefficient vector $\boldsymbol{\theta}_a^*$. Formally, for all t, we have the expected reward at time t with arm a as $\mathbb{E}[r_{t,a}|\mathbf{x}_{t,a}] = \boldsymbol{\theta}_a^{*\mathrm{T}}\mathbf{x}_{t,a}$. Here the dot product of $\boldsymbol{\theta}_a^*$ and $\mathbf{x}_{t,a}$ could also be succinctly expressed as $\langle\boldsymbol{\theta}_a^*, \mathbf{x}_{t,a}\rangle$. At each round t, we observe the realized reward $r_{t,a} = \langle\boldsymbol{\theta}_a^*, \mathbf{x}_{t,a}\rangle + \epsilon_t$ where ϵ_t is the noise term.

Basically, LinUCB applies ridge regression technique to estimate the true coefficients. Let $D_a \in \mathbb{R}^{m_a\times d}$ denote the context of the historical observations when arm a is selected and $\mathbf{r}_a \in \mathbb{R}^{m_a}$ denote the relative rewards. The regularised least-square estimator for $\boldsymbol{\theta}_a$ could be expressed as:

$$\hat{\boldsymbol{\theta}}_a = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\arg\min}\left(\sum_{i=1}^{m_a}(r_{i,a} - \langle\boldsymbol{\theta}, D_a(i,:)\rangle)^2 + \lambda||\boldsymbol{\theta}||_2^2\right) \tag{3.5}$$

where λ is the penalty factor of the ridge regression. The solution to Equation 3.5 is:

$$\hat{\boldsymbol{\theta}}_a = (D_a^{\mathrm{T}} D_a + \lambda I_d)^{-1} D_a^{\mathrm{T}} \mathbf{r}_a \tag{3.6}$$

[83] derived a confidence interval that contains the true expected reward with probability at least $1 - \delta$:

$$\left| \hat{\boldsymbol{\theta}}_a^{\mathrm{T}} \mathbf{x}_{t,a} - \mathbb{E}[r_{t,a} | \mathbf{x}_{t,a}] \right| \leq \alpha \sqrt{\mathbf{x}_{t,a}^{\mathrm{T}} (D_a^{\mathrm{T}} D_a + \lambda I_d) \mathbf{x}_{t,a}}$$

for any $\delta > 0$, where $\alpha = 1 + \sqrt{ln(2/\delta)/2}$. Following the rule of optimism in the face of uncertainty for linear bandits (OFUL), this confidence bound leads to a reasonable arm-selection strategy: at each round t, pick an arm by

$$a_t = argmax_{a \in \mathcal{A}_t} \left(\hat{\boldsymbol{\theta}}_a^{\mathrm{T}} \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^{\mathrm{T}} A_a^{-1} \mathbf{x}_{t,a}} \right) \tag{3.7}$$

where $A_a = D_a^{\mathrm{T}} D_a + \lambda I_d$. The parameter λ could be tuned to a suitable value in order to improve the algorithm's performance. Line 13 and 14 in Algorithm 1 provide an iterative way to update the arm-related matrices A_a and b_a. In the remaining content we will denote the weighted 2-norm $\sqrt{\mathbf{x}_{t,a}^{\mathrm{T}} A_a^{-1} \mathbf{x}_{t,a}}$ as $||\mathbf{x}_{t,a}||_{A_a^{-1}}$ for the sake of simplicity.

Regret Bound of LinUCB

Existing research works (e.g., [50, 84]) on deriving the regret bound of LinUCB are based on the following four assumptions:

1. The true coefficient $\boldsymbol{\theta}^*$ is shared by all arms.

2. The error term ϵ_t follows 1-sub-Gaussian distribution for each time point.

3. $\{\alpha_t\}_{i=1}^n$ is a non-decreasing sequence with $\alpha_1 \geq 1$.

4. $||\mathbf{x}_{t,a}||_2 < L$, $||\boldsymbol{\theta}^*||_2 < M$ for all time points and arms.

For assumption 1, since there is only one unified $\boldsymbol{\theta}$, we change the notation of D_a, $\mathbf{r}_a$ to D_t and $\mathbf{r}_t$ to denote the historical observations up to time t for all arms. The matrix A_a will be denoted as A_t accordingly. For assumption 3, following [50] and [84], we modify α in Algorithm 1 to be a time dependent sequence to get a suitable confidence set for $\boldsymbol{\theta}^*$ at each round, but use a fixed and tuned α in the experiment part to make the online computation more efficient.

To derive the regret bound, the first step is to construct a confidence set $\mathcal{C}_t \in \mathbb{R}^d$ for the true coefficient. At each round t, a natural choice is to make $\mathcal{C}_t$ centered at $\hat{\boldsymbol{\theta}}_{t-1}$. [50] shows that the confidence ellipsoid could be a suitable choice for constructing the confidence region, which is defined as follows:

$$\mathcal{C}_t = \{\boldsymbol{\theta} \in \mathbb{R}^d : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t-1}||_{A_{t-1}} < \alpha_t\}$$

The key point is how to obtain an appropriate α_t at each round to make $\mathcal{C}_t$ contain the true parameter $\boldsymbol{\theta}^*$ with high probability and be as small as possible simultaneously. [50] takes the advantages of the martingale techniques and derives a confidence bound in terms of the weighted 2-norm shown in Lemma 1.

Lemma 1. (Theorem 2 in [50]) Suppose the noise term is 1-sub-Gaussian distributed, let $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that for all $t \in \mathbb{N}^+$,

$$||\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_t||_{A_t} \leq \sqrt{\lambda}||\boldsymbol{\theta}^*||_2 + \sqrt{2log(|A_t|^{1/2}|\lambda I_d|^{-1/2}\delta^{-1})} \tag{3.8}$$

The RHS of Equation 3.8 gives an appropriate selection of α_t for the confidence ellipsoid. Under the fact that $\theta^* \in \mathcal{C}_t$ and the optimistic arm selection rule of LinUCB we could further bound the regret at each round with high probability by $r_t = \langle \boldsymbol{\theta}^*, \mathbf{x}_{t,a} \rangle - \langle \hat{\boldsymbol{\theta}}, \mathbf{x}_{t,a} \rangle \leq 2\alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}}$. Summing up the regret at each round, the following corollary gives a $\tilde{\mathcal{O}}(dlog(T))$ cumulative regret bound up to time T.

Corollary 1. (Corollary 19.3 in [85]) Under the assumptions above, the expected regret of LinUCB with $\delta = 1/T$ is bounded by

$$R_T \leq Cd\sqrt{T}log(TL) \tag{3.9}$$

where C is a suitably large constant.

4 Achieving User-side Fairness in Bandit-based Recommendation

4.1 Introduction

Personalized recommendation based on multi-arm bandit (MAB) algorithms has become a popular topic of research and shown to lead to high utility and efficiency [22] as it dynamically adapts the recommendation strategy based on feedback. However, it is also known that such personalization could incur biases or even discrimination that can influence decisions and opinions [23, 24]. Recently researchers have started taking fairness and discrimination into consideration in the design of MAB based personalized recommendation algorithms [25, 26, 27]. However, they focused on the fairness of the recommended items (e.g., services provided by small or large companies) instead of the customers who received those items. For example, [26] focused on individual fairness, i.e., "treating similar individuals similarly," and considered the individual as an arm with the aim of ensuring the probability of selecting an arm is equal to the probability with which the arm has the best quality realization. [25] aimed to achieve group fairness over items by ensuring the probability distribution from which items are sampled satisfies certain fairness constraints at all time steps. In this chapter, we aim to develop novel algorithms to ensure fair and ethical treatment of customers with different profile attributes (e.g., gender, race, education, disability, and economic conditions) in a contextual bandit based personalized recommendation.

Consider the personalized educational video recommendation in Table 4.3 as an illustrative example. Table 4.1 shows two students, Alice and Bob, having the same profiles except for the gender. Table 4.2 shows potential videos and Table 4.3 shows recommendations

Student	Gender	Grade	GPA	...
Alice	female	9th	2.6	...
Bob	male	9th	2.6	...
...	...	...	...	...

Table 4.1: Students.

Video	Gender of speaker	rating	length	...
2501	female	4.3	4 minutes	...
0964	male	4.3	6 minutes	...
...	...	...	...	...

Table 4.2: Videos.

by a personalized recommendation algorithm. Focusing on the fairness of the video would ensure that videos featuring female speakers have similar chances of being recommended as those featuring male speakers. However, one group of students could benefit more from the recommended videos than the other group, therefore yielding to an unequal improvement of the learning performances. In our work, rather than focusing on the fairness of the item being recommended, i.e., the video, we focus on the user-side fairness in terms of the reward, i.e., the improvement of student's learning performance after watching the recommended video. We want to ensure that both male students and female students who share similar profiles will receive a similar reward regardless of the video being recommended, such that they both benefit from the video recommendations and improve their learning performance equally.

We study how to achieve the user-side fairness in the classic contextual bandit algorithm. The contextual bandit framework [48], which is used to sequentially recommend

Student	Video	Reward
Alice	2501	0.60
Bob	0964	0.80
...	...	...

Table 4.3: Recommendations.

items to a customer based on her contextual information, is able to fit user preferences, address the cold-start problem by balancing the exploration and exploitation trade-off in recommendation systems, and simultaneously adapt the recommendation strategy based on feedback to maximize the customer's learning performance. However, such a personalized recommendation system could induce an unfair treatment of certain customers which could lead to discrimination. We develop a novel fairness aware contextual bandit algorithm such that customers will be treated fairly in personalized learning.

We train our fair contextual bandit algorithm to detect discrimination, that is, whether or not a group of customers is being privileged in terms of reward received. Our fair contextual bandit algorithm then measures to what degree each of the items (arms in bandits) contributes to the discrimination. Furthermore, in order to counter the discrimination, if any, we introduce a fairness penalty factor. The goal of this penalty factor is to maintain a balance between fairness and utility, by ensuring that the arm picking strategy will not incur discrimination whilst achieving good utility. Finally, we compare our algorithm against the traditional LinUCB both theoretically and empirically and we show that our approach not only achieves group-level fairness in terms of reward, but also yields comparable effectiveness.

Overall, our contributions are two-fold. First, we develop a fairness aware contextual bandit algorithm that achieves user-side fairness in terms of reward and is robust against factors that would otherwise increase or incur discrimination. Secondly, we provide a theoretical regret analysis to show that our algorithm has a regret bound higher than LinUCB up to only an additive constant.

4.2 Fairness Aware Contextual Bandits

We focus on how to achieve user-side fairness in contextual bandit based recommendation and present our fair contextual bandit algorithm, called Fair-LinUCB and derive its regret bound.

4.2.1 Problem Formulation

We define a sensitive attribute $S \in \mathbf{x}_{t,a}$ with domain values $\{s^+, s^-\}$ where s^+ (s^-) is the value of the privileged (protected) group. Let T_s denote a time index subset such that the users being treated at time points in T_s all hold the same sensitive attribute value s. We introduce the group-level cumulative mean reward as $\bar{r}^s = \frac{1}{|T_s|}\sum_{t\in T_s} r_{t,a}$. Specifically, $\bar{r}^{s^+}$ denotes the cumulative mean reward of the individuals with sensitive attribute $S = s^+$, and $\bar{r}^{s^-}$ denotes the cumulative mean reward of all individuals having the sensitive attribute $S = s^-$.

We define the group fairness in contextual bandits as $\mathbb{E}[\bar{r}^{s^+}] = \mathbb{E}[\bar{r}^{s^-}]$, more specifically, the expected mean reward of the protected group and that of the unprotected group should be equal. A recommendation algorithm incurs group-level unfairness in regards to a sensitive attribute S if $|\mathbb{E}[\bar{r}^{s^+}] - \mathbb{E}[\bar{r}^{s^-}]| > \tau$ where $\tau \in \mathbb{R}^+$ reflects the tolerance degree of unfairness.

4.2.2 Fair-LinUCB Algorithm

We describe our fair LinUCB algorithm and show its pseudo code in Algorithm 2. The key difference from the traditional LinUCB is the strategy of choosing an arm during recommendation (shown in Line 12 of Algorithm 2). In the remaining of this section, we explain how this new strategy achieves user-side group-level fairness.

Algorithm 2 Fair-LinUCB

Input: α , $\gamma \in \mathbb{R}^+$
$\bar{r}^{s^+}, \bar{r}^{s^-} \leftarrow 0$
for t = 1,2,3,..., T **do**
    Observe features of all arms $a \in \mathcal{A}_t : \mathbf{x}_{t,a} \in \mathbb{R}^d$
    for $a \in \mathcal{A}_t$ **do**
        if a is new **then**
            $A_a \leftarrow \lambda \mathbf{I}_d$ (d-Dimension identity matrix)
            $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$ (d-Dimension zero vector)
            $\bar{r}_a^{s^+}, \bar{r}_a^{s^-} \leftarrow 0$
        end if
        $\hat{\boldsymbol{\theta}}_a \leftarrow A_a^{-1}\mathbf{b}_a$
        $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^{\mathrm{T}}\mathbf{x}_{t,a} + \alpha||\mathbf{x}_{t,a}||_{A_a^{-1}} + \mathcal{L}(\gamma, F_a)$
    end for
    Choose arm $a_t = argmax_{a\in\mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_{t,a_t}
    $A_a \leftarrow A_a + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^{\mathrm{T}}$
    $\mathbf{b}_a \leftarrow \mathbf{b}_a + r_{t,a_t}\mathbf{x}_{t,a_t}$
    if $S_t = s^+$ **then**
        update $\bar{r}^{s^+}, \bar{r}_a^{s^+}$ with r_{t,a_t}
    else
        update $\bar{r}^{s^-}, \bar{r}_a^{s^-}$ with r_{t,a_t}
    end if
end for

Given a sensitive attribute S with domain values $\{s^+, s^-\}$, the goal of our fair contextual bandit is to minimize the cumulative mean reward difference between the protected group and the privileged group while preserving its efficiency. Note that Fair-LinUCB can be extended to the general setting of multiple sensitive attributes $S_j \in \boldsymbol{S} = \{S_1, S_2, ..., S_l\}$ where $\boldsymbol{S} \subset \mathbf{x}_{t,a}$ and each S_j can have multiple domain values. In order to measure the unfairness at the group-level, our Fair-LinUCB algorithm will keep track of both cumulative mean rewards along the time, e.g., $\bar{r}^{s^+}$ and $\bar{r}^{s^-}$. We capture the orientation of the bias (i.e., towards which group the bias is leaning) through the sign of the cumulative mean reward difference. By doing so, Fair-LinUCB is able to know which group is being discriminated and which group is being privileged.

When running context bandits for recommendation, each arm may generate a reward discrepancy and therefore contribute to the unfairness to some degree. Fair-LinUCB captures the reward discrepancy at the arm level by keeping track of the cumulative mean reward generated by each arm a for both groups s^+ and s^-. Specifically, let $\bar{r}_a^{s^+}$ denote the average of the rewards generated by arm a for the group s^+, and let $\bar{r}_a^{s^-}$ denote the average of the rewards generated by arm a for the group s^-. The bias of an arm is thus the difference of both averages: $\Delta_a = (\bar{r}_a^{s^+} - \bar{r}_a^{s^-})$. Finally, by combining the direction of the bias and the amount of the bias induced by each arm a, we define the fairness penalty term as $F_a = -sign(\bar{r}^{s^+} - \bar{r}^{s^-}) \cdot \Delta_a$, and exert onto the UCB value in our fair contextual bandit algorithm. Note that the lesser an arm contributes to the bias, the smaller the penalty.

As a result, if an arm has a high UCB but incurs bias, its adjusted UCB value will decrease and it will be less likely to be picked by the algorithm. In contrast, if an arm has a small UCB but is fair, its adjusted UCB value will increase, and it will be more likely to be picked by the algorithm, thereby reducing the potential unfairness in recommendation.

Different from the traditional LinUCB that picks the arm to solely maximize the UCB, our Fair-LinUCB accounts for the fairness of the arm and picks the arm that maximizes the summation of the UCB and the fairness. Formally, we show the modified arm selection criteria in Equation 4.1.

$$p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^{\mathrm{T}} \mathbf{x}_{t,a} + \alpha ||\mathbf{x}_{t,a}||_{A_a^{-1}} + \mathcal{L}(\gamma, F_a) \tag{4.1}$$

We adopt a linear mapping function $\mathcal{L}$ with input parameters γ and F_a to transform the fairness penalty term proportionally to the size of its confidence interval. Specifically,

$$\mathcal{L}(\gamma, F_a) = \frac{\alpha_t ||\mathbf{x}_{t,a_m}||_{A_t^{-1}}}{2}(F_a + 1)\gamma \tag{4.2}$$

$$a_m = argmin_{a \in \mathcal{A}_t} ||\mathbf{x}_{t,a}||_{A_a^{-1}} \tag{4.3}$$

Assuming that the reward generated is in the range $[0, 1]$, the fairness penalty F_a lies in $[-1, 1]$. When designing the coefficient of the linear mapping function, we choose a_m to be the arm with the smallest confidence interval to guarantee a unified fairness calibration among all the arms. Under the effect of $\mathcal{L}$, the range of the fairness penalty is mapped from $[-1, 1]$ to $[0,\ \gamma\alpha_t ||\mathbf{x}_{t,a_m}||_{A_t^{-1}}]$, which implies a similar scale with the confidence interval. In our empirical evaluations, we show how γ controls fairness-accuracy trade-off on the practical performance of Fair-LinUCB.

4.2.3 Regret Analysis

In this section, We prove that our Fair-LinUCB algorithm has a $\tilde{\mathcal{O}}(dlog(T))$ regret bound under certain assumptions with carefully chosen parameters. We adopt the regret

analysis framework of linear contextual bandit and introduce a mapping function on the fairness penalty term. By applying the mapping function $\mathcal{L}$ we make our fairness penalty term possess the similar scale with the half length of the confidence interval. Thus we can merge the regret generated by UCB term and fairness term together and derive our regret bound.

Theorem 4. Under the same assumptions shown in Section 3.3, further assuming γ is a moderate small constant with $\gamma \leq \Gamma$, there exists $\delta \in (0, 1)$ such that with probability at least $1 - \delta$ Fair-LinUCB achieves the following regret bound:

$$R_T \leq \sqrt{2Tdlog(1 + TL^2/(d\lambda))} \times (2 + \Gamma)(\sqrt{\lambda}M + \sqrt{2log(1/\delta) + dlog(1 + TL^2/(d\lambda))}) \quad (4.4)$$

Proof. We first introduce three technical lemmas from [50] and [85] to help us complete the proof of Theorem 4.

Lemma 2. (Lemma 11 in appendix of [50]) If $\lambda \geq max(1, L^2)$, the weighted L2-norm of feature vector could be bounded by : $\sum_{t=1}^{T} ||\mathbf{x}_{t,a}||^2_{A_t^{-1}} \leq 2log\frac{|A_t|}{\lambda^d}$

Lemma 3. (Lemma 10 in appendix of [50]) The determinant $|A_t|$ could be bounded by: $|A_t| \leq (\lambda + tL^2/d)^d$.

Lemma 4. (Theorem 20.5 in [85]) With probability at least $1 - \delta$, for all the time point $t \in \mathbb{N}^+$ the true coefficient $\boldsymbol{\theta}^*$ lies in the set:

$$\mathcal{C}_t = \{\boldsymbol{\theta} \in \mathbb{R}^d : ||\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}||_{A_t} \leq \sqrt{\lambda}M + \sqrt{2log(1/\delta) + dlog(1 + TL^2/(d\lambda))}\} \quad (4.5)$$

In Fair-LinUCB, the range of fairness term is $[-1, 1]$, we apply a linear mapping function $\mathcal{L}(\gamma, x) = \frac{\alpha_t ||\mathbf{x}_{t,a_m}||_{A_t^{-1}}}{2}(x+1)\gamma$ to map the range of $\mathcal{L}(\gamma, F_a)$ to $[0,\ \gamma\alpha_t ||\mathbf{x}_{t,a_m}||_{A_t^{-1}}]$, where $a_m = argmin_{a \in \mathcal{A}_t} ||\mathbf{x}_{t,a}||_{A_a^{-1}}$.

According to the rule, the regret at each time t is bounded by:

$$
\begin{aligned}
reg_t &= \mathbf{x}_{t,a}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_t - \mathbf{x}_{t,a}^{\mathrm{T}} \boldsymbol{\theta}^* \\
&\leq \mathbf{x}_{t,a}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_t + \alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} + \mathcal{L}(\gamma, F_a) - \mathbf{x}_{t,a}^{\mathrm{T}} \boldsymbol{\theta}^* \\
&\leq \mathbf{x}_{t,a}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_t + \alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} + \mathcal{L}(\gamma, F_a) - (\mathbf{x}_{t,a}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_{\boldsymbol{t}} - \alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}}) \\
&\leq 2\alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} + \mathcal{L}(\gamma, 1) \\
&= 2\alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} + \gamma\alpha_t ||\mathbf{x}_{t,a_m}||_{A_t^{-1}} \\
&\leq 2\alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} + \gamma\alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} \\
&\leq (2+\Gamma)\alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}}
\end{aligned}
$$

The second line above is derived based on the theoretic result in Lemma 1 and following the selection rule of the Fair-LinUCB algorithm, specifically, $\mathbf{x}_{t,a^*}^{\mathrm{T}} \boldsymbol{\theta}^* \leq \mathbf{x}_{t,a^*}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_t + \alpha_t ||\mathbf{x}_{t,a^*}||_{A_t^{-1}} \leq \mathbf{x}_{t,a^*}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_t + \alpha_t ||\mathbf{x}_{t,a^*}||_{A_t^{-1}} + \mathcal{L}(\gamma, F_{a^*}) \leq \mathbf{x}_{t,a}^{\mathrm{T}} \hat{\boldsymbol{\theta}}_t + \alpha_t ||\mathbf{x}_{t,a}||_{A_t^{-1}} + \mathcal{L}(\gamma, F_a)$. Note that Lemma 1 can be equally applied here because the estimator $\hat{\theta}_t$ is still a valid ridge regression estimator at each round.

Summing up the regret at each bound, with probability at least $1-\delta$ the cumulative regret up to time T is bounded by:

$$
R_T = \sum_{t=1}^{T} reg_t \leq \sqrt{T \sum_{t=1}^{T} reg_t^2} \leq (2+\Gamma)\alpha_T \sqrt{T \sum_{t=1}^{T} ||\mathbf{x}_{t,a}||_{A_t^{-1}}^2} \tag{4.6}
$$

Since $\{\alpha_t\}_{i=1}^n$ is a non-decreasing sequence, we can enlarge each element α_t to α_T to obtain the inequalities in Equation 4.6. By applying the inequalities from Lemma 2 and 3 we could further relax the regret bound up to time T to:

$$\begin{aligned} R_T &\leq (2+\Gamma)\alpha_T\sqrt{2Tlog\frac{|A_t|}{\lambda^d}} \\ &\leq (2+\Gamma)\alpha_T\sqrt{2Td(log(\lambda+TL^2/d)-log\lambda)} \\ &= (2+\Gamma)\alpha_T\sqrt{2Tdlog(1+TL^2/(d\lambda))} \end{aligned} \tag{4.7}$$

Following the result of Lemma 1, by loosing the determinant of A_t according to Lemma 3, Lemma 4 provides a suitable choice for α_T up to time T. By plugging in the RHS from Equation 4.5 we get the regret bound shown in Theorem 4:

$$R_T \leq \sqrt{2Tdlog(1+TL^2/(d\lambda))} \times (2+\Gamma)(\sqrt{\lambda}M + \sqrt{2log(1/\delta)+dlog(1+TL^2/(d\lambda))})$$

□

Corollary 2. Setting $\delta = 1/T$, the regret bound in Theorem 4 could be simplified as $R_T \leq C'd\sqrt{T}log(TL)$.

Comparing Corollary 2 with Corollary 1 (for LinUCB), we can see the regret bound of Fair-LinUCB is worse than the original LinUCB only up to an additive constant. This perfectly matches the intuition that Fair-LinUCB is able to keep aware of the fairness and guarantee there is no reward gap between different subgroups or individuals, however, it suffers from a relatively higher regret.

4.3 Experimental Evaluation

4.3.1 Experiment Setup

4.3.1.1 Simulated Dataset

There are presently no publicly available datasets that fits our environment. We therefore generate one simulated dataset for our experiments by combining the following two publicly available datasets.

- **Adult dataset**: The Adult dataset [86] is used to represent the students (or bandit players). It is composed of 31,561 instances: 21,790 males and 10,771 females, each having 8 categorical variables (work class, education, marital status, occupation, relationship, race, sex, native-country) and 3 continuous variables (age, education number, hours per week), yielding an overall of 107 features after one-hot encoding.

- **YouTube dataset**: The Statistics and Social Network of YouTube Videos [1] dataset is used to represent the items to be recommended (or arms). It is composed of 1,580 instances each having 6 categorical features (age of video, length of video, number of views, rate, ratings, number of comments), yielding a total of 25 features after one-hot encoding. We add a 26^{th} feature used to represent the gender of the speaker in the video which is drawn from a Bernoulli distribution with the probability of success as 0.5.

The feature contexts $\mathbf{x}_{t,a}$ used throughout the experiment is the concatenation of both the student feature vector and the video feature vector. In our experiments we choose the sensitive attribute to be the **gender of adults**, and we therefore focus on the unfairness on

[1] https://netsg.cs.sfu.ca/youtubedata/

the group-level for the male group and female group. Furthermore, we assume that a male student prefers a video featuring a male and a female student prefers a video featuring a female speaker. Thus, in order to maintain the linear assumption of the reward function, we add an extra binary variable in the feature context vector that represents whether or not the gender of the student matches the gender of the speaker in the video. Overall, $\mathbf{x}_{t,a}$ contains a total of 134 features.

For our experiments, we use a subset of 5,000 random instances from the Adult dataset, which is then split into two subsets: one for training and one for testing. The training subset is composed of 1,500 male individuals and 1,500 female individuals whilst the testing subset is composed of 1000 males and 1000 females. Similarly, a subset of YouTube dataset is used as our pool of videos to recommend (or arms). The subset contains 30 videos featuring a male speaker and 70 videos featuring a female speaker.

4.3.1.2 Reward Function

We compare our Fair-LinUCB against the original LinUCB using a simple reward function wherein we manually set the $\boldsymbol{\theta}^*$ coefficients. The reward r is defined as

$$r = \theta_1^* \cdot x_1 + \theta_2^* \cdot x_2 + \theta_3^* \cdot x_3$$

where $\theta_1^* = 0.3$, $\theta_2^* = 0.4$, $\theta_3^* = 0.3$ and $x_1 =$ video rating, $x_2 =$ education level, $x_3 =$ gender match. The remaining $d-3$ coefficients are set to 0. Hence, only these three features matter to generate our true reward. The gender match is set to 1 if both the student gender and the gender of the video match, and 0 otherwise. The education level is divided into 5 subgroups each represented by a value ranging from 0.0 to 1.0 with a higher education level

yielding a higher value. In our setup, the education level is used to represent the strength of the student. Similarly, the video rating varies from 0 to 1.0, and is used to represent the educational quality of the video. Evidently, a higher reward is generated when the gender of the student matches the gender of the video.

4.3.1.3 Evaluation Metrics

Throughout our experiments we measure the effectiveness of the algorithms through the average utility loss. Since we know the true reward function, we can derive the optimal reward at each round t. We can thus define

$$\text{utility loss} = \frac{1}{T}\sum_{t=1}^{T}(r_{t,a^*} - r_{t,a})$$

where r_{t,a^*} is the optimal reward at round t by choosing arm a^* and $r_{t,a}$ is the observed reward by the algorithm after picking arm a.

We measure the fairness of the algorithms through the absolute value of the difference between the cumulative mean reward ($\bar{r}_t$, as introduced in Section 4.2.1) of the male group and female group:

$$\text{reward difference} = |\bar{r}_t^{s^+} - \bar{r}_t^{s^-}|$$

Additionally, for all following figures the left hand side plots the cumulative mean reward during the training phase whilst the right hand side reflects the cumulative mean reward over the testing dataset. Note that the contextual bandit continues to learn throughout both phases.

4.3.1.4 Baselines

As existing fair bandits algorithms focus on item-side fairness, we mainly compare our Fair-LinUCB against LinUCB in terms of utility-fairness trade-off in our evaluations. We also report a comparison with a simple fair LinUCB method that suppresses the unfairness by removing the sensitive attribute and all its correlated attributes from the context. We name this method as Naive in our evaluation.

4.3.2 Comparison with Baselines

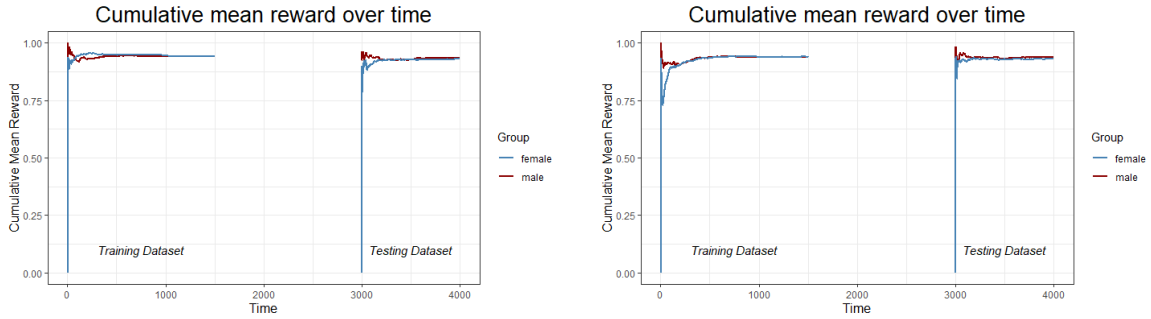
4.3.2.1 Comparison with LinUCB

Our first experiment compares the performances of the traditional LinUCB against our Fair-LinUCB, using the reward function r described in the previous section. Figure 4.2 plots the cumulative mean reward of both the male and female groups over time. We can notice that the cumulative mean rewards of both groups suffer a discrepancy with LinUCB, and the outcome can therefore be considered unfair towards the male group. Indeed, as shown on Figure 4.2a the cumulative mean reward of the female group (0.839) is greater than the cumulative mean reward of the male group (0.802), yielding a reward difference of 0.037. The utility loss incurred is 0.050. In contrast, Fair-LinUCB is able to seal the reward discrepancy with a γ coefficient set to 3 (Figure 4.2b). Our algorithm thereby achieves a cumulative mean reward of 0.819 for both the male group and the female group, which yields a reward difference of 0.0, while incurring a utility loss of 0.052. Our Fair-LinUCB outperforms the traditional LinUCB in terms of reward difference while suffering a slight loss of utility. The comparison results are summarized in the first two rows of Table 4.4.

To evaluate how the inclusion or exclusion of sensitive attributes affects the fairness-

Table 4.4: Comparison of three algorithms under reward function r.

	Utility Loss	Reward difference
Fair-LinUCB ($\gamma = 3$)	0.052	**0.000**
LinUCB	0.050	0.037
Naive	**0.046**	0.035



(a) (b)

Figure 4.1: LinUCB (a) vs Fair-LinUCB $\gamma = 3$ (b) with reward function r_2.

utility tradeoff, we compare LinUCB against Fair-LinUCB with a modified reward function:

$$r_2 = \theta_1^* \cdot x_1 + \theta_2^* \cdot x_2$$

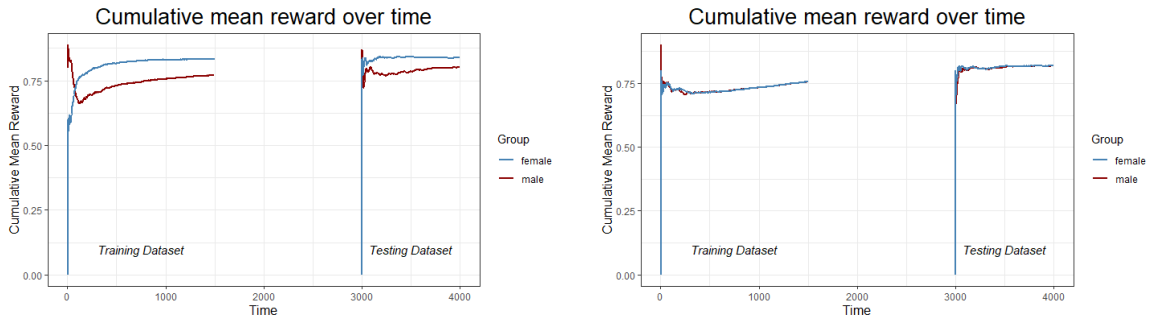
where $\theta_1^* = 0.5$ and $\theta_2^* = 0.5$ and $x_1 =$ video rating, $x_2 =$ education level The remaining $d - 2$ coefficients are set to 0. r_2 is not dependent upon the gender match attribute and expects to incur zero or small discrepancy between both groups. As depicted on Figure 4.1, both LinUCB and Fair-LinUCB show a very low cumulative mean reward discrepancy. Specifically, LinUCB incurs a utility loss of 0.037 and a reward difference of 0.006, while Fair-LinUCB incurs 0.034 utility loss and a reward difference of 0.008. Furthermore, in this case, although Fair-LinUCB has additional constraints for the arm picking strategy due to the fairness penalty, it does not induce any loss of utility when compared to LinUCB.

4.3.2.2 Comparison with Naive

Naive method tries to achieve fairness by removing from the context the sensitive attribute and the features that are highly correlated with the sensitive attribute. In our experiment, we first compute the correlation matrix of all the user's features and then remove the gender feature as well as all features that are highly correlated with it. Specifically, features that have a correlation coefficient greater than 0.3 were removed, which include the following: is male, is female, is divorced, is married, is widowed, is a husband, has an administrative clerical job, has a salary less than 50k. We report in the last row of Table 4.4 the utility loss and reward difference of Naive with reward function r.

We can see the reward discrepancy between the male and female groups from the Naive method is 0.035, thus showing it cannot completely remove discrimination. The utility loss from the Naive method is 0.046, which is only slightly smaller than LinUCB and Fair-LinUCB. In fact, as shown in Table 4.5, Fair-LinUCB with $\gamma = 2$ can outperform the Naive method in terms of both fairness and utility. In short, removing the gender information and highly correlated features from the context does not necessarily close the gap of the reward difference.

In summary, although LinUCB learns to pick the arm that maximizes the reward given a particular context, we have seen that it could incur discrimination towards a group of users in some cases. Fair-LinUCB is capable of detecting when unfairness occurs, and will adapt its arm picking strategy accordingly so as to be as fair as possible and reduce any reward discrepancy. When a reward discrepancy is not detected, our algorithm does not need to adjust the arm picking strategy and therefore performs as well as the traditional LinUCB.



(a) (b)

Figure 4.2: LinUCB (a) vs Fair-LinUCB $\gamma = 3$ (b) with reward function r.

Table 4.5: Impact of γ on the fairness-utility trade-off.

	Utility Loss	Reward difference
$\gamma = 0$	0.050	0.037
$\gamma = 1$	0.040	0.016
$\gamma = 2$	**0.035**	0.004
$\gamma = 3$	0.052	**0.000**
$\gamma = 4$	0.081	**0.000**

4.3.3 Impact of γ on Fairness-Utility Trade-off

The γ coefficient introduced in Section 4.2.2 controls the weight of the fairness penalty that the algorithm will exert onto the UCB value. Indeed, as shown in Equation (4.2), γ is used to adjust the upper bound of the linear mapping function $\mathcal{L}(\gamma, F_a)$. Thus, when the γ coefficient increases, the range of the fairness penalty increases proportionally which will consequently increase the UCB value in Equation 4.1. The γ coefficient therefore reflects the significance of the fairness of Fair-LinUCB. However, as γ becomes larger, the fairness penalty becomes out of proportion to the extent of neglecting the importance of the UCB value, thereby decreasing the utility of the algorithm.

To evaluate the fairness-utility trade-off of Fair-LinUCB, we compare several γ values and report the fairness and utility loss in Table 4.5. With a γ equal to 0, our algorithm behaves as a traditional LinUCB, therefore it incurs discrimination (reward difference mea-

sured at 0.037), and a utility loss of 0.050 is reported. We can observe that when γ increases slightly, the algorithm improves the reward difference and loss of utility. Specifically, a reward difference of 0.016 is achieved for $\gamma = 1$ with a utility loss of 0.040, and a reward difference of 0.004 with a utility loss of 0.035 is achieved with $\gamma = 2$. Although the utility losses are improved, they both remain not fair. In our best case scenario, with $\gamma = 3$, the algorithm is completely fair, i.e., reward difference is 0.000, with a utility loss of 0.052. Finally, when the γ coefficient is too large, the algorithm prioritizes fairness over utility, resulting in a fair algorithm that suffers a greater loss of utility. For example, with a γ set to 4, Fair-LinUCB incurs a utility loss of 0.081.

4.3.4 Impact of Arm and User Distributions

In certain cases the distribution of the arms (videos) or the users can significantly impact the cumulative mean reward of some groups of users, and therefore incur the large reward difference. In our experiment, given the reward function r, we first explore the impact of the ratio of gender arms, i.e., videos by female or male speakers, and then we investigate the impact of the order of the data in which the algorithm learns. The following results discuss our findings. We explore the effect of three different arm ratio values: (1) 70% male and 30% female, (2) 50% male and 50% female, and (3) 30% male and 70% female. Table 4.6 reports the utility loss, reward difference, as well as both the cumulative mean reward for the male and female groups. As observed with the LinUCB performances, the arm ratio induces unfairness on some user group. Indeed, when there is a majority of male arms, it appears that the male user group will benefit more and will have a higher cumulative mean reward. Likewise, when the arms have more females than males, the female user group will benefit more than the male user group, and will therefore have a higher cumulative mean

Table 4.6: Impact of different arm ratio on the fairness and utility.

Arm ratio m:f	Utility Loss	Reward difference	Male cmr	Female cmr
LinUCB				
7:3	0.061	0.029	0.824	0.795
1:1	0.053	0.012	0.824	0.812
3:7	**0.050**	0.037	0.802	0.839
Fair-LinUCB $\gamma = 3$				
7:3	0.087	0.001	0.784	0.783
1:1	0.162	**0.000**	0.709	0.709
3:7	0.052	**0.000**	0.819	0.819

reward. Although having a balanced ratio of male and female arms minimizes the reward difference, it is not always feasible or convenient to adjust the arms distribution in practice.

We ran the same experiment with Fair-LinUCB with $\gamma = 3$. As we can see, in all three cases, Fair-LinUCB yields a very low reward difference. Indeed, our Fair-LinUCB learns which group is being discriminated and adjusts its arm picking strategy accordingly so as to remove any discrimination, it however suffers a higher utility loss than LinUCB. Note that a γ different than 3 could yield a better utility loss for the ratios 7:3 and 1:1.

Thus, as opposed to a traditional LinUCB which only learns to maximize the reward given a context, our Fair-LinUCB learns how to achieve fairness at the same time, making it robust against factors that would otherwise induce unfairness.

It is also our intuition that the order of the data in which LinUCB learns to recommend an item could affect its recommendation choice or arm pick.

In these experiments, we use the 70% male and 30% female arms setting, and we manually change the order of the training data. In the first setting, we manually set the order of the students in the training data by having all 1,500 female students followed by the 1,500 males instances. In the second setting we order the data by having all 1,500 male instances first, followed by the 1,500 female instances. The test data remains shuffled. We

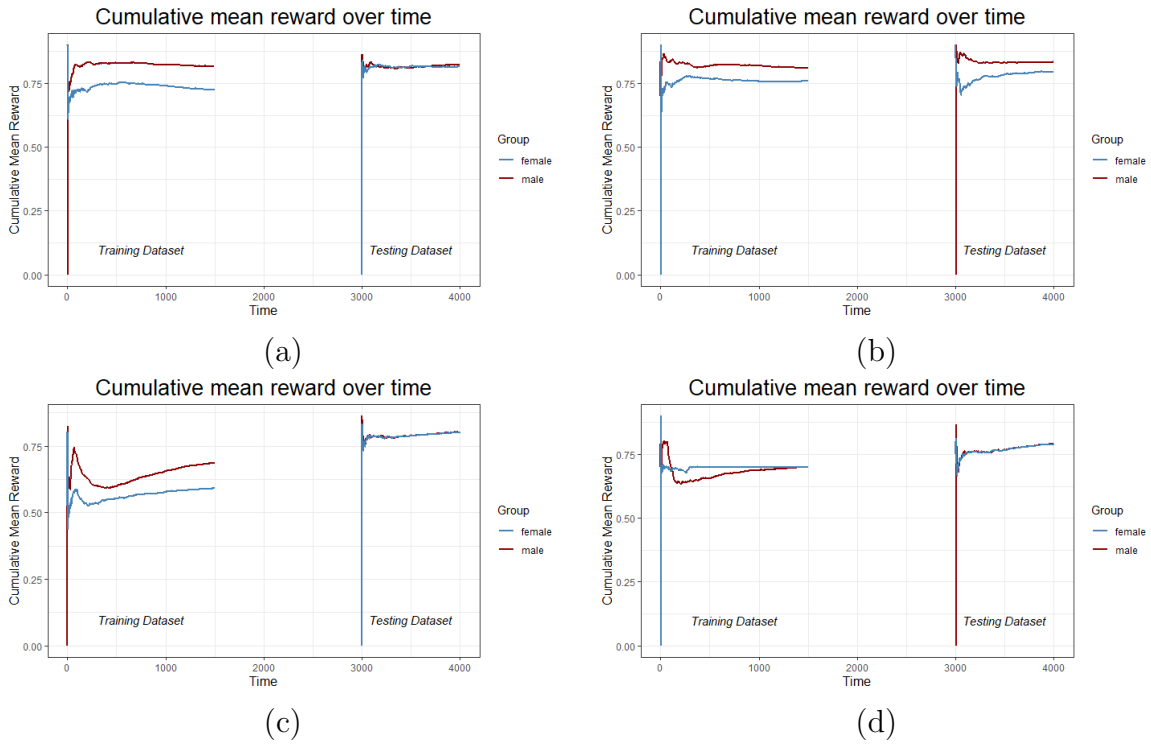


Figure 4.3: Impact of the order of the data on the performances.

then compare LinUCB with Fair-LinUCB in order to see the impact on the learning strategy of both algorithms.

We ran the traditional LinUCB and report the cumulative mean reward of the male user group and female user group over time. Figure 4.3 shows the impact of the order of the data on the performances. Specifically, (a)LinUCB: 1,500 females followed by 1,000 males. (b)LinUCB: 1,500 males followed by 1,000 females. (c)FairLinUCB with $\gamma = 3$: 1,500 females followed by 1,000 males. (d)FairLinUCB with $\gamma = 3$: 1,500 males followed by 1,000 females. As shown in Figure 4.3a and Figure 4.3b, overall the male group gets a higher cumulative mean reward than the female group. Particularly, the male group achieves 0.822 against 0.816 for the female group in Figure 4.3a and 0.834 against 0.795 in Figure 4.3b. However, we notice that the reward discrepancy is much higher in the second scenario as compared to the first one. From Figure 4.3a, it appears that learning to recommend videos to all females

students prior to recommending videos to any male students affects the recommendation process positively (i.e., it yields a higher cumulative mean reward for the female group). Thus, the order of the training data can sometimes affect the recommendation process of LinUCB, which can impact the recommendation outcomes and may also induce discrimination towards one group.

We ran the same experiments with Fair-LinUCB, using a γ coefficient of 3, and we report our results in Figure 4.3c and Figure 4.3d. We notice that in both situations our Fair-LinUCB remains very fair, that is, we do not observe a cumulative mean reward discrepancy between the male and female user group. In the former setting, both groups achieve a cumulative mean reward of 0.802 against 0.789 in the latter, both yielding a cumulative mean reward difference of 0.00. In addition, we notice that regardless of the order of the training data our Fair-LinUCB performs equivalently in both scenarios. However, the gain in fairness also induces a loss of utility. Indeed, in the first setting LinUCB achieves 0.052 utility loss against 0.070 for Fair-LinUCB. In the second setting, LinUCB achieves 0.057 against 0.082 for Fair-LinUCB. Thus, our results indicate that Fair-LinUCB is able to close the reward discrepancy and is robust against scenarios that might otherwise induce unfairness.

4.4 Summary

Previous research have shown that personalized recommendation can be highly effective at a cost of introducing unfairness. In this chapter, we have proposed a fair contextual bandit algorithm for personalized recommendation. While current research in fair recommendation mainly focus on how to achieve fairness on the items that are being recommended, our work differs by focusing on fairness on the individuals whom are being recommended an

item. Specifically, we aim to recommend items to users while insuring that both the protected group and privileged group improve their learning performance equally. Our developed Fair-LinUCB improves upon the state-of-the-art LinUCB algorithm by automatically detecting unfairness, and adjusting its arm-picking strategy such that it maximizes the fairness outcome. We further provided a regret analysis of our fair contextual bandit algorithm and demonstrate that the regret bound is only worse than LinUCB up to an additive constant. Finally, we evaluate the performances of our Fair-LinUCB against that of LinUCB by comparing both their effectiveness and degree of fairness. Experimental evaluations showed that our Fair-LinUCB achieves competitive effectiveness while outperforming LinUCB in terms of fairness. We further showed that our algorithm is robust against numerous factors that would otherwise induce or increase discrimination in the traditional LinUCB algorithm. The early version of this work is published at BigData 2021 [87] and HCIS 2022 [88].

5 Achieving User-side Counterfactual Fairness in Bandit-based Recommendation

5.1 Introduction

Fairness in machine learning has been a research subject with rapid growth recently. Although there are many works focusing on fairness in personalized recommendation [25, 26, 27], how to achieve individual fairness in bandit recommendation still remains a challenging task. We focus on online recommendation, e.g., customers are being recommended items, and consider the setting where customers arrive in a sequential and stochastic manner from an underlying distribution and the online decision model recommends a chosen item for each arriving individual based on some strategy. The challenge here is how to choose the arm at each step to maximize the expected reward while achieving user-side fairness for customers, i.e., customers who share similar profiles will receive similar rewards regardless of their sensitive attributes and items being recommended.

Recently researchers have started taking fairness and discrimination into consideration in the design of personalized recommendation algorithms [25, 26, 27, 40, 54, 55, 56, 57, 58]. Among them, [40] was the first paper of studying fairness in classic and contextual bandits. It defined fairness with respect to one-step rewards and introduced a notion of meritocratic fairness, i.e., the algorithm should never place higher selection probability on a less qualified arm (e.g., job applicant) than on a more qualified arm. The following works along this direction include [54] for infinite and contextual bandits, [55] for reinforcement learning, [26] for the simple stochastic bandit setting with calibration based fairness. However,

all existing works require some fairness constraint on arms at every round of the learning process, which is different from our user-side fairness setting. One recent work [88] focused on achieving user-side fairness in bandit setting, but it only purposed a heuristic way to achieve correlation based group level fairness and didn't incorporate causal inference and counterfactual fairness into bandits.

By incorporating causal inference into bandits, we first propose the d-separation based upper confidence bound bandit algorithm (D-UCB), based on which we then propose the fair causal bandit (F-UCB) for achieving the counterfactual individual fairness. Our work is inspired by recent research on causal bandits [89, 90, 91, 92, 93], which studied how to learn optimal interventions sequentially by representing the relationship between interventions and outcomes as a causal graph along with associated conditional distributions. For example, [93] developed the causal UCB (C-UCB) that exploits the causal relationships between the reward and its direct parents. However, different from previous works, our algorithms adopt soft intervention [94] to model the arm selection strategy and leverage the d-separation set identified from the underlying causal graph, thus greatly reducing the amount of exploration needed to achieve low cumulative regret. We show that our D-UCB achieves $\tilde{O}(\sqrt{|\mathbf{W}| \cdot T})$ regret bound where T is the number of iterations and $\mathbf{W}$ is a set that d-separates arm/user features and reward R in the causal graph. As a comparison, the C-UCB achieves $\tilde{O}(\sqrt{|Pa(R)| \cdot T})$ where $Pa(R)$ is the parental variables of R that is a trivial solution of the d-separation set. In our F-UCB, we further achieve counterfactual fairness in each round of exploration. Counterfactual fairness requires the expected reward an individual would receive keeps the same if the individual's sensitive attribute were changed to its counterpart. The introduced counterfactual reward combines two interventions, a soft intervention on the arm selection and a hard intervention on the sensitive attribute. The

F-UCB achieves counterfactual fairness in online recommendation by picking arms from a subset of arms at each round in which all the arms satisfy counterfactual fairness constraint. Our theoretical analysis shows F-UCB achieves $\tilde{O}(\frac{\sqrt{|\mathbf{W}|T}}{\tau-\Delta_{\pi_0}})$ cumulative regret bound where τ is the fairness threshold and Δ_{π_0} denotes the maximum fairness discrepancy of a safe policy π_0, i.e., a policy that is fair across all rounds.

We conduct experiments on the Email Campaign data [93] whose results show the benefit of using the d-separation set from the causal graph. Our D-UCB incurs less regrets than two baselines, the classic UCB which does not leverage any causal information as well as the C-UCB. In addition, we validate numerically that our F-UCB maintains good performance while satisfying counterfactual individual fairness in each round. On the contrary, the baselines fail to achieve fairness with significant percentages of recommendations violating fairness constraint. We further conduct experiments on the Adult-Video dataset and compare our F-UCB with another user-side fair bandit algorithm Fair-LinUCB [88]. The results demonstrate the advantage of our causal based fair bandit algorithm on achieving individual level fairness in online recommendation.

5.2 Achieving Counterfactual Fairness in Bandit

In this section, we present our D-UCB and F-UCB bandit algorithms. The online recommendation is commonly modeled as a contextual multi-armed bandit problem, where each customer is a "bandit player", each potential item a has a feature vector $\mathbf{a} \in \mathcal{A}$ and there are a total number of k items[1]. For each customer arrived at time $t \in [T]$ with feature vector $\mathbf{x}_t \in \mathcal{X}$, the algorithm recommends an item with features $\mathbf{a}$ based on vector $\mathbf{x}_{t,a}$

[1]We use $\mathbf{a}$ to represent the feature vector of item/arm a, and they may be used interchangeably when the context is unambiguous.

which represents the concatenation of the user and the item feature vectors ($\mathbf{x}_t$, $\mathbf{a}$), observes the reward r_t (e.g., purchase), and then updates its recommendation strategy with the new observation. There may also exist some intermediate features (denoted by $\mathbf{I}$) that are affected by the recommended item and influence the reward, such as the user feedback about relevance and quality.

5.2.1 Modeling Arm Selection via Soft Intervention

In bandit algorithms, we often choose an arm that maximizes the expectation of the conditional reward, $a_t = \text{argmax}_a \, \mathbb{E}[R|\mathbf{x}_{t,a}]$. The arm selection strategy could be implemented by a functional mapping from $\mathcal{X}$ to $\mathcal{A}$, and after each round the parameters in the function get updated with the newest observation tuple.

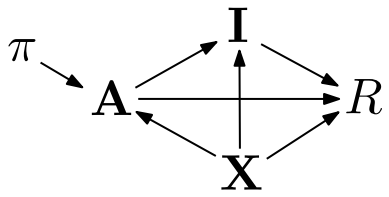


Figure 5.1: Graph structure for contextual bandit recommendation. Node π denotes the soft intervention conducted on arm selection.

We advocate the use of the causal graph and soft interventions as a general representation of any bandit algorithm. We consider the causal graph $\mathcal{G}$, e.g., as shown in Figure 5.1, where $\mathbf{A}$ represents the arm features, $\mathbf{X}$ represents the user features, R represents the reward, and $\mathbf{I}$ represents some intermediate features between $\mathbf{A}$ and R. Since the arm selection process could be regarded as the structural equation of $\mathbf{X}$ on $\mathbf{A}$, we treat $\mathbf{X}$ as $\mathbf{A}$'s parents. Then, the reward R is influenced by the arm selection, the contextual user features, as well as some intermediate features, so all the three factors are parents of R. In this setting, it is natural to treat the update of the arm selection policy as a soft intervention π performed on

the arm features $\mathbf{A}$. Each time when an arm selection strategy is learned, the corresponding soft intervention is considered to be conducted on $\mathbf{A}$ while user features $\mathbf{X}$ and all other relationships in the causal graph are unchanged.

There are several advantages of modeling arm selection learning using the soft intervention. First, it can capture the complex causal relationships between context and reward without introducing strong assumptions, e.g., linear reward function, or Gaussian/Bernoulli prior distribution, which are often not held in practice. Second, it is flexible in terms of the functional form. For example, it can be of any function type, and it can be independent or dependent upon the target variable's existing parents and can also include new variables that are not the target variable's parents. Third, the soft intervention can be either deterministic, i.e., fixing the target variable to a particular constant, or stochastic, i.e., assigns to the target variable a distribution with probabilities over multiple states. As a result, most existing and predominant bandit algorithms could be described using this framework. Moreover, based on this framework we could propose new bandit algorithms by adopting different soft interventions.

Formally, let Π_t be the arm selection policy space at time $t \in [T]$, and $\pi \in \Pi_t$ be a specific policy. The implementation of policy π is modeled by a soft intervention. Denoting by $R(\pi)$ the post-interventional value of the reward after performing the intervention, the expected reward under policy π, denoted by μ_π, is given by $\mathbb{E}[R(\pi)|\mathbf{x}_t]$. According to the σ-calculus [94], it can be further decomposed as follows:

$$\mu_\pi = \mathbb{E}[R(\pi)|\mathbf{x}_t] = \sum_{\mathbf{a}} P_\pi(\mathbf{a}|\mathbf{x}_t) \cdot \mathbb{E}[R(\mathbf{a})|\mathbf{x}_t] = \mathbb{E}_{\mathbf{a}\sim\pi}\left[\mathbb{E}[R(\mathbf{a})|\mathbf{x}_t]\right] \tag{5.1}$$

where $P_\pi(\mathbf{a}|\mathbf{x}_t)$ is a distribution defined by policy π. As can be seen, once a policy is given,

the estimation of μ_π depends on the estimation of $\mathbb{E}[R(\mathbf{a})|\mathbf{x}_t]$ (denoted by μ_a). Note that μ_a represents the expected reward when selecting an arm a, which is still a post-intervention quantity and needs to be expressed using observational distributions in order to be computable. In the following, we propose a d-separation based estimation method and based on which we develop our D-UCB algorithm. For the ease of representation, our discussions in the following subsections assume deterministic policies, in principle the above framework could be applied to stochastic policies as well.

5.2.2 D-UCB Algorithm

Let $\mathbf{W} \subseteq \mathbf{A} \cup \mathbf{X} \cup \mathbf{I}$ be a subset of nodes that d-separates reward R from features $(\mathbf{A} \cup \mathbf{X}) \backslash \mathbf{W}$ in the causal graph. Such set always exists since $\mathbf{A} \cup \mathbf{X}$ and $Pa(R)$ are trivial solutions. Let $\mathbf{Z} = \mathbf{W} \backslash (\mathbf{A} \cup \mathbf{X})$. Using the do-calculus [21], we can decompose μ_a as follows.

$$\begin{aligned}
\mu_a &= \mathbb{E}[R|do(\mathbf{a}), \mathbf{x}_t] = \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{z}, do(\mathbf{a}), \mathbf{x}_t] P(\mathbf{z}|do(\mathbf{a}), \mathbf{x}_t) \\
&= \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{z}, \mathbf{a}, \mathbf{x}_t] P(\mathbf{z}|\mathbf{a}, \mathbf{x}_t) = \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{z}, \mathbf{a}, \mathbf{x}_t] P(\mathbf{z}|\mathbf{x}_{t,a}) \\
&= \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{w}] P(\mathbf{z}|\mathbf{x}_{t,a}) \qquad (5.2)
\end{aligned}$$

where the last step is due to the d-separation. Similar to [93], we assume that distribution $P(\mathbf{z}|\mathbf{x}_{t,a})$ is known based on previous knowledge used to build the causal graph. Then, by using a sample mean estimator (denoted by $\hat{\mu}_{\mathbf{w}}(t)$) to estimate $\mathbb{E}[R|\mathbf{w}]$ based on the

observational data up to time t, the estimated reward mean is given by

$$\hat{\mu}_{\pi}(t) = \mathbb{E}_{\mathbf{a}\sim\pi}\left[\sum_{\mathbf{Z}} \hat{\mu}_{\mathbf{w}}(t) \cdot P(\mathbf{z}|\mathbf{x}_{t,a})\right] \tag{5.3}$$

Subsequently, we propose a causal bandit algorithm based on d-separation, called D-UCB. Since there is always uncertainty on the reward given a specific policy, in order to balance exploration and exploitation we follow the rule of optimistic in the face of uncertainty (OFU) in D-UCB algorithm. The policy taken at time t will lead to the highest upper confidence bound of the expected reward, which is given by

$$\pi_t = \underset{\pi\in\Pi_t}{\operatorname{argmax}}\, \mathbb{E}_{\mathbf{a}\sim\pi}[UCB_a(t)] \tag{5.4}$$

$$UCB_a(t) = \sum_{\mathbf{Z}} UCB_{\mathbf{w}}(t) P(\mathbf{z}|\mathbf{x}_{t,a}) \tag{5.5}$$

Since $\hat{\mu}_{\mathbf{w}}(t)$ is an unbiased estimator and the error term of the reward is assumed to be sub-Gaussian distributed, the $1-\delta$ upper confidence bound of $\mu_{\mathbf{w}}(t)$ is given by

$$UCB_{\mathbf{w}}(t) = \hat{\mu}_{\mathbf{w}}(t) + \sqrt{\frac{2\log(1/\delta)}{1 \vee N_{\mathbf{w}}(t)}} \tag{5.6}$$

After taking the policy, we will have new observations on r_t and $\mathbf{w}_t$. The sample mean estimator is then updated accordingly:

$$\hat{\mu}_{\mathbf{w}}(t) = \frac{1}{T_{\mathbf{w}}(t)} \sum_{k=1}^{t} r_t \mathbb{1}_{\mathbf{w}_k=\mathbf{w}} \quad \text{where} \quad T_{\mathbf{w}}(t) = \sum_{k=1}^{t} \mathbb{1}_{\mathbf{w}_k=\mathbf{w}} \tag{5.7}$$

We hypothesize that the choice of d-separation set $\mathbf{W}$ would significantly affect the

Algorithm 3 D-UCB: Causal Bandit based on d-separation

1: Input: Policy space Π, confidence level parameter δ, original causal Graph $\mathcal{G}$ with domain knowledge
2: Find the d-separation set $\mathbf{W}$ with minimum subset $\mathbf{Z}$ in terms of domain space.
3: **for** $t = 1, 2, 3, ..., T$ **do**
4: Obtain the optimal policy π_t following Eq. (5.4).
5: Take action $\mathbf{a}_t \sim \pi_t$ and observe a real-valued payoff r_t and a d-separation set value $\mathbf{w}_t$.
6: Update $\hat{\mu}_{\mathbf{w}}(t)$ for all $\mathbf{w} \in \mathbf{W}$ following Eq. (5.7).
7: **end for**

regret of the D-UCB. To this end, we analyze the upper bound of the cumulative regret $\mathcal{R}_T$. The following theorem shows that, the regret upper bound depends on the domain size of d-separation set $\mathbf{W}$.

Theorem 5 (Regret bound of D-UCB)**.** Given a causal graph $\mathcal{G}$, with probability at least $1 - 2\delta T|\mathbf{W}| - \exp(-\frac{|\mathbf{W}|\log^3(T)}{32\log(1/\delta)})$, the regret of D-UCB is bounded by

$$\mathcal{R}_T \leq \sqrt{|\mathbf{W}|T\log(T)}log(T) + \sqrt{32|\mathbf{W}|T\log(1/\delta)}$$

where $|\mathbf{W}|$ is the domain space of set $\mathbf{W}$.

Proof Sketch. The proof of Theorem 5 follows the general regret analysis framework of the UCB algorithm [95]. By leveraging d-separation decomposition of the expected reward, we split the cumulative regret into two terms and bound them separately. Since there are less terms to traverse when summing up and bounding the uncertainty caused by exploration-exploitation strategy, D-UCB is supposed to obtain lower regret than the original UCB algorithm and C-UCB algorithm. By setting $\delta = 1/T^2$, it is easy to show that D-UCB algorithm achieves $\tilde{O}(\sqrt{|\mathbf{W}| \cdot T})$ regret bound. Refer to Appendix for proof details. □

Algorithm 3 shows the pseudo code of the D-UCB. In Line 2, according to Theorem 5, we first determine the d-separation set $\mathbf{W}$ with the minimum domain space. In Line 4

we leverage causal graph and the observational data up to time t to find the optimal policy $\pi_t = \text{argmax}_{\pi \in \Pi_t} \mathbb{E}_{\mathbf{a} \sim \pi}[UCB_a(t)]$. In Line 5, we take action $\mathbf{a}_t \sim \pi_t$ and observe a real-valued payoff r_t, and in Line 6, we update the observational data with $\mathbf{a}_t$ and r_t.

Remark. Determining the minimum d-separation set has been well studied in causal inference [96]. We leverage the algorithm of finding a minimum cost separator [97] to identify $\mathbf{W}$. The discovery procedure usually requires the complete knowledge of the causal graph. However, in the situation where the d-separation set to be used as well as the associated conditional distributions $P(\mathbf{z}|\mathbf{x}_{t,a})$ are given, the remaining part of the algorithm will work just fine without the causal graph information. Moreover, the assumption of knowing $P(\mathbf{z}|\mathbf{x}_{\mathbf{t},\mathbf{a}})$ follows recent research works on causal bandit. Generalizing the causal bandit framework to partially/completely unknown causal graph setting is a much more challenging but important task. A recent work [98] tried to generalize causal bandit algorithm based on causal trees/forests structure.

To better illustrate the long-term regret of causal bandit algorithm, suppose the set $\mathbf{A} \cup \mathbf{U} \cup \mathbf{I}$ includes N variables that are related to the reward and the d-separation set $\mathbf{W}$ includes n variables. If each of the variable takes on 2 distinct values, the number of deterministic policies can be as large as 2^N for traditional bandit algorithm, leading to a $\mathcal{O}(\sqrt{2^N T})$ regret bound. On the other hand, our proposed causal algorithms exploit the knowledge of the d-separation set $\mathbf{W}$ and achieves $\mathcal{O}(\sqrt{2^n T})$ regret, which implies a significant reduction regarding to the regret bound if $n << N$. If the number of arm candidates is much smaller than the domain space of $\mathbf{W}$, our bound analysis could be easily adjusted to this case using a subspace of $\mathbf{W}$ that corresponds to the arm candidates.

5.2.3 Counterfactual Fairness

Now, we are ready to present our fair UCB algorithm. Rather than focusing on the fairness of the item being recommended (e.g., items produced by small companies have similar chances of being recommended as those from big companies), we focus on the user-side fairness in terms of reward, i.e., individual users who share similar profiles will receive similar rewards regardless of their sensitive attributes and items being recommended such that they both benefit from the recommendations equally. To this end, we adopt counterfactual fairness as our fairness notion.

Consider a sensitive attribute $S \in \mathbf{X}$ in the user's profile. Counterfactual fairness concerns the expected reward an individual would receive assuming that this individual were in different sensitive groups. In our context, this can be formulated as the counterfactual reward $\mathbb{E}[R(\pi, s^*)|\mathbf{x}_t]$ where two interventions are performed simultaneously: soft intervention π on the arm selection and hard intervention $do(s^*)$ on the sensitive attribute S, while conditioning on individual features $\mathbf{x}_t$. Denoting by $\Delta_\pi = \mathbb{E}[R(\pi, s^+)|\mathbf{x}_t] - \mathbb{E}[R(\pi, s^-)|\mathbf{x}_t]$ the counterfactual effect of S on the reward, a policy that is counterfactually fair is defined as follows.

Definition 6. A policy π is counterfactually fair for an individual arrived if $\Delta_\pi = 0$. The policy is τ- counterfactually fair if $|\Delta_\pi| \leq \tau$ where τ is the predefined fairness threshold.

To achieve counterfactual fairness in online recommendation, at round t, we can only pick arms from a subset of arms for the customer (with feature $\mathbf{x}_t$), in which all the arms satisfy counterfactual fairness constraint. The fair policy subspace $\Phi_t \subseteq \Pi_t$ is thus given by $\Phi_t = \{\pi : \Delta_\pi \leq \tau\}$.

However, the counterfactual fairness is a causal quantity that is not necessarily uniden-

tifiable from observational data without the knowledge of structure equations [99]. In [9], the authors studied the criterion of identification of counterfactual fairness given a causal graph and provided the bounds for unidentifiable counterfactual fairness. According to Proposition 1 in [9], our counterfactual fairness is identifiable if $\mathbf{X}\backslash\{S\}$ are not descendants of S. In this case, similar to Eq. (5.1), we have that $\mathbb{E}[R(\pi, s^*)|\mathbf{x}_t] = \mathbb{E}_{\mathbf{a}\sim\pi}[\mathbb{E}[R(\mathbf{a}, s^*)|\mathbf{x}_t]]$ where $s^* \in \{s^+, s^-\}$. Similar to Eq. (5.2), we denote $\mu_{a,s^*} = \mathbb{E}[R(a, s^*)|\mathbf{x}_t]$, which can be decomposed using the do-calculus as

$$\mu_{a,s^*} = \mathbb{E}[R(a, s^*)|\mathbf{x}_t] = \sum_{\mathbf{Z}} \mathbb{E}[R|s^*, \mathbf{w}\backslash s_t] \cdot P(\mathbf{z}|s^*, \mathbf{x}_{t,a}\backslash s_t) \tag{5.8}$$

where $\mathbf{w}\backslash s_t$ and $\mathbf{x}_{t,a}\backslash s_t$ represent all values in $\mathbf{w}$ and $\mathbf{x}_{t,a}$ except s_t respectively. Note that s^* is the sensitive attribute value in the counterfactual world which could be different from the observational value s_t. The estimated counterfactual reward can be calculated as

$$\hat{\mu}_{a,s^*}(t) = \sum_{\mathbf{Z}} \hat{\mu}_{\mathbf{w}^*}(t) \cdot P(\mathbf{z}|s^*, \mathbf{x}_{t,a}\backslash s_t)$$

where $\mathbf{w}^* = \{s^*, \mathbf{w}\backslash s_t\}$ and $\hat{\mu}_{\mathbf{w}^*}(t)$ is again the sample mean estimator based on the observational data up to time t. The estimated counterfactual discrepancy of a policy is

$$\hat{\Delta}_\pi(t) = |\mathbb{E}_{\mathbf{a}\sim\pi}[\hat{\mu}_{a,s^+}(t)] - \mathbb{E}_{\mathbf{a}\sim\pi}[\hat{\mu}_{a,s^-}(t)]| \tag{5.9}$$

In the case where μ_{a,s^*} is not identifiable, based on Proposition 2 in [9] we derive the lower and upper bounds of μ_{a,s^*} as presented in the following theorem. Please refer to Appendix for the proof.

Theorem 6. Given a causal graph as shown in Figure 5.1, if there exists a non-empty set

$\mathbf{B} \subseteq \mathbf{X} \backslash \{S\}$ which are descendants of S, then $\mu_{a,s^*} = \mathbb{E}[R(a, s^*)|\mathbf{x}_t]$ is bounded by

$$\mu_{a,s^*} \leq \sum_{\mathbf{Z}} \max_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{w} \backslash s_t]\} \cdot P(\mathbf{z}|\mathbf{x}_{t,a}) \ ,$$
$$\mu_{a,s^*} \geq \sum_{\mathbf{Z}} \min_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{w} \backslash s_t]\} \cdot P(\mathbf{z}|\mathbf{x}_{t,a})$$

5.2.4 F-UCB Algorithm

Taking the estimation error of the counterfactual discrepancy into consideration, we could also use the high probability upper confidence bound of the counterfactual effect to build the conservative fair policy subspace $\bar{\Phi}_t = \{\pi : UCB_{\Delta_\pi}(t) \leq \tau\}$ where

$$UCB_{\Delta_\pi}(t) = \hat{\Delta}_\pi(t) + \sum_{\mathbf{Z}} \sqrt{\frac{8 \log(1/\delta)}{1 \vee N_{\mathbf{w}}(t)}} P(\mathbf{z}|\mathbf{x}_{t,a}) \tag{5.10}$$

which is derived based on the fact that the sum of two independent sub-Gaussian random variables is still sub-Gaussian distributed. Thus, the learning problem can be formulated as the following constrained optimization problem:

$$\min \mathcal{R}_T = \sum_{t=1}^{T} \left(\mathbb{E}_{\mathbf{a} \sim \pi_t^*}[\mu_a] - \mathbb{E}_{\mathbf{a} \sim \pi_t}[\mu_a] \right) \quad \text{s.t.} \ \ \forall t, \ \pi_t \in \bar{\Phi}_t$$

where π_t^* is defined as the optimal policy in the policy space Π_t at each round, which is the same in D-UCB setting. A safe policy π_0 refers to a a feasible solution under the fair policy subspace at each round, i.e., $\pi_0 \in \Pi_t$ such that $\Delta_{\pi_0} \leq \tau$ for each $t \in [T]$.

This optimization can be solved similarly by following the rule of OFU. Algorithm 4

depicts our fair bandit algorithm called the F-UCB. Different from the D-UCB algorithm, F-UCB only picks arm from $\bar{\Phi}_t$ at each time t. In Line 5, we compute the estimated reward mean and the estimated fairness discrepancy. In Line 6, we determine the fair policy subspace $\bar{\Phi}_t$, and in Line 7, we find the optimal policy $\pi_t = \text{argmax}_{\pi \in \bar{\Phi}_t} \mathbb{E}_{\mathbf{a} \sim \pi}[UCB_a(t)]$.

Algorithm 4 F-UCB: Fair Causal Bandit

1: Input: Policy space Π, fairness threshold τ, confidence level parameter δ, original causal Graph $\mathcal{G}$ with domain knowledge
2: Find the d-separation set $\mathbf{W}$ with minimum subset $\mathbf{Z}$ in terms of domain space.
3: **for** $t = 1, 2, 3, ..., T$ **do**
4: **for** $\pi \in \Pi_t$ **do**
5: Compute the estimated reward mean using Eq. (5.3) and the estimated fairness discrepancy using Eq. (5.9).
6: **end for**
7: Determine the conservative fair policy subspace $\bar{\Phi}_t$.
8: Find the optimal policy following Eq. (5.4) within $\bar{\Phi}_t$.
9: Take action $\mathbf{a}_t \sim \pi_t$ and observe a real-valued payoff r_t and a d-separation set value $\mathbf{w}_t$.
10: Update $\hat{\mu}_{\mathbf{w}}(t)$ for all $\mathbf{w} \in \mathbf{W}$.
11: **end for**

The following regret analysis shows that, the regret bound of F-UCB is larger than that of D-UCB as expected, and it is still influenced by the domain size of set $\mathbf{W}$.

Theorem 7 (Regret bound of fair causal bandit)**.** Given a causal graph $\mathcal{G}$, let $\delta_E = 4|\mathbf{W}|T\delta$ and Δ_{π_0} denote the maximum fairness discrepancy of a safe policy π_0 across all rounds. Setting $\alpha_c = 1$ and $\alpha_r = \frac{2}{\tau - \Delta_{\pi_0}}$, with probability at least $1 - \delta_E$, the cumulative regret of F-UCB is bounded by:

$$\mathcal{R}_T \leq (\frac{2}{\tau - \Delta_{\pi_0}} + 1) \times \left(2\sqrt{2T|\mathbf{W}|\log(1/\delta_E)} + 4\sqrt{T\log(2/\delta_E)\log(1/\delta_E)}\right)$$

Proof Sketch. Our derivation of the regret upper bound of F-UCB follows the proof idea of bandits with linear constraints [100], where we treat counterfactual fairness as a linear

constraint. By leveraging the knowledge of a feasible fair policy at each round and properly designing the numerical relation of the scale parameters α_c and α_r, we are able to synchronously bound the cumulative regret of reward and fairness discrepancy term. Merging these two parts of regret analysis together leads to a unified bound of the F-UCB algorithm. By setting δ_E to $1/T^2$ we can show F-UCB achieves $\tilde{O}(\frac{\sqrt{|\mathbf{W}|T}}{\tau-\Delta_{\pi_0}})$ long-term regret. The detailed proof is reported in Appendix. □

Remark. In Theorem 7, α_c and α_r refer to the scale parameters that control the magnitude of the confidence interval for sample mean estimators related to reward and fairness term respectively. The values taken in Theorem 7 is one feasible solution with α_c taking the minimum value under the constraint domain space.

The general framework we proposed (Eq. (5.1)) can be applied to any policy/function class. However, the D-UCB and F-UCB algorithms we proposed still adopt the deterministic policy following the classic UCB algorithm. Thus, the construction of $\bar{\Phi}_t = \{\pi : UCB_{\Delta_\pi}(t) \leq \tau\}$ can be easily achieved as the total number of policies are finite. In this chapter we also assume discrete variables, but in principle the proposed algorithms can also be extended to continuous variables by employing certain approximation approaches, e.g., neural networks for estimating probabilities and sampling approaches for estimating integrals. However, the regret bound analysis may not apply as $|\mathbf{W}|$ will become infinite in the continuous space.

5.3 Experiment

In this section, we conduct experiments on two datasets and compare the performance of D-UCB and F-UCB with UCB, C-UCB and Fair-LinUCB in terms of the cumulative regret. We also demonstrate the fairness conformance of F-UCB and the violations of other

algorithms.

5.3.1 Email Campaign Dataset

We adopt the Email Campaign data as used in previous works [93]. The dataset is constructed based on the online advertising process. Its goal is to determine the best advertisement recommendation strategy for diverse user groups to improve their click through ratio (CTR), thus optimize the revenue generated through advertisements. We construct the causal graph following the domain knowledge and one of the recent research works on causal bandit [93]. Figure 5.2 shows the topology of the causal graph. We use X_1, X_2, X_3 to denote three user profile attributes, *gender*, *age* and *occupation*; A_1, A_2, A_3 to denote three arm features, *product*, *purpose*, *send-time* that could be intervened; I_1, I_2, I_3, I_4 to denote *Email body template*, *fitness*, *subject length*, and *user query*; and R to denote the reward that indicates whether users click the advertisement. The reward function is $R = 1/12(I_1+I_2+I_3+A_3)+\mathcal{N}(0,\sigma^2)$, where $\sigma = 0.1$. In Figure 5.2, nodes with blue frame denote the variables that can be intervened. The node with red frame is the sensitive attribute. Light shaded nodes denote the minimal d-separation set. In our experiment, we set $\delta = 1/t^2$ for each $t \in [T]$. Specifically, Table 5.1 shows the attributes of Email Campaign data and their domain values. Table 5.2 shows the conditional probabilities of $P(I_4 = i|X_1, X_2, X_3)$. The following equations are the conditional distributions for the remaining variables.

$$P(I_2 = 1|A_1, A_2, I_4) = (A_1 + A_2 + I_4)/12 \qquad P(I_1 = 1|A_1, A_2, I_2) = (A_1 + A_2 + I_2)/10$$

$$P(I_3 = 1|I_2 = 1) = 0.4 \qquad P(I_3 = 1|I_2 = 2) = 0.6$$

Table 5.1: Variables in Email campaign data.

Variables	Domain Value
Click (R)	$(0, 1)$
Gender (X_1)	$(1, 2)$
Age (X_2)	$(1, 2)$
Occupation (X_3)	$(1, 2)$
Product (A_1)	$(1, 2, 3)$
Propose (A_2)	$(1, 2, 3, 4)$
Send time (A_3)	$(1, 2, 3)$
Email body template (I_1)	$(1, 2)$
Fitness (I_2)	$(1, 2)$
Subject length (I_3)	$(1, 2, 3, 4)$
User query (I_4)	$(1, 2)$

Table 5.2: Conditional probabilities of $P(I_4 = i|X_1, X_2, X_3)$.

(X_1, X_2, X_3)	$i = 1$	$i = 2$	$i = 3$	$i = 4$
(0,0,0)	0.4	0.3	0.2	0.1
(0,0,1)	0.3	0.4	0.2	0.1
(0,1,0)	0.6	0.1	0.2	0.1
(0,1,1)	0.5	0.2	0.2	0.1
(1,0,0)	0.1	0.3	0.2	0.4
(1,0,1)	0.1	0.4	0.2	0.3
(1,1,0)	0.1	0.1	0.2	0.6
(1,1,1)	0.1	0.2	0.2	0.5

Figure 5.3 plots the cumulative regrets of different bandit algorithms along T. For each bandit algorithm, the online learning process starts from initialization with no previous observation. Figure 5.3 shows clearly all three causal bandit algorithms perform better than UCB. This demonstrates the advantage of applying causal inference in bandits. Moreover, our D-UCB and F-UCB outperform C-UCB, showing the advantage of using d-separation set in our algorithms. The identified d-separation set $\mathbf{W}$ (*send time*, *fitness*, and *template*) and the domain space of $\mathbf{Z}$ (*fitness* and *template*) significantly reduce the exploration cost in D-UCB and F-UCB.

Remark. Note that in Figure 5.3, for the first 2000 rounds, F-UCB has lower cumulative regret than D-UCB. A possible explanation is that fair constraint may lead to a policy sub-

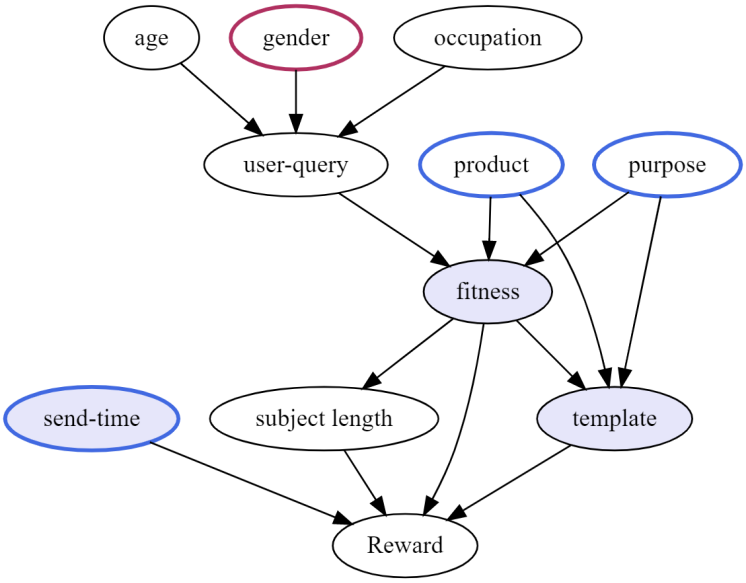


Figure 5.2: Graph structure under Email Campaign data.

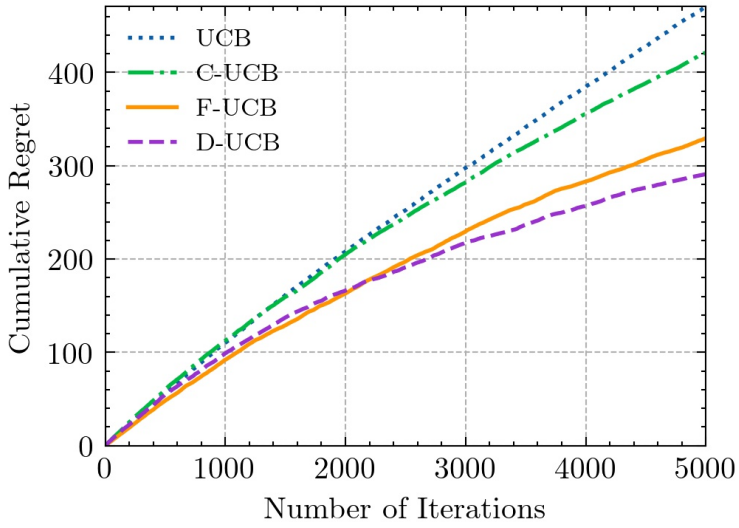


Figure 5.3: Comparison of bandit algorithms ($\tau = 0.3$ for F-UCB).

space that contains many policies with high reward. As the number of explorations increase, D-UCB gains more accurate reward estimations for each policy in the whole policy space and eventually outperforms F-UCB.

Table 5.3 shows how the cumulative regret of F-UCB ($T = 5000$ rounds) varies with the fairness threshold τ. The values in Table 5.3 (and Table 5.4) are obtained by averaging

Table 5.3: Comparison results for Email campaign data.

τ	Cumulative Regret of F-UCB	Unfair Decisions			
		UCB	C-UCB	D-UCB	F-UCB
0.1	392.12	3030	3176	3473	0
0.2	363.55	1383	1487	1818	0
0.3	355.21	482	594	739	0
0.4	317.80	141	185	234	0
0.5	313.89	18	27	47	0

the results over 5 trials. The larger the τ, the smaller the cumulative regret. In the right block of Table 5.3, we further report the number of fairness violations of the other three algorithms during the exploration of $T = 5000$ rounds, which demonstrates the need of fairness aware bandits. In comparison, our F-UCB achieves strict counterfactual fairness in every round.

5.3.2 Adult-Youtube Video Dataset

Following the setting of [88], we generate one simulated dataset for our experiments by combining the following two publicly available datasets.

- **Adult dataset**: The Adult dataset [86] is used to represent the students (or bandit players). It is composed of 31,561 instances: 21,790 males and 10,771 females, each having 8 categorical variables (work class, education, marital status, occupation, relationship, race, sex, native-country) and 3 continuous variables (age, education number, hours per week). We select 4 variables, *age, sex, race, income*, as user features in our experiments and binarize their domain values due to data sparsity issue.

- **YouTube dataset**: The Statistics and Social Network of YouTube Videos [2] dataset is used to represent the items to be recommended (or arms). It is composed of 1,580 instances each having 6 categorical features (age of video, length of video, number

[2]https://netsg.cs.sfu.ca/youtubedata/

of views, rate, ratings, number of comments). We select four of those variables (age, length, ratings, comments) and binarize them for a suitable size of the arm pool.

For our experiments, we use a subset of 10,000 random instances from the Adult dataset, which is then split into two subsets: one for graph construction and the other for online recommendation. Similarly, a subset of YouTube dataset is used as our pool of videos to recommend. The subset contains 16 video types (arms) representing different domain values of the 4 binarized arm features.

The feature contexts $\mathbf{x_{t,a}}$ used throughout the experiment is the concatenation of both the student feature vector and the video feature vector. Four elements in $\mathbf{x_{t,a}}$ are selected according to domain knowledge as the variables that will determine the value of the reward. A linear reward function is then applied to build this mapping relation from those selected variables to the reward variable. In our experiments we choose the sensitive attribute to be the **gender of adults**, and focus on the individual level fairness discrepancy regarding to both male and female individuals. For the Adult-Video experiment setting, we construct the causal graph using a causal discovery software Tetrad (https://www.ccd.pitt.edu/tools/).

We further compare the performance of F-UCB algorithm with Fair-LinUCB [88] on Adult-Video dataset. We select 10,000 instances and use half of the data as the offline data to construct causal graph and adopt the other half to be user sequence and arm candidates for online recommendation. The causal graph constructed from the training data is shown in Figure 5.4, where $\mathbf{X} = \{age, sex, race, income\}$ denote user features, $\mathbf{A} = \{length, ratings, views, comments\}$ denote video features. Bold nodes denote direct parents of the reward and red nodes denote the sensitive attribute. The minimum d-separation set for this graph topology is $\mathbf{W} = \{$*age, income, ratings, views*$\}$. The reward function is set as $R = 1/5(age + income + ratings + views) + \mathcal{N}(0, \sigma^2)$, where $\sigma = 0.1$. We

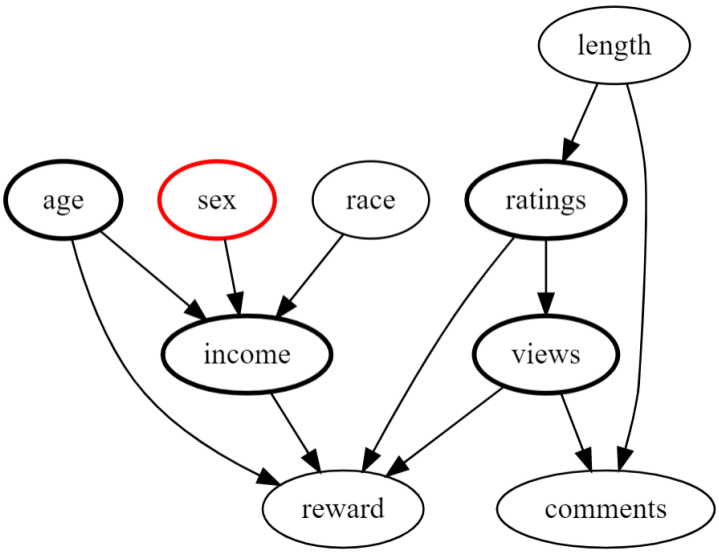


Figure 5.4: Graph structure for Adult-Video data.

Table 5.4: Comparison results for Adult-Video data.

τ	Regret	Unfair Decisions	
	F-UCB	F-UCB	Fair-LinUCB
0.1	361.43	0	2053
0.2	332.10	0	1221
0.3	323.12	0	602
0.4	303.32	0	82
0.5	296.19	0	6

set $\delta = 1/t^2$ for each $t \in [T]$. The cumulative regret is added up through 5000 rounds.

We observe from Table 5.4 a high volume of unfair decisions made by Fair-LinUCB under strict fairness threshold (nearly forty percent of the users are unfairly treated when $\tau = 0.1$). This implies Fair-LinUCB algorithm can not achieve individual level fairness when conducting online recommendation compared to F-UCB. On the other hand, the cumulative regret for Fair-LinUCB is around 250 over 5000 rounds, which is slightly better than F-UCB. This is because we use the same linear reward setting as [88] in our experiment and Lin-UCB based algorithm will better catch the reward distribution under this setting.

5.4 Summary

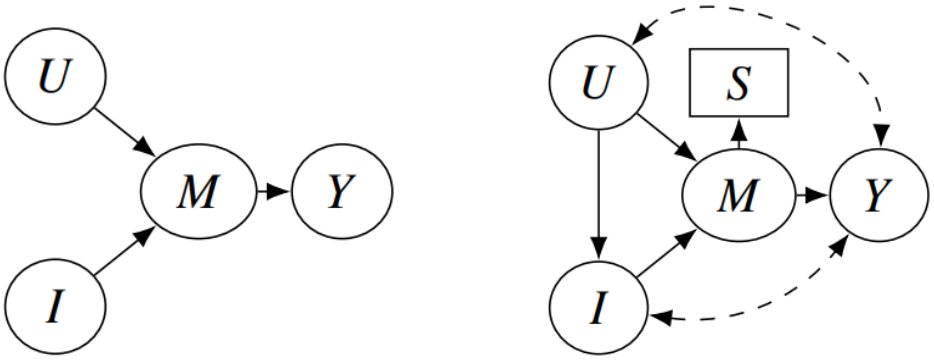
In this chapter, we studied how to learn optimal interventions sequentially by incorporating causal inference in bandits. We developed D-UCB and F-UCB algorithms which leverage the d-separation set identified from the underlying causal graph and adopt soft intervention to model the arm selection strategy. Our F-UCB further achieves counterfactual individual fairness in each round of exploration by choosing arms from a subset of arms satisfying counterfactual fairness constraint. Our theoretical analysis and empirical evaluation show the effectiveness of our algorithms against baselines. The early version of this work is published at AAAI 2022 [101].

6 Dealing with Confounding and Sample Selection Biases in Recommendation

6.1 Introduction

Recommender systems provide personalized services for users seeking information and play an increasingly important role in online applications. However, the user-item interaction data, which are used to train recommender systems and then generated by the deployed systems, often have both selection and confounding biases. The confounding bias arises when hidden variables determine user/item features and an outcome variable simultaneously. For example, popularity bias is one classic instance of confounding bias. It occurs when items are over-displayed and therefore have more chances to be seen as well as clicked by users. Under popularity bias, the click through rate (CTR) of the users does not accurately reflect the users' true preference on an over-exposed item. Additionally, the selection mechanism, e.g., choosing users based on a certain time or location, can lead to sample selection bias. Several attempts have been made to alleviate such biases from both causal and counterfactual inference perspectives [28, 29, 30]. However, previous work has focused on dealing with one specific source of bias rather than handling multiple sources simultaneously. Neglecting the presence of both confounding and sample selection biases leads to poor recommendation performance.

Figure 6.1 shows abstract causal graph structures of a recommendation system under confounding and selection biases. Our formulation is based on the structural causal model (SCM) [21], which describes the causal mechanisms of a system as a set of nodes and struc-



(a) (b)

Figure 6.1: Abstract causal graph structures of an online recommendation system. (a)Conventional Recommender. (b)Biased Recommender.

tural equations. Figure 6.1a illustrates the graph structure of a traditional recommendation model, where **U** represents user features, **I** represents item features, **M** represents a matching score model, and a binary outcome variable Y represents user clicks, e.g., $Y = 1$ for clicked and $Y = 0$ for not clicked. Figure 6.1b demonstrates the graph structure of a recommendation model under confounding and selection biases. The dashed bi-directed edges indicate the existence of the common cause factors of user/item features and user clicks, which might be unobserved. Node S depicts the (biased) selection mechanism of the recommended item. Specifically, $S = 1$ indicates the tuple is selected into the observational data, and $S = 0$ otherwise.

In causal inference based personalized recommendation, one core problem is to derive conditional causal effects in the form of $P(Y = y|do(\mathbf{I} = \mathbf{i}), \mathbf{U} = \mathbf{u})$, which represents the causal effect of the intervention (recommending item with features **i**) on the outcome y for a user with features **u**. Note that the *do* operator simulates the interventions or treatments that force item attributes **I** to take certain values **i**. The intervention/treatment in the recommender system can be thought of as referring to a recommendation strategy, i.e., how the system selects, organizes, and shows item features to some specific users.

Two of the most common obstacles in deriving conditional causal effects are confounding and selection biases. Confounding bias often arises from lack of control over the decision making process. Formally, it denotes the difference between the interventional quantity $P(y|do(\mathbf{i}), \mathbf{u})$ and its probabilistic counterpart $P(y|\mathbf{i}, \mathbf{u})$. In practice, the most common method for controlling confounding bias is the backdoor adjustment [21]. Selection bias arises because of preferential exclusion of samples, which makes the observed data no longer a true representation of the underlying population. As a result, biased estimates of the causal effect will be produced. To correct selection bias, we must derive unbiased estimates from the observed distribution $P(\mathbf{V}|S = 1)$ instead of from $P(\mathbf{V})$ (which is unavailable), where $\mathbf{V}$ denotes the observed variables in a causal graph.

In this chapter, we formulate both confounding and selection biases and show that they can be separately mitigated by conditioning on a bias adjustment set that satisfies certain criteria. In particular, we show $P(y|do(\mathbf{i}), \mathbf{u})$ can be identified and recovered when unbiased external data over a subset of variables $\mathbf{T}$ (i.e., $P(\mathbf{T})$) is available. We develop a debiased recommendation algorithm, called dREC, within the scope of the structural causal model that can achieve accurate prediction under the presence of both biases. We also present a general statistical procedure based on inverse probability weighting (IPW) to estimate the adjustment formula from the biased data.

We further show under the presence of confounding and selection biases how to derive path-specific effects, i.e., the effect of changing $\mathbf{I}$ from $\mathbf{i}_1$ to $\mathbf{i}_2$ on y transmitted along a path set π, and counterfactual effect, i.e., the probability that event $Y = y$ would be observed had $\mathbf{I}$ been $\mathbf{i}$, given that we actually observed $\mathbf{I}$ to be $\mathbf{i}'$. We compare the performance of our debiased approach with baselines with datasets under different biased selection settings. Experimental results show the effectiveness of our approach.

6.2 Debiased Recommendation

6.2.1 Overview

In personalized recommendation, given a user u with features $\mathbf{u}$, we aim to recommend an item with features $\mathbf{i}$ such that the expected reward is maximized. Formally, we have $\mathbf{i} = \text{argmax}_{\mathbf{i} \in D_{\mathcal{I}}} P(y|do(\mathbf{i}), \mathbf{u})$. Conditional causal effects can be calculated based on the backdoor criteria when there is no sample selection bias.

$$P(y|do(\mathbf{i}), \mathbf{u}) = \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, \mathbf{u}) P(\mathbf{z}|\mathbf{u}) \tag{6.1}$$

However, the challenge is to estimate $P(y|do(\mathbf{i}), \mathbf{u})$ from a selection-biased distribution. In this section, we first give an overview of our algorithm. We then present our derived selection-backdoor criterion and procedure of identifying an intervention set and adjustment pair in Section 6.2.2. We describe an estimation procedure based on inverse probability weighting in Section 6.2.3.

Algorithm 5 Debiased Recommendation under Confounding and Selection Biases (dREC)

1: **Input:** Historical observation data $\mathcal{H}$ with item features $\mathbf{I}$, user features $\mathbf{U}$, click Y, and $P(\mathbf{T})$ of externally and unbiasedly measured features $\mathbf{T}$.
2: **Initialization:** Construct causal graph $\mathcal{G}_s$ based on $\mathcal{H}$.
3: $\mathbf{I}, \mathbf{Z}, \mathbf{Z}^{\intercal} \leftarrow \mathcal{F}(\mathcal{G}_s, \mathbf{I}, \mathbf{U}, Y, \mathbf{T})$. (Algorithm 6)
4: **for** each user u with features $\mathbf{u}$ **do**
5: **for** $\mathbf{i} = 1, 2, ...|D_{\mathcal{I}}|$ **do**
6: Compute $P(y|do(\mathbf{i}), \mathbf{u})$ using Equation 6.4.
7: **end for**
8: Recommend to user u an item with features $\mathbf{i} = \text{argmax}_{\mathbf{i} \in D_{\mathcal{I}}} P(y|do(\mathbf{i}), \mathbf{u})$.
9: **end for**

Algorithm 5 shows the pseudo code of our debiased personalized recommendation (dREC) under the existence of confounding and selection biases. The input of our dREC algorithm consists of two parts: $P(\mathbf{V}|S = 1)$ as a distribution collected under selection bias, and $P(\mathbf{T})$ as a distribution of a subset of the variables $\mathbf{T}$. In line 2, we first build

a causal graph $\mathcal{G}$ based on observed data and then construct the augmented graph $\mathcal{G}_s$ by adding a node S representing a binary indicator of entry into the observed data. In line 3, we call the function $\mathcal{F}$ to find a valid adjustment pair $(\mathbf{Z}, \mathbf{Z}^{\intercal})$ and intervention set $\mathbf{I}$ so that we can eliminate confounding and selection biases simultaneously from $P(y|do(\mathbf{i}), \mathbf{u})$. In line 6, we perform the estimation procedure based on inverse probability weighting to calculate conditional causal effects $P(y|do(\mathbf{i}), \mathbf{u})$. In line 8, we recommend to user u an item with features $\mathbf{i} = \text{argmax}_{\mathbf{i} \in D_{\mathcal{I}}} P(y|do(\mathbf{i}), \mathbf{u})$, where $D_{\mathcal{I}}$ is the space of identified intervention features $\mathbf{I}$ from Algorithm 6.

6.2.2 Identification under Confounding and Selection Biases

We present our main theoretical result in Theorem 8. It shows that the conditional causal effects $P(Y|do(\mathbf{I}), \mathbf{U})$ can be identified and recovered under confounding and selection biases if and only if an adjustment pair $(\mathbf{Z}, \mathbf{Z}^{\intercal})$ satisfying certain graphical criterion can be identified.

Theorem 8 (**Generalized Adjustment for Conditional Intervention**)**.** Given a causal diagram $\mathcal{G}$ augmented with selection variable S, variable sets $\mathbf{I}, \mathbf{U}, \mathbf{Z}$, outcome Y, a set of externally and unbiasedly measured variables $\mathbf{T}$, and a set $\mathbf{Z}^{\intercal} \subseteq \mathbf{Z} \cap \mathbf{T}$, for every model compatible with $\mathcal{G}$, we have

$$P(y|do(\mathbf{i}), \mathbf{u}) = \sum_{\mathbf{z}} P(y|\mathbf{i}, \mathbf{z}, \mathbf{u}, S = 1) P(\mathbf{z} \backslash \mathbf{z}^{\intercal} | \mathbf{z}^{\intercal}, \mathbf{u}, S = 1) P(\mathbf{z}^{\intercal}|\mathbf{u}) \tag{6.2}$$

if and only if $(\mathbf{Z}, \mathbf{Z}^{\intercal})$ satisfies the following generalized adjustment criterion:

1. No element in $\mathbf{Z}$ is a descendant in $\mathcal{G}_{\overline{\mathbf{I}}}$ of any $W \notin \mathbf{I}$ lying on a proper causal path from $\mathbf{I}$ to $\mathbf{Y}$.

2. All non-causal paths in $\mathcal{G}$ from $\mathbf{I}$ to $\mathbf{Y}$ are blocked by $\mathbf{Z} \cup \mathbf{U}$ and S.

3. $\mathbf{Z}^{\intercal}$ d-separates $\mathbf{Y}$ from S in the proper backdoor graph, i.e., $(\mathbf{Y} \perp\!\!\!\perp S|\mathbf{Z}^{\intercal})_{\mathcal{G}^{pbd}_{\mathbf{YI}}}$.

$(\mathbf{Z}, \mathbf{Z}^{\intercal})$ is said to be an adjustment pair for recovering the conditional causal effect of $\mathbf{I}$ on Y given $\mathbf{U}$.

Generally speaking, condition (1) prevents causal paths from being compromised by conditioning on an element in $\mathbf{Z}$. Condition (2) requires all non-causal paths to be blocked by $\mathbf{Z} \cap \mathbf{U}$ and S. Condition (3) ensures that the influence of the selection mechanism on the outcome is nullified by $\mathbf{Z}^{\intercal}$. Essentially, if the set $(\mathbf{U}, \mathbf{Z}, S)$ blocks all the non-causal paths, and $\mathbf{z}^{\intercal}$ d-separates Y and S, we are able to unbiasedly estimate $P(y|do(\mathbf{i}), \mathbf{u})$ under the presence of confounding and selection biases.

Proof sketch. Our derivation of generalized adjustment for conditional intervention is based on the proof idea of Theorem 2 in [46]. If we can find an adjustment pair $(\mathbf{Z}, \mathbf{Z}^{\intercal})$ that satisfies the three graphical conditions in Theorem 8 for $\mathbf{I}, Y$ in $\mathcal{G}$, then $P(y|do(\mathbf{i}), \mathbf{u})$ can be expressed as Equation 6.2. Note that $P(y|\mathbf{i}, \mathbf{z}, S = 1)$ and $P(\mathbf{z} \backslash \mathbf{z}^{\intercal}|\mathbf{z}^{\intercal}, S = 1)$ can be directly computed from the biased observational data $P(\mathbf{v}|S = 1)$, and $P(\mathbf{z}^{\intercal}|\mathbf{u})$ from the external unbiased observational data $P(\mathbf{t})$ given a certain user profile. The target condition intervention thus can be computed from the confounded and sample selection biased data. If $P(y|do(\mathbf{i}), \mathbf{u})$ is computable and recoverable by the adjustment expression in Equation 6.2, according to Theorem 1, $(\mathbf{Z}, \mathbf{Z}^{\intercal})$ is a valid adjustment pair. Please refer to Appendix for the detailed proof. □

Remark. Equation 6.2 is a significant extension of Equation 6.1 and it deals with both confounding and selection biases by allowing the use of unbiased data over a subset of the

covariates. Note that the target distribution cannot be recovered from selection biased data if the selection node is outcome-dependent. This can be regarded as a corollary of Theorem 1, where S and Y cannot be d-separated by arbitrary intervention sets. More specifically, if an arrow exists from Y to S, the conditional independence of S and Y given $\mathbf{I}$ is no longer valid.

Algorithm 6 Identify Intervention Set and Adjustment Pair

1: **Input:** Causal graph $\mathcal{G}_s$, Item features $\mathbf{I}$, User features $\mathbf{U}$, Outcome Y, Available external unbiased features $\mathbf{T}$.
2: **Initialize:** $\mathbf{Z} \leftarrow \emptyset$, $\mathbf{Z}^\intercal \leftarrow \emptyset$.
3: **for all** $\mathbf{I}_s \subseteq \mathbf{I}$ starting with the largest size **do**
4: **if** An adjustment pair can be found according to Theorem 9 **then**
5: **return** $(\mathbf{I}_s, \mathbf{Z}, \mathbf{Z}^\intercal)$.
6: **else**
7: Apply function ListAdjPairs$(\mathcal{G}_s, \mathbf{I}_s, Y, S, \mathbf{V}, \mathbf{T})$ [46]
8: **if** List of Adjustment Pairs is not empty **then**
9: **return** $(\mathbf{I}_s, \mathbf{Z}, \mathbf{Z}^\intercal)$ with the least cost.
10: **end if**
11: **end if**
12: **end for**

Algorithm 6 shows our procedure of identifying an adjustment pair for $P(y|do(\mathbf{i}), \mathbf{u})$. In line 3-12, we search over the subset space of $\mathbf{I}$ and return the intervention set $\mathbf{I}_s$ and its corresponding adjustment pair $(\mathbf{Z}, \mathbf{Z}^\intercal)$. Although Theorem 8 shows the generalized criterion for adjustment pairs, one challenge is how to find them systematically and efficiently. It is clear that any algorithm that aims to output all adjustment pairs will take exponential time. In line 4, we call an explicit procedure to find one valid adjustment pair based on Theorem 9.

Theorem 9. $(\mathbf{Z}, \mathbf{Z}^\intercal)$ is an adjustable pair if

$$\mathbf{Z} = An(\mathbf{I} \cup Y \cup \{S\})_{\mathcal{G}^{pbd}_{\mathbf{I}Y}} \cap \mathbf{C},\ \mathbf{Z}^\intercal = (An(\{S\} \cup Y)_{\mathcal{G}^{pbd}_{\mathbf{I}Y}} \cap \mathbf{T}) \cap \mathbf{C}$$

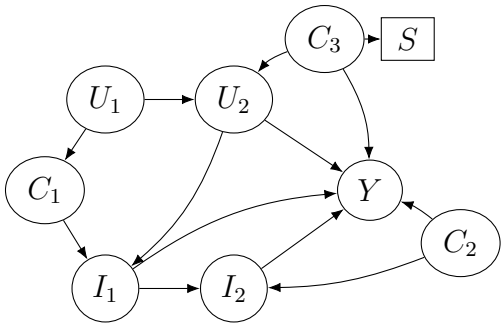


Figure 6.2: Illustrative example of implementing dREC algorithm.

where $\mathbf{C} = \mathbf{V} \backslash (\mathbf{I} \cup Y \cup De((De(\mathbf{I}))_{\mathcal{G}_{\overline{\mathbf{I}}}} \backslash \mathbf{I}) \ \cap \ An(Y)_{\mathcal{G}_{\overline{\mathbf{I}}}})$.

Applying Theorem 9, we are able to construct one admissible pair efficiently with $\mathcal{O}(n+m)$ time complexity where n, m are the number of variables and edges in $\mathcal{G}$, respectively. If the explicit adjustment pair cannot be identified, we then apply the procedure ListAdjPairs developed in [46] that outputs all sets for generalized adjustment. ListAdjPairs runs with polynomial delay. If no adjustment pair exists for the current intervention set $\mathbf{I}_s$, we move to another candidate feature subset for intervention.

We next presents a computation complexity analysis of our dREC algorithm. Specifically, line 3 in Algorithm 5 calls Algorithm 6 to find a valid adjustment pair, which works with $\mathcal{O}(2^{|\mathbf{I}|} \cdot n(n+m))$ complexity in the worst case by applying function LISTADJPAIRS for each candidate intervention set. Line 4-9 estimates the condition causal effect for each user-item pair, which takes $\mathcal{O}(|D_{\mathcal{I}}| \cdot |u|)$ complexity, where $|\mathbf{I}|$ and $|D_{\mathcal{I}}|$ are the ordinality of the original item feature set and the identified intervention set returned by Algorithm 6, $|u|$ denotes the number of users. Thus dREC algorithm achieves $\mathcal{O}(max(2^{|\mathbf{I}|} \cdot n(n+m), |D_{\mathcal{I}}| \cdot |u|))$ computation complexity in general.

Figure 6.2 shows an illustrative example of a causal graph for personalized recommendation. In Figure 6.2, U_1, U_2 represent the user features, I_1, I_2 represent the item features, Y

denotes the outcome, C_1, C_2, C_3 denote the potential adjustment covariates, among which C_3 affects selection mechanism S. Suppose the intervention set is $\mathbf{I} = \{I_1, I_2\}$, the user feature set is $\mathbf{U} = \{U_1, U_2\}$. Since the node Y cannot be d-separated from S conditioned on $(\mathbf{U}, \mathbf{I})$, $P(Y|do(\mathbf{i}), \mathbf{u})$ cannot be recovered from observational distribution $P(\mathbf{v}|S = 1)$. However, by incorporating the unbiased observation of $C_3 \subseteq \mathbf{T}$, we can identify one admissible adjustment pair $(\mathbf{Z}, \mathbf{Z}^\intercal) = (\{C_2, C_3\}, C_3)$ from potential adjustment covariates. The causal quantity can thus be identified and recovered from the observational distribution by applying Equation 6.2. We also emphasize that the truncated factorization based on adjustment pair $(\mathbf{Z}, \mathbf{Z}^\intercal)$ is robust towards some unobserved confounding. For example, if there is one hidden confounder between I_2 and C_2, our debiased formula still holds.

6.2.3 Estimation Based on Inverse Probability Weighting

The direct calculation of $P(y|do(\mathbf{i}), \mathbf{u})$ in Equation 6.2 involves finding the conditional probability of Y given $\mathbf{I}$ for each stratum defined by the possible values of the covariates $\mathbf{Z}$ and user features $\mathbf{U}$. This presents computational and sample complexity challenges because the number of different strata may be huge with the cardinality of $\mathbf{Z}$ and $\mathbf{U}$. As a result, the number of samples in the training data falling under each stratum is too small to provide a reliable estimate of the conditional distribution. In this section, we follow the widely adopted robust statistical estimation procedure, inverse probability weighting estimation [102], to construct an estimator for conditional causal effects in the presence of selection bias using the generalized adjustment given in Theorem 8.

With the absence of sample selection bias, given observed i.i.d. data $\mathcal{H} = \{(\mathbf{u}_t, \mathbf{i}_t, \mathbf{z}_t, y_t)\}_{t=1}^T$, the IPW estimator for $P(y|do(\mathbf{i}), \mathbf{u})$ is given by $\hat{\mu} = \frac{1}{T}\sum_{t=1}^T w_t \mathbb{1}_{\mathbf{i}_t=\mathbf{i}} y_t$, where $\mathbb{1}_{\mathbf{i}_t=\mathbf{i}}$ is the indicator function, $w_t = \frac{1}{\hat{P}(\mathbf{i}_t|\mathbf{z}_t,\mathbf{u}_t)\hat{P}(\mathbf{u}_t)}$, and $\hat{P}(\mathbf{i}_t|\mathbf{z}_t, \mathbf{u}_t))$ is the estimator of the propensity score

that is estimated from data by some parametric model.

Given a valid adjustment pair $(\mathbf{Z}, \mathbf{Z}^\intercal)$, $P(y|do(\mathbf{i}), \mathbf{u})$ can be rewritten as follows:

$$\begin{aligned}
&\sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, \mathbf{u}, S=1) P(\mathbf{z} \backslash \mathbf{z}^\intercal | \mathbf{z}^\intercal, \mathbf{u}, S=1) P(\mathbf{z}^\intercal | \mathbf{u}) \\
&= \sum_{\mathbf{Z}} \frac{P(y, \mathbf{i}, \mathbf{z}|\mathbf{u}, S=1)}{P(\mathbf{i}|\mathbf{z}, \mathbf{u}, S=1)} \frac{P(\mathbf{z}^\intercal|\mathbf{u})}{P(\mathbf{z}^\intercal|\mathbf{u}, S=1)} \\
&= \sum_{\mathbf{Z}} \frac{P(y, \mathbf{i}, \mathbf{z}, \mathbf{u}|S=1)}{P(\mathbf{i}|\mathbf{z}, \mathbf{u}, S=1)} \frac{P(S=1|\mathbf{u})}{P(S=1|\mathbf{z}^\intercal, \mathbf{u})}
\end{aligned} \tag{6.3}$$

Given observed data $\{(\mathbf{u}_t, \mathbf{i}_t, \mathbf{z}_t, y_t)\}_{t=1}^T$ under selection bias from $P(\mathbf{v}|S=1)$, we can obtain a reliable estimate of the propensity score $P(\mathbf{i}|\mathbf{z}, \mathbf{u}, S=1)$ as well as the inverse probability of the selection weight conditioned on a certain user profile $\frac{P(S=1|\mathbf{u})}{P(S=1|\mathbf{u}, \mathbf{z}^\intercal)}$, from selection biased data and additional unbiased external data. Thus, the conditional causal effects under selection bias can be estimated by:

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} w_t^c \cdot w_t^s \mathbb{1}_{\mathbf{i}_t = \mathbf{i}} y_t \tag{6.4}$$

where $w_t^c = \frac{1}{P(\mathbf{i}_t|\mathbf{z}_t, \mathbf{u}_t, S=1)}$, and $w_t^s = \frac{P(S=1|\mathbf{u}_t)}{P(S=1|\mathbf{u}_t, \mathbf{z}_t^\intercal)}$. In practice, these quantities can be estimated by some parametric models like logistic regression or neural networks. According to Equation 6.4, the loss function in the training can be written as:

$$\mathcal{L}_{dRec} = \frac{1}{T} \sum_{t=1}^{T} w_t^c \cdot w_t^s \cdot L(\hat{y}_t, y_t) \tag{6.5}$$

where L denotes the predominant Mean Squared Error (MSE) metric $L(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$.

6.3 Extension

6.3.1 Path-specific Causal Effect

In this section, we aim to measure the path-specific causal effect under the sample selection mechanism. Due to the orthogonality of confounding and sample selection biases,

it is non-trivial to derive the expression of path-specific causal effect under sample selection bias, and the graphical condition for causal effect identification.

Definition 7 (Path-specific Causal Effect)**.** The path-specific causal effect measures the effect of changing $\mathbf{I}$ from $\mathbf{i}_1$ to $\mathbf{i}_2$ on an outcome Y transmitted along a path set π.

$$PTE_{\pi}(\mathbf{i}_2, \mathbf{i}_1) = P(Y = y|do(\mathbf{i}_2|_{\pi}, \mathbf{i}_1|_{\bar{\pi}})) - P(Y = y|do(\mathbf{i}_1)) \tag{6.6}$$

where π denotes a subset of causal paths from $\mathbf{I}$ to Y and $\bar{\pi}$ denotes the causal paths not in π.

The condition under which the path-specific effect can be estimated from the observational data is known as the recanting witness criterion.

Definition 8 (Recanting Witness Criterion)**.** Given a path set π pointing from $\mathbf{I}$ to Y, let W be a node in $\mathcal{G}$ such that: i) there exists a path from $\mathbf{I}$ to W which is a segment of a path in π; ii) there exists a path from W to Y which is a segment of a path in π; iii) there exists another path from W to Y which is not a segment of any path in π. Then, the recanting witness criterion for the path-specific treatment effect is satisfied with W as a witness.

The path-specific causal effect can be computed from the observational data if and only if the recanting witness criterion is not satisfied [103]. Note that to calculate the second term $P(y|do(\mathbf{i}_1))$ in the presence of confounding bias and selection bias, we can directly follow the adjustment formula shown in Equation 3.3 and obtain

$$P(y|do(\mathbf{i}_1)) = \sum_{\mathbf{Z}} P(y|\mathbf{i}_1, \mathbf{z}, S = 1)P(\mathbf{z} \backslash \mathbf{z}^{\intercal}|\mathbf{z}^{\intercal}, S = 1)P(\mathbf{z}^{\intercal}) \tag{6.7}$$

We then aim to compute the second term $P(Y = y|do(\mathbf{i}_2|_{\pi}, \mathbf{i}_1|_{\bar{\pi}}))$ in the presence of confounding and selection biases if some unbiased observations can be further collected. We

give our main theorem of computing and recovering the the second term of path-specific causal effect under selection biases as follows:

Theorem 10 (Path-specific Causal Effect under Selection Bias)**.** In a Markovian model, the path-specific treatment effect under selection bias is recoverable if the recanting witness criterion is not satisfied and the generalized adjustment criterion is satisfied simultaneously. Specifically, the first term $P(Y = y|do(\mathbf{i}_2|_\pi, \mathbf{i}_1|_{\bar{\pi}}))$ in Equation 6.6 is given by

$$\sum_{\mathbf{Z}} \left(\sum_{\mathbf{PA}_\pi} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z}, S = 1) P(y|\mathbf{pa}_\pi, \mathbf{i}_1, \mathbf{z}, S = 1) \right) \times P(\mathbf{z} \backslash \mathbf{z}^\intercal|\mathbf{z}^\intercal, S = 1) P(\mathbf{z}^\intercal) \tag{6.8}$$

Proof sketch. Let $\mathbf{PA}$ denote Y's parent nodes along all causal paths, $\mathbf{PA}_\pi$ denote Y's parent nodes that lie in π, and $\mathbf{PA}_{\bar{\pi}}$ denote the remaining parents along the causal paths. We can compute $P(Y = y|do(\mathbf{i}_2|_\pi, \mathbf{i}_1|_{\bar{\pi}}))$ by adjusting on a valid covariate set $\mathbf{Z}$, which is shown in the following equation:

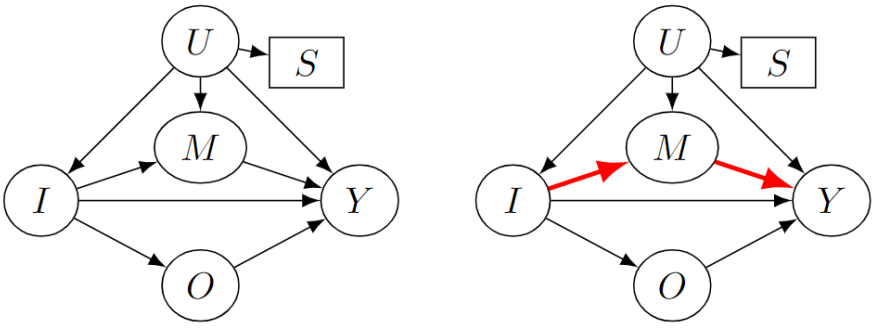
$$\sum_{\mathbf{Z} \cup \mathbf{PA}} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z}) P(\mathbf{pa}_{\bar{\pi}}|\mathbf{i_1}, \mathbf{z}) P(y|\mathbf{pa}, \mathbf{z}) P(\mathbf{z}) = \sum_{\mathbf{Z}} \left(\sum_{\mathbf{PA}_\pi} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z}) P(y|\mathbf{pa}_\pi, \mathbf{i}_1, \mathbf{z}) \right) P(\mathbf{z}) \tag{6.9}$$

We then derive Equation 6.8 based on Equation 6.9 and the condition that we are able to find a valid adjustment pair $(\mathbf{Z}, \mathbf{Z}^\intercal)$ that satisfies the generalized adjustment criterion. Please refer to Appendix for the detailed proof. □

We next use the causal graph in Figure 6.3 to further illustrate the computation process. To recover causal effect from selection and confounding biases, it is obvious to identify the adjustment pair $(\mathbf{Z}, \mathbf{Z}^\intercal) = (U, U)$. Setting $\pi = I \to M \to Y$, the path-specific

causal effect transmitted along π is $PTE_{\pi}(i_2, i_1) = P(Y = y|do(i_2|_{\pi}, i_1|_{\bar{\pi}})) - P(Y = y|do(i_1))$. The first term can be computed from observational distribution in the following procedure:

$$
\begin{aligned}
& P(Y = y|do(i_2|_{\pi}, i_1|_{\bar{\pi}})) \\
&= \sum_{U,M,O} P(m|i_2, u)P(o|i_1)P(y|i_1, u, m, o)P(u) \\
&= \sum_{U,M,O} P(m|i_2, u)P(o|i_1, u)P(y|i_1, u, m, o)P(u) \\
&= \sum_{U,M,O} \frac{P(m|i_2, u)}{P(m|i_1, u)} P(m, o|i_1, u)P(y|i_1, u, m, o)P(u) \\
&= \sum_{U,M,O} \frac{P(m|i_2, u)}{P(m|i_1, u)} P(y, m, o|i_1, u)P(u) \\
&= \sum_{U,M} \frac{P(m|i_2, u)}{P(m|i_1, u)} P(y, m|i_1, u)P(u) \\
&= \sum_{U,M} P(m|i_2, u)P(y|i_1, m, u)P(u) \\
&= \sum_{U} \left(\sum_{M} P(m|i_2, u, S = 1)P(y|i_1, m, u, S = 1) \right) P(u)
\end{aligned}
$$



(a) Original graph (b) Graph with specific path

Figure 6.3: Illustrative example of computing path-specific treatment effect.

6.3.2 Counterfactual Effect

Causal interventions only consider post-interventional distributions, but counterfactual inference considers both the real world without the intervention and the counterfactual world with the intervention. Particularly, we want to answer: what will the user's CTR be had they been recommended an item with different features? The counterfactual effect is expressed as $P(y_{\mathbf{i}}|\mathbf{i}') = P(Y(do(\mathbf{I} = \mathbf{i}))|\mathbf{I} = \mathbf{i}')$. We now extend the backdoor criterion for counterfactual effect to backdoor adjustment under selection bias.

Theorem 11 (Counterfactual Effect under Selection Bias)**.** If $(\mathbf{Z}, \mathbf{Z}^{\intercal})$ satisfies the generalized adjustment criterion in Theorem 8, the counterfactual effect $P(y_{\mathbf{i}}|\mathbf{i}')$ is identifiable and can be recovered from biased observational data by

$$P(y_{\mathbf{i}}|\mathbf{i}') = \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, S = 1)P(\mathbf{z} \backslash \mathbf{z}^{\intercal}|\mathbf{z}^{\intercal}, \mathbf{i}', S = 1)P(\mathbf{z}^{\intercal}) \tag{6.10}$$

Proof. We follow the regime in [104] to add a new node W with the same set of parents as $\mathbf{I}$ to generate the counterfactual graph $\mathcal{G}'$. The following lemma demonstrates that computing the counterfactual effect given a graph $\mathcal{G}$ is equivalent to calculating the related conditional intervention in $\mathcal{G}'$.

Lemma 5 (Equivalent Conditional Intervention [104])**.** The estimand of the counterfactual effect $P(y_{\mathbf{i}}|\mathbf{i}')$ is equal to that of the conditional intervention $P(y|w, do(\mathbf{i}))$ by replacing all occurrences of w with $\mathbf{i}'$.

We next demonstrate the derivation of estimating the conditional intervention $P(y|w, do(\mathbf{i}))$ from observation data under selection bias. First, since W is independent of $\mathbf{I}$, we follow the

rules of conditional probability and obtain:

$$P(y|w, do(\mathbf{i})) = P(y, w|do(\mathbf{i}))/P(w|do(\mathbf{i})) = P(y, w|do(\mathbf{i}))/P(w) \tag{6.11}$$

Based on the construction of $\mathcal{G}'$ and the first two conditions of the generalized adjustment criterion, it is straightforward to derive $(Y \perp\!\!\!\perp W|\mathbf{Z})_{\mathcal{G}'_{\bar{\mathbf{I}}}}$. Thus, the first term in Equation 6.11 can be marginalized on $\mathbf{Z}$ as follows:

$$\begin{aligned} P(y|w, do(\mathbf{i})) &= \Big(\sum_{\mathbf{z}} P(y|\mathbf{z}, do(\mathbf{i}))P(w, \mathbf{z})\Big)/P(w) \\ = \Big(\sum_{\mathbf{z}} P(y|\mathbf{z}, \mathbf{i})P(w, \mathbf{z})\Big)/P(w) &= \sum_{\mathbf{z}} P(y|\mathbf{z}, \mathbf{i})P(\mathbf{z}|w) \end{aligned} \tag{6.12}$$

We then recover the target distribution from selection biased data. Based on the third condition of the generalized adjustment criterion and the fact that node W cannot be the decedent of $\mathbf{Z}$, we further decompose $P(\mathbf{z}|w)$ by leveraging unbiased data $\mathbf{Z}^{\intercal}$. We have:

$$\begin{aligned} P(y|w, do(\mathbf{i})) &= \sum_{\mathbf{z}} P(y|\mathbf{z}, \mathbf{i})P(\mathbf{z}\backslash\mathbf{z}^{\intercal}|w, \mathbf{z}^{\intercal})P(\mathbf{z}^{\intercal}|w) \\ &= \sum_{\mathbf{z}} P(y|\mathbf{z}, \mathbf{i}, S=1)P(\mathbf{z}\backslash\mathbf{z}^{\intercal}|w, \mathbf{z}^{\intercal}, S=1)P(\mathbf{z}^{\intercal}) \end{aligned} \tag{6.13}$$

Finally, substituting w with $\mathbf{i}'$ in Equation 6.13 leads to the result in Equation 6.10. □

Figure 6.4 shows an illustrative example for calculating counterfactual effect with the presence of sample selection bias. It is easy to identify $(\mathbf{Z}, \mathbf{Z}^{\intercal}) = (U, U)$. We thus have:

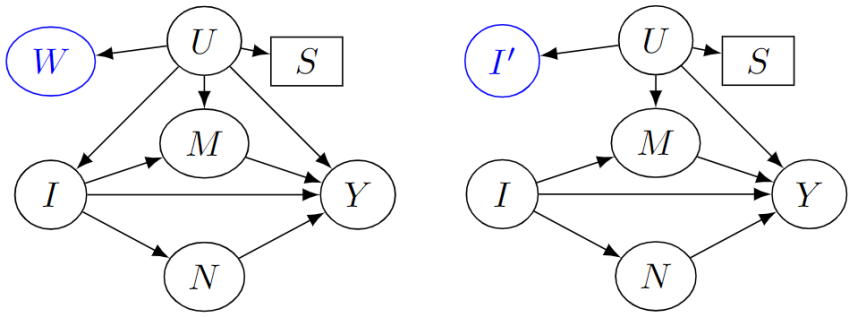
$$P(y|w, do(i)) = P(y, w|do(i))/P(w|do(i)) = P(y, w|do(i))/P(w) \tag{6.14}$$

Furthermore we can decompose $P(y, w|do(i))$ as:

$$\begin{aligned}
&\sum_{U,M,N} P(n|do(i))P(m|u, do(i))P(w|u)P(u)P(y|u, m, n, do(i)) \\
&= \sum_{U,M,N} P(m, n|u, do(i))P(y|u, m, n, do(i))p(w, u) \\
&= \sum_{U,M,N} P(y, m, n|u, do(i))P(w, u) \\
&= \sum_{U} P(y|u, do(i))P(w, u)
\end{aligned}$$

Since S and Y are d-separated by U and we are able to have the unbiased observation of U, plugging in the above equation to Equation 6.14 and replacing W with I' leads to the following results.

$$P(y|w, do(i)) = \sum_{U} P(y|u, do(i))P(u|i') = \sum_{U} P(y|U, do(i), S = 1)P(u) \tag{6.15}$$



(a) Graph augmented with node W (b) Counterfactual Graph $\mathcal{G}'$

Figure 6.4: Illustrative example of computing counterfactual effect.

6.4 Empirical Evaluation

We compare the performance of our dREC algorithm with baselines on Adult-Video dataset under the task of personalized recommendation. We follow the settings of [101] by combining two publicly available datasets: Adult dataset [86] and Youtube video dataset [1]. The user feature set $\mathbf{U}$ includes `age(a)`, `education(e)`, `sex`, `marital-status`, `workclass`, `hours`, and `income`. The item feature set $\mathbf{I}$ includes `length(l)`, `ratings(r)`, `views(v)`, and `comments(c)`. We generate clicks from the concatenation of user/item features $(\mathbf{U}, \mathbf{I})$ by the following procedure:

$$\begin{aligned} clicks &\sim Bernoulli(p) \\ p_d &= \mathcal{K}_1(a, e, v, l, c) + \mathcal{K}_2(\mathbb{1}_a, \mathbb{1}_e) \\ p_s &= \mathcal{K}_1(a, e, v, l, c) + \mathcal{K}_2(\mathbb{1}_a, \mathbb{1}_e) + \epsilon \end{aligned} \tag{6.16}$$

where $\mathbb{1}_a = 1$ if `age = view = length = 1`, otherwise $\mathbb{1}_a = 0$; $\mathbb{1}_e = 1$ if `education = comments = 1`, otherwise $\mathbb{1}_e = 0$; $\mathcal{K}_1(\mathbf{x}) = 0.1\mathbf{x}$, $\mathcal{K}_2(\mathbf{x}) = 0.25\mathbf{x}$. p_d denotes the click probability for deterministic case. We inject a noise term ϵ in p_s to simulate stochastic situation, which is sampled from the truncated normal distribution $\mathcal{N}_t(0, 0.1)$.

We randomly select 80 percent of users from the Adult dataset and recommend them a video from the Youtube video dataset. We generate user click information to form training data under selection bias. We divide `age` into 9 subgroups, then include a tuple with a certain probability to generate the biased dataset. We use $S = [\alpha, \beta]$ to describe the selection mechanism where α denotes the selection probability of the 10-30 year old group, and β denotes the others. We set $S_1 = [0.2, 0.8]$ and $S_2 = [0.7, 0.8]$ to simulate one skewed selection

[1] https://netsg.cs.sfu.ca/youtubedata/

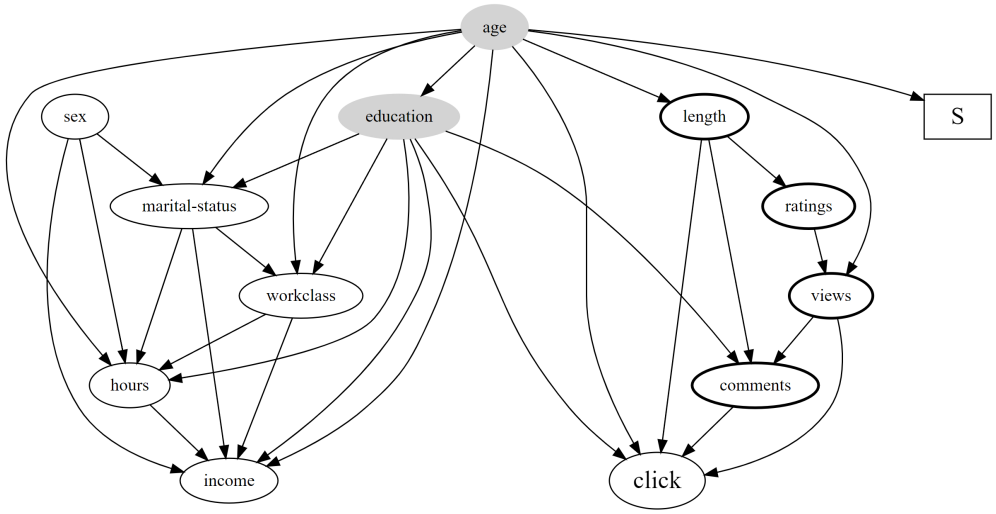


Figure 6.5: Causal graph for Adult-Youtube video Dataset.

Table 6.1: Comparison results in the Adult-Video dataset. Lower MAE and higher Hit@1 (PR@5) mean better results.

	Deterministic			Stochastic		
	CP	Biased	dREC	CP	Biased	dREC
MAE, S1	0.177	0.172	**0.169**	0.180	**0.177**	**0.177**
Hit@1, S1	0.231	0.277	**0.292**	0.193	0.199	**0.269**
PR@5, S1	0.263	0.301	**0.361**	0.247	0.285	**0.318**
MAE, S2	0.167	**0.164**	**0.164**	0.169	0.167	**0.165**
Hit@1, S2	0.273	0.316	**0.322**	0.212	0.230	**0.246**
PR@5, S2	0.299	**0.407**	0.397	0.248	0.360	**0.364**

scenario and one slightly skewed scenario, respectively. We construct the causal graph from the training data using the TETRAD software [105]. Figure 6.5 shows the constructed causal graph where the light-shaded nodes (`age`, `education`) denote the adjustment variables $\mathbf{Z}$, `age` is $\mathbf{Z}^{\intercal}$, and S denotes the selection mechanism. We compare our proposed method, dREC, with two baselines: conditional probability (CP) based on $P(Y = 1|\mathbf{U} = \mathbf{u}, \mathbf{I} = \mathbf{i})$, and biased estimate (Biased) based on $P(Y = 1|\mathbf{U} = \mathbf{u}, do(\mathbf{I} = \mathbf{i}))$. We evaluate the performance using three metrics: prediction accuracy based on MAE, Hit@1, and Precision@5.

Table 6.1 shows our experimental result. For all of the results, we run our experiments five times and report the average values. We summarize the comparison results based on t-test and find out that dREC achieves best results. For both deterministic and stochastic situations

under biased selection mechanisms S_1 and S_2, the p-values of testing dREC against the two baselines are less than 0.05 in terms of all three metrics in 21 out of 24 comparison cases. Note that when the slightly skewed mechanism S_2 is applied, under the deterministic situation, dREC obtains slightly lower Precision@5 (0.397) than Biased (0.407). This is because there are three estimated terms in dREC (Equation 6.2), but only two estimated terms in Biased (Equation 6.1). More estimation terms induce more uncertainty.

6.5 Evaluation of Path-specific Effect and Counterfactual Effect

We show the accuracy of our debiased approach using the Adult dataset [86]. We first construct the causal graph using TETRAD [105] software. Figure 6.6 shows the constructed causal graph. Node S denotes the selection mechanism. Node `workclass(w)` is the variable under intervention. Node `income` denotes the outcome variable. Light-shaded nodes {`age`, `education`, `marital-status`, `sex`} and `age` denote the adjustment pair $\mathbf{Z}$ and $\mathbf{Z}^{\intercal}$.

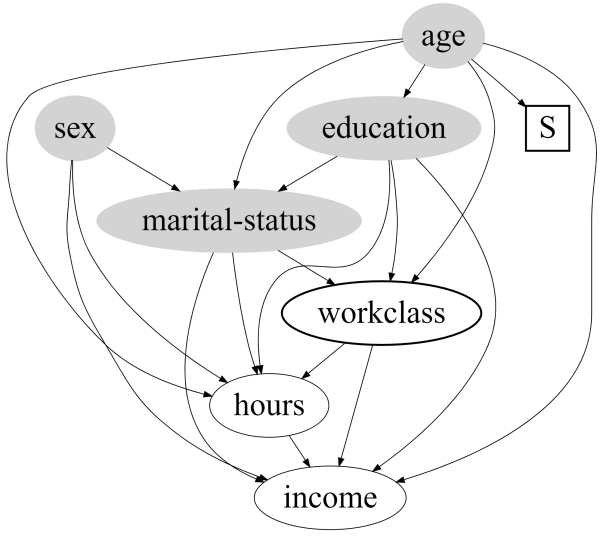


Figure 6.6: Causal graph for Adult Dataset.

We divide `age` into 5 subgroups, then include a tuple with a certain probability to generate the biased dataset. S_i $(i = 1, 2, 3)$ represents three selection mechanisms. In par-

ticular, let α denote the selection probability of the 20-60 year old group, and β denote the others. The selection mechanism can thus be represented as $S = [\alpha, \beta]$. We set $S_1 = [0.2, 0.8]$, $S_2 = [0.8, 0.2]$ to simulate two skewed selection mechanisms and set $S_3 = [0.8, 0.8]$ to simulate a uniform selection mechanism with no selection bias. To calculate the path-specific causal effect, we set the target path set π as `workclass` $\rightarrow$ `income`.

Table 6.2: Comparison results in Adult dataset.

	Causal Effect				*Path-Specific Effect*			*Counterfactual Effect*		
	CP	Truth	Biased	**dREC**	Truth	Biased	**dREC**	Truth	Biased	**dREC**
$w = 1, S_1$	0.256	0.242	0.215	0.243	0.155	0.129	0.156	0.223	0.194	0.202
$w = 1, S_2$	0.302	0.242	0.254	0.244	0.155	0.165	0.155	0.223	0.233	0.226
$w = 1, S_3$	0.288	0.242	0.242	0.242	0.155	0.154	0.154	0.223	0.226	0.218

We compare our debiased method with baselines on answering three causal queries: causal effect $P(Y = 1|do(w))$, path-specific causal effect $P(Y = 1|do(w_2|_\pi, w_1|_{\bar{\pi}}))$ and counterfactual effect $P((Y = 1)_{w_2}|w_1)$. We calculate each effect five times and report their average in Table 6.2. The column "Truth" denotes the ground truth value by calculating an effect based on the population data. The column "Biased" denotes estimating an effect using biased data without mitigating selection bias. The column "dREC" denotes estimating an effect using biased data and external observation with our derived formula. The column "CP" denotes estimating a causal effect using simple conditional probability estimator. The experimental results show our debiased method consistently improves the prediction accuracy of three causal queries, especially under biased sample selection scenarios S_1 and S_2.

6.6 Summary

In this chapter we studied both confounding and sample selection biases in recommendation systems and develops a causal based debiased recommendation algorithm that

simultaneously controls for confounding and selection biases via some auxiliary external data. We presented sufficient and necessary graphical conditions for conditional causal effects, path-specific effects, and counterfactual effects. We also derived a procedure to estimate an adjustment under confounding and selection biases based on the inverse probability weighting technique. Our empirical evaluation shows the effectiveness of our approach.

7 Robustly Improving Bandit Algorithms with Confounded and Selection Biased Offline Data

7.1 Introduction

The past decade has seen the rapid development of contextual bandit as a legit framework to model interactive decision-making scenarios, such as personalized recommendation [83], online advertising [106, 107], and anomaly detection [108]. The key challenge in a contextual bandit problem is to select the most beneficial item (i.e. the corresponding arm or intervention) according to the observed context at each round. In practice it is common that the agent has additional access to logged data from various sources, which may provide some useful information. A key issue is how to accurately leverage offline data such that it can efficiently assist the online decision-making process. However, one inevitable problem is that there may exist compound biases in the offline dataset, probably due to the data collection process, the existence of unobserved variables, the policies implemented by the agent, and so on. As a consequence, blindly fitting a model without considering those biases will lead to an inaccurate estimator of the reward distribution for each arm, ending up inducing a negative impact on the online learning phase.

To overcome this limitation and make good use of the offline data for online bandit learning, we formulate our framework from a causal inference perspective. Causal inference provides a family of methods to infer the effects of actions from a combination of data and qualitative assumptions about the underlying mechanism. Based on Pearl's structural causal model [21] we can derive a truncated factorization formula that expresses the target causal

quantity with probability distributions from the data. Appropriately adopting that prior knowledge on the reward distribution of each arm can accelerate the learning speed and achieve lower cumulative regret for online bandit algorithms.

Previous studies along this direction [70, 72, 69] only focused on one specific bias and have not dealt with compound biases in the offline data. It was shown in [44] that biases could be classified into confounding bias and selection bias based on the causal structure they imply. Due to the orthogonality of confounding and selection bias, simply deconfounding and estimating causal effects in the presence of selection bias using observational data is in general impractical without further assumptions, such as strong graphical conditions [45] or the accessibility of external unbiasedly measured data [109]. In this chapter, we address this limitation by non-parametrically bounding the target conditional causal effect when recoverability and identifiability can not be satisfied simultaneously. We propose two novel strategies to extract prior causal bounds for the reward distribution of each arm and use them to effectively guide the bandit agent to learn a nearly-optimal decision policy. We demonstrate that our approach could further reduce cumulative regret and is resistant to different levels of compound biases in offline data.

Our contributions can be summarized into three parts: 1) We derive causal bounds for conditional causal effects under confounding and selection biases based on c-component factorization and substitute intervention methods; 2) we propose a novel framework that leverages the prior causal bound obtained from biased offline data to guide the arm-picking process in bandit algorithms, thus robustly decreasing the exploration of sub-optimal arms and reducing the cumulative regret; and 3) we develop two contextual bandit algorithms (LinUCB-PCB and OAM-PCB) and one non-contextual bandit algorithm (UCB-PCB) that are enhanced with prior causal bounds. We theoretically show under mild conditions all

three bandit algorithms achieve lower regret than their non-causal counterparts. We also conduct an empirical evaluation to demonstrate the effectiveness of our method under the linear contextual bandit setting.

7.2 Algorithm Framework

An overview of our framework is illustrated in Figure 7.1. Our algorithm framework leverages the observational data to derive a prior causal bound for each arm to mitigate the cold start issue in online bandit learning, thus reducing the cumulative regret. In the offline evaluation phase, we call our bounding conditional causal effect (BCE) algorithm (shown in Algorithm 7) to obtain the prior causal bound for each arm given a user's profile. Then in the online bandit phase, we apply adapted contextual bandit algorithms with the prior causal bounds as input.

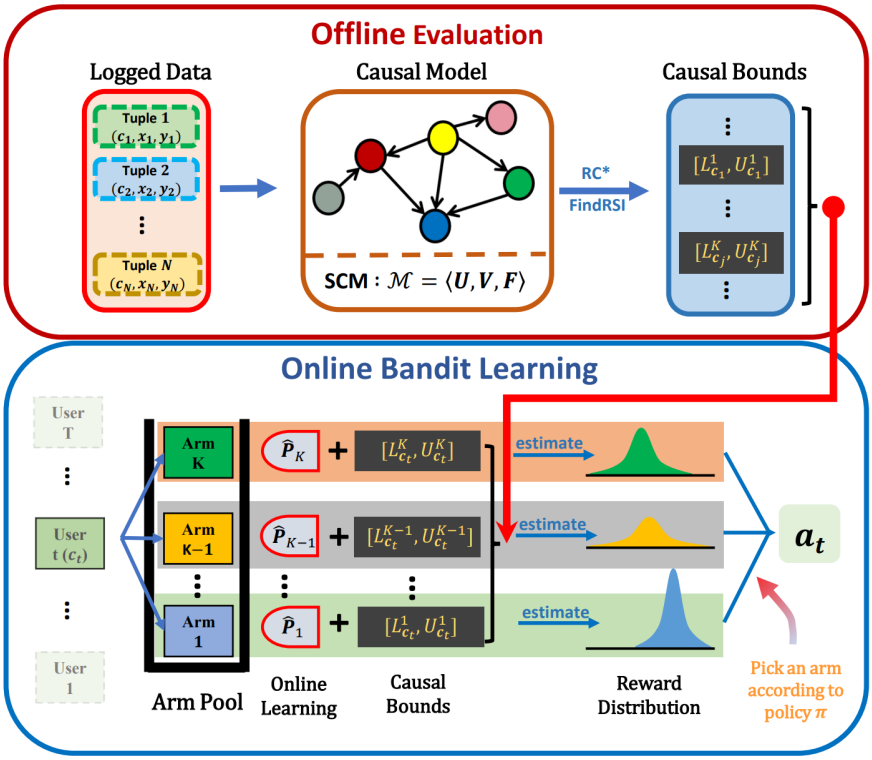


Figure 7.1: An illustration graph of our proposed framework.

Let $\mathbf{C} \in \mathcal{C}$ denote the context vector, where $\mathcal{C}$ denotes the domain space of $\mathbf{C}$. We use Y to denote the reward variable and $\mathbf{X} \in \mathcal{X}$ to denote the intervention variables. At each time $t \in [T]$, a user arrives and the user profile $\mathbf{c}_t$ is revealed to the agent. The agent pulls an arm a_t with features $\mathbf{x}_{a_t}$ following the policy $\pi \in \Pi$, and $\pi : \mathcal{C} \to \mathcal{X}$ could be treated as a mapping function from the context space $\mathcal{C}$ to the arm feature space $\mathcal{X}$. The agent then receives the reward Y_{a_t}. We use $u_{a_t} = \mathbb{E}[Y|do(\mathbf{X} = \mathbf{x}_{a_t}), \mathbf{c}_t]$ to denote the expected mean reward of pulling arm with feature value $\mathbf{x}_{a_t}$ given the user context. In the contextual bandit setting, the agent aims to approximate the optimal policy $\pi^* = \text{argmax}_{\pi \in \Pi} \mu_\pi$ and minimize the cumulative regret simultaneously (an arm choice in $\mathcal{X}$ can be seen as a policy in Π in the deterministic setting). Specifically, the goal of the agent is to pull an arm at each round, update the policy, and minimize the cumulative regret $R(T) = \sum_{t=1}^{T} (\mu_t^* - \mathbb{E}[Y_{a_t}])$.

When the offline observational data are available, we can leverage them to reduce explorations in the online phase. However, under the circumstances that the causal effect is either unidentifiable or nonrecoverable from the observational data, blindly using the observational data might even have a negative effect on the online learning phase. Our approach is to derive a causal bound for the desired causal effect from the biased observational data. We will further show even when the observational data could only lead to loose causal bounds, we can still guarantee our approach is no worse than conventional bandit algorithms.

7.3 Deriving Causal Bounds under Confounding and Selection Biases

In this section, we focus on bounding the effects of conditional interventions in the presence of confounding and sample selection biases. To tackle the identifiability issue of a conditional intervention $P(Y = y|do(\mathbf{x}), \mathbf{c})$, the cond-identify algorithm [110] provides a

complete procedure to compute conditional causal effects using observational distributions.

$$P(Y = y|do(\mathbf{x}), \mathbf{c}) = P_{\mathbf{x}}(y|\mathbf{c}) = \frac{P_{\mathbf{x}}(y, \mathbf{c})}{P_{\mathbf{x}}(\mathbf{c})} \tag{7.1}$$

where $P_{\mathbf{x}}(y|\mathbf{c})$ is the abbreviated form of the conditional causal effect $P(Y = y|do(\mathbf{x}), \mathbf{c})$. [110] showed that if the numerator $P_{\mathbf{x}}(y, \mathbf{c})$ is identifiable, then $P_{\mathbf{x}}(y|\mathbf{c})$ is also identifiable. In the contextual bandit setting, because none of the variables in $\mathbf{C}$ is a descendant of variables in $\mathbf{X}$, the denominator $P_{\mathbf{x}}(\mathbf{c})$ can be reduced to $P(\mathbf{c})$ following the causal topology. Since $P(\mathbf{c})$ is always identifiable and can be accurately estimated, we do not need to consider the situation where neither $P_{\mathbf{x}}(y, \mathbf{c})$ nor $P_{\mathbf{x}}(\mathbf{c})$ is identifiable but $P_{\mathbf{x}}(y|\mathbf{c})$ is still identifiable. Thus the conditional causal effect $P_{\mathbf{x}}(y|\mathbf{c})$ in Equation 7.1 is identifiable **if and only if** $P_{\mathbf{x}}(y, \mathbf{c})$ is identifiable. However, the cond-identify algorithm [110] is not applicable for the scenario with the presence of selection bias.

In this chapter we develop novel approaches for deriving bounds of conditional causal effects in the presence of both confounding and selection biases. Specifically, we allow the existence of unobserved confounders, which are denoted using dashed bi-directed arrows in the causal graph $\mathcal{G}$. We also introduce the selection node S depicting the data selection mechanism in the offline evaluation phase. With slight abuse of the notation, we denote $\mathcal{G}$ to be the causal graph augmented with S in the remaining sections. By adopting state-of-the-art causal discovery techniques on the offline dataset we assume the causal graph is accessible by the agent and remains invariant through the offline evaluation phase and online learning phase.

Algorithm 7 shows our algorithm framework of bounding conditional causal effects under confounding and selection biases. We develop two methods, c-component factorization

and substitute intervention, and apply each to derive a bound for conditional causal effect separately. We then compare the two causal bounds and return the tighter upper/lower bound. Specifically, lines 5-10 in Algorithm 7 decompose the target causal effect following c-component factorization and recursively call our RC* algorithm (shown in Algorithm 8) to bound each c-factor. Lines 11-15 search over recoverable intervention space and find valid substitute interventions to bound the target causal effect. Lines 16-18 compare two derived causal bounds and take the tighter upper/lower bound as the output causal bounds.

Algorithm 7 Bounding Conditional Causal Effect

1: **function** BCE($\mathbf{x}, \mathbf{c}, y, \mathcal{G}, \mathcal{H}$)
2: **Input:** Intervention variables $\mathbf{X} = \mathbf{x}$, context vector $\mathbf{C} = \mathbf{c}$, outcome variable $Y = y$, causal graph $\mathcal{G}$.
3: **Output:** Causal bound $[L_{\mathbf{x},\mathbf{c}}, U_{\mathbf{x},\mathbf{c}}]$ of the conditional intervention $P_{\mathbf{x}}(y|\mathbf{c})$.
4: **Initialization:** $[L_q, U_q] = [0, 1], [L_w, U_w] = [0, 1]$
5: // C-component Factorization
6: Decompose $P_{\mathbf{x}}(y, \mathbf{c}) = \sum_{\mathcal{D}\setminus\{\mathbf{Y},\mathbf{C}\}} \prod_{i=1}^{l} Q[\mathcal{D}_i]$ following Equation 3.4.
7: **for** each $\mathcal{D}_i$ **do**
8: $L_{Q(\mathcal{D}_i)}, U_{Q(\mathcal{D}_i)} = \text{RC}^*(\mathcal{D}_i, P(\mathbf{v}|S = 1), \mathcal{G})$
9: **end for**
10: Update L_q, U_q according to Theorem 12.
11: // Substitute Intervention
12: $\mathcal{D} = \text{FindRSI}(\mathbf{x}, \mathbf{c}, y, \mathcal{G})$
13: **if** $\mathcal{D} \neq \phi$ **then**
14: Update L_w, U_w according to Theorem 13.
15: **end if**
16: // Comparing Bounds
17: Calculate estimated values $\hat{L}_q, \hat{L}_w, \hat{U}_q, \hat{U}_w$ based on $\mathcal{H}$.
18: **return** $L_{\mathbf{x},\mathbf{c}} = \max\{\hat{L}_q, \hat{L}_w\}$, $U_{\mathbf{x},\mathbf{c}} = \min\{\hat{U}_q, \hat{U}_w\}$

7.3.1 Bounding via C-component Factorization

To derive the causal bound based on c-component factorization, we decompose the target intervention into c-factors and call RC* algorithm to recover each c-factor. The RC* algorithm shown in Algorithm 8 is designed based on the RC algorithm in [80] to accommo-

date for non-recoverable situations. When the c-factor $Q[\mathbf{E}]$ is recoverable, the RC* algorithm returns an expression of $Q[\mathbf{E}]$ using biased distribution $P(\mathbf{v}|S=1)$.

Specifically, lines 4-6 in Algorithm 8 marginalize out the non-ancestors of $\mathbf{E} \cup S$ since they do not affect the recoverability results. From Lemma 3 in [109], each c-component in line 7 is recoverable since none of them contains ancestors of S. Line 13 calls the Identify function proposed by [111] that gives a complete procedure to determine the identifiability of $Q[\mathbf{E}]$. When $Q[\mathbf{E}]$ is identifiable, Identify$(\mathbf{E}, \boldsymbol{C}_i, Q[\boldsymbol{C}_i])$ returns a closed form expression of $Q[\mathbf{E}]$ in terms of $Q[\boldsymbol{C}_i]$. In line 15, if none of the recoverable c-components $\boldsymbol{C}_i$ contains $\mathbf{E}$, we replace the distribution P by dividing the recoverable quantity $\prod_i Q[\boldsymbol{C}_i]$ and recursively run the RC* algorithm on the graph $\mathcal{G}_{\mathbf{V}\setminus\boldsymbol{C}}$. Under certain situations where line 8 in RC* Algorithm fails ($\boldsymbol{C} = \emptyset$), the corresponding $Q[\mathbf{E}]$ cannot be computed from the biased observational data in theory. These situations are referred to as nonrecoverable situations. We address this nonrecoverable challenge by non-parametrically bounding the targeted causal quantity. In this case, RC* returns a bound $[L_{Q(\mathbf{E})}, U_{Q(\mathbf{E})}]$ for $Q[\mathbf{E}]$. The bound for $P_{\mathbf{x}}(y, \mathbf{c})$ is derived by summing up the estimator/bounds of those c-factors following Equation 3.4.

Theorem 12 (Causal Bound from RC* algorithm)**.** Given a conditional intervention $P_{\mathbf{x}}(y|\mathbf{c})$, the causal bounds derived by calling RC* algorithm for each c-factor are:

$$
\begin{aligned}
L_q &= \sum_{\mathcal{D}\setminus\{\mathbf{Y},\mathbf{C}\}} \prod_{i=1}^{l} L_{Q[D_i]}/P_{\mathbf{x}}(\mathbf{c}) \\
U_q &= \sum_{\mathcal{D}\setminus\{\mathbf{Y},\mathbf{C}\}} \prod_{i=1}^{l} U_{Q[D_i]}/P_{\mathbf{x}}(\mathbf{c})
\end{aligned}
\tag{7.2}
$$

Note that in line 9 of RC* algorithm, we bound the target c-component by $[0, 1]$ since under semi-Markovian models it is challenging to find a tight bound for $Q[\mathbf{E}]$ when $\boldsymbol{C} = \emptyset$.

Algorithm 8 RC* Algorithm

function RC*($\mathbf{E}, P, \mathcal{G}$)
Input: $\mathbf{E}$ a c-component, P a distribution and $\mathcal{G}$ a causal graph.
Output: Causal bound $[L_{Q[\mathbf{E}]}, U_{Q(\mathbf{E})}]$ for $Q[\mathbf{E}]$.
if $\mathbf{V} \backslash (An(\mathbf{E}) \cup An(S)) \neq \phi$ **then**
    return RC*($\mathbf{E}, \sum_{\mathbf{V} \backslash (An(\mathbf{E}) \cup An(S))} P, \mathcal{G}_{(An(\mathbf{E}) \cup An(S))}$)
end if
Let $\boldsymbol{C}_1, ..., \boldsymbol{C}_k$ be the c-components of $\mathcal{G}$ that contains no ancestors of S and $\boldsymbol{C} = \cup_{i \in [k]} \boldsymbol{C}_i$.
if $\boldsymbol{C} = \emptyset$ **then**
    Bound $Q[\mathbf{E}]$ with $U_{Q(\mathbf{E})} = 1, L_{Q(\mathbf{E})} = 0$.
    return $L_{Q(\mathbf{E})}, U_{Q(\mathbf{E})}$
end if
if $\mathbf{E}$ is a subset of some $\boldsymbol{C}_i$ and Identify($\mathbf{E}, \boldsymbol{C}_i, Q[\boldsymbol{C}_i]$) does not return FAIL **then**
    return $L_{Q(\mathbf{E})} = U_{Q(\mathbf{E})} =$ Identify($\mathbf{E}, \boldsymbol{C}_i, Q[\boldsymbol{C}_i]$)
end if
return RC*($\mathbf{E}, \frac{P}{\prod_i Q[\boldsymbol{C}_i]}, \mathcal{G}_{\mathbf{V} \backslash \boldsymbol{C}}$)

One future direction is to further apply a non-parametric bounding technique similar to [9]. That is, choosing certain probability distributions in the truncated formula that are the source of unrecoverability. Then we set a variable set with specific domain values to allow the related probability distributions to achieve their maximum/minimum values.

7.3.2 Bounding via Substitute Interventions

From previous discussion we find that RC* algorithm may return a loose bound when we fail to recover most of the c-factors. In order to obtain a tight causal bound that is robust under various graph conditions, we develop another novel strategy to bound $P_{\mathbf{x}}(y, \mathbf{c})$. Our key idea is to search over the substitute recoverable interventions with a larger intervention space. Note that for a variable set $\mathbf{W}$ such that $\mathbf{W} \cap \mathbf{X} = \emptyset$ in the contextual bandit setting, we can perform marginalization on $\mathbf{W}$ and derive $P_{\mathbf{x}}(y, \mathbf{c}) = \sum_{\mathbf{W}} P_{\mathbf{x}}(y, \mathbf{c}|\mathbf{w}) P(\mathbf{w})$. We can

further bound $P_{\mathbf{x}}(y, \mathbf{c})$ by

$$\begin{aligned} P_{\mathbf{x}}(y, \mathbf{c}) &\leq \max_{\mathbf{w}^* \in \mathbf{W}} P_{\mathbf{x}}(y, \mathbf{c}|\mathbf{w}^*) \\ P_{\mathbf{x}}(y, \mathbf{c}) &\geq \min_{\mathbf{w}^* \in \mathbf{W}} P_{\mathbf{x}}(y, \mathbf{c}|\mathbf{w}^*) \end{aligned} \tag{7.3}$$

We then investigate whether the action/observation exchange rule of do-calculus [21] and the corresponding graph conditions could be extended in the presence of selection bias and list the results in the following lemma.

Lemma 6 (Action/Observation Rule under Selection Bias)**.** If the graphical condition $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}, S|\mathbf{X}, \mathbf{W})_{\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z(W)}}}}$ is satisfied in $\mathcal{G}$, the following equivalence between two post-interventional distributions holds:

$$P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{w}), \mathbf{z}, S = 1) = P(y|do(\mathbf{x}), \mathbf{W}, \mathbf{z}, S = 1) \tag{7.4}$$

where $\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{Z}}}$ represents the causal graph with the deletion of both incoming and outgoing arrows of $\mathbf{X}$ and $\mathbf{Z}$ respectively. $\mathbf{Z(W)}$ is the set of $\mathbf{Z}$-nodes that are not ancestors of variables in $\mathbf{W}$ in $\mathcal{G}_{\overline{\mathbf{X}}}$.

Following the general action/observation exchange rules in Equation 7.4, if $(Y, \mathbf{C} \perp\!\!\!\perp \mathbf{W}, S|\mathbf{X})_{\mathcal{G}_{\overline{\mathbf{X}}\underline{\mathbf{W}}}}$, we can replace $P_{\mathbf{x}}(y, \mathbf{c}|\mathbf{w}^*)$ with $P_{\mathbf{x},\mathbf{w}^*}(y, \mathbf{c})$ and derive the bound for $P_{\mathbf{x}}(y, \mathbf{c})$ as shown in Theorem 13.

Theorem 13 (Causal Bound with Substitute Interventions)**.** Given a set of variables corresponding to recoverable substitute interventions: $\mathcal{D} = \{\mathbf{W}|P_{\mathbf{x},\mathbf{w}}(y, \mathbf{c}) \text{ is recoverable}\}$, the

target conditional intervention $P_{\mathbf{x}}(y|\mathbf{c})$ is bounded by

$$
\begin{aligned}
L_w &= \max_{\mathbf{W}\in\mathcal{D}} \min_{\mathbf{w}^*\in\mathbf{W}} P_{\mathbf{x},\mathbf{w}^*}(y,\mathbf{c})/P_{\mathbf{x}}(\mathbf{c}) \\
U_w &= \min_{\mathbf{W}\in\mathcal{D}} \max_{\mathbf{w}^*\in\mathbf{W}} P_{\mathbf{x},\mathbf{w}^*}(y,\mathbf{c})/P_{\mathbf{x}}(\mathbf{c})
\end{aligned}
\tag{7.5}
$$

Algorithm 9 Finding Recoverable Substitute Interventions

function FindRSI($\mathbf{x}, \mathbf{c}, y, \mathcal{G}$)
Input: Causal graph $\mathcal{G}$, target intervention $P_{\mathbf{x}}(y, \mathbf{c})$.
Output: A valid variable set $\mathcal{D} = \{\mathbf{W}|P_{\mathbf{x},\mathbf{w}}(y,\mathbf{c})$ is recoverable and could be expressed in terms of biased observational distributions following Equation 3.3$\}$.
Initialize: $\mathcal{D} \leftarrow \emptyset$.
for all $\mathbf{W}$ such that $\mathbf{W} \cap \mathbf{X} = \emptyset$, starting with the smallest size of $\mathbf{W}$ **do**
    if a valid adjustment set can be found according to Theorem 2 **then**
        $\mathcal{D} = \mathcal{D} \cup \{\mathbf{W}\}$
    end if
end for

We list our procedure of finding all the recoverable substitute interventions in Algorithm 9. Basically the main function FindRSI in Algorithm 9 returns a set containing all admissible variables, each of which corresponding to a recoverable intervention with a larger intervention space.

Next, we give an illustration example on how to run our BCE algorithm to get prior causal bounds. Figure 7.2 shows a causal graph constructed from offline data, where nodes U_1, U_2 and X_1, X_2 depict user features and item features respectively, Y denotes the outcome variable, S denotes the selection variable, and I_1 denotes an intermediate variable. The bi-directed dashed line indicates there exist unobserved confounders that affect both I_1 and Y simultaneously. To bound the conditional causal effect $p_{x_1,x_2}(y|u_1,u_2)$ via c-component factorization, we first identify the set $\mathcal{D} = An(\mathbf{Y})_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{X}}} = \{Y, U_1, U_2\}$. The target intervention

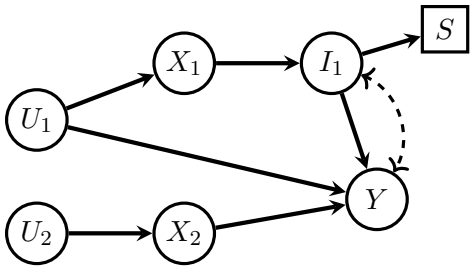


Figure 7.2: Causal graph for synthetic data.

could be expressed as

$$p_{x_1,x_2}(y|u_1,u_2) = p_{x_1,x_2}(y,u_1,u_2)/p(u_1,u_2) = (Q[Y]\cdot Q[U_1]\cdot Q[U_2])/p(u_1,u_2) \tag{7.6}$$

We then call RC* algorithm to bound each c-component and return the bound for each arm according to Theorem 12. For bounding causal effects via substitute interventions, we call FindRSI to find a valid variable set $\mathcal{D} = \{I_1\}$. According to Theorem 13, we can obtain the bound for each arm.

7.4 Online Bandit Learning with Prior Causal Bounds

In this section we show how to incorporate our derived causal bounds to online contextual bandit algorithms. We focus on the stochastic contextual bandit setting with linear reward function. Under the linear assumption, the binary reward is generated by $P(Y_t = 1) = \langle \boldsymbol{\theta}, \mathbf{x}_{t,a} \rangle + \eta_t$ where $\boldsymbol{\theta} \in \mathbb{R}^d$, $\mathbf{x}_{t,a} \in \mathbb{R}^d$ denotes the context vector related to the concatenation of user and arm feature vector at time t and the noise term η_t follows sub-Gaussian distribution for $t \in [T]$. Let $a_t = \text{argmax}_{a\in\mathcal{A}}\, \mathbb{E}[Y_{a\sim\pi(\mathbf{c}_t)}]$, the expected cumulative

regret of a policy π is defined as:

$$\mathbb{E}[R_\pi(T)] = \sum_{t=1}^{T} \langle \boldsymbol{\theta}, \mathbf{x}_{t,a^*} \rangle - \sum_{t=1}^{T} \mathbb{E}[Y_{a_t}]$$

We next conduct regret analysis and prove our strategy could consistently improve the long-term regret with the guide of the pruned arm set and a prior causal bound for each arm's reward distribution.

7.4.1 LinUCB Algorithm with Prior Causal Bounds

LinUCB [82] is one of the most widely used stochastic contextual bandit algorithms that assume the expected reward of each arm a is linearly dependant on its d-dimensional feature vector $\mathbf{x}_{t,a}$ with an unknown coefficient $\boldsymbol{\theta}_a$ at time t. We develop the LinUCB-PCB algorithm that includes a modified arm-picking strategy, clipping the original upper confidence bounds with the prior causal bounds obtained from the offline evaluation phase. Algorithm 10 shows the pseudo-code of our LinUCB-PCB algorithm. The truncated upper confidence bound shown in line 12 of Algorithm 10 contains strong prior information about the true reward distribution implied by the prior causal bound, thus leading to a lower asymptotic regret bound.

Theorem 14. Let $||\mathbf{x}||_2$ define the L-2 norm of a context vector $\mathbf{x} \in \mathbb{R}^d$ and

$$L = \max_{a, \mathbf{c} \in \{\mathcal{A}, \mathcal{C}\}, U_{a,\mathbf{c}} \geq \langle \boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}} \rangle} ||\mathbf{x}_{a,\mathbf{c}}||_2$$

The expected regret of LinUCB-PCB algorithm is bounded by:

$$R_T \leq Cd\sqrt{T}log(TL) \tag{7.7}$$

Algorithm 10 LinUCB algorithm with Prior Causal Bounds (LinUCB-PCB)

Input: Time horizon T, arm set $\mathcal{A}$, prior causal bounds $\{[L_{a,\mathbf{c}}, U_{a,\mathbf{c}}]\}_{a,\mathbf{c}\in\{\mathcal{A},\mathcal{C}\}}$, $\alpha \in \mathbb{R}^+$.
for t = 1,2,3,...,T **do**
  Observe contextual features $\mathbf{x}_{t,a} \in \mathbb{R}^d$ for all arms
  for $a \in \mathcal{A}$ **do**
    if a is new **then**
      $A_a \leftarrow \mathbf{I_d}$, $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$
    end if
    $\hat{\boldsymbol{\theta}}_a \leftarrow A_a^{-1}\mathbf{b}_a$
    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^{\mathrm{T}}\mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^{\mathrm{T}}A_a^{-1}\mathbf{x}_{t,a}}$
    $U_{t,a} \leftarrow U_{a,\mathbf{c}_t} : \mathbf{x}_{a,\mathbf{c}_t} \triangleq [\mathbf{x}_a, \mathbf{c}_t] = \mathbf{x}_{t,a}$
    $\overline{UCB}_a(t) \leftarrow \min\{p_{t,a}, U_{t,a}\}$ //Truncated UCB
  end for
  Pull arm $a_t \leftarrow argmax_{a\in\mathcal{A}}\overline{UCB}_a(t)$, and observe a reward r_{t,a_t}
  $A_{a_t} \leftarrow A_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^{\mathrm{T}}$, $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_{t,a_t}\mathbf{x}_{t,a_t}$
end for

where C is a suitably large constant.

We follow the standard procedure of deriving the expected regret bound for linear contextual bandit algorithms in [50] and [112]. Please refer to Appendix for the proof. We next discuss the potential improvement in regret that can be achieved by applying LinUCB-PCB algorithm in comparison to original LinUCB algorithm.

Theorem 15. If there exists an arm a such that $U_{a,\mathbf{c}_t} < \langle\boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}_t}\rangle$ at a round $t \in [T]$, LinUCB-PCB is guaranteed to achieve lower cumulative regret than LinUCB algorithm.

Proof. To prove Theorem 15, we first introduce a Lemma shown as follows:

Lemma 7 (Reduced Arm Exploration Set)**.** Given an arm a with $U_{a,\mathbf{c}} < \langle\boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}}\rangle$, we have $P(a_t = a) = 0, \forall t \in [T]$.

We prove Lemma 7 by contradiction. Given an arm a with $U_{a,\mathbf{c}} < \langle\boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}}\rangle$ and suppose the agent pulls the arm a at round t. Based on the optimism in the face of uncertainty

(OFU) principle we have $\langle \boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}} \rangle < \overline{UCB}_{a^*,\mathbf{c}} < \overline{UCB}_{a,\mathbf{c}} \leq U_{a,\mathbf{c}}$, which contradicts with the fact $U_{a,\mathbf{c}} < \langle \boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}} \rangle$. Thus for all the rounds $t \in [T]$ we have $P(a_t = a) = 0$.

Lemma 7 basically states that although the optimal reward given a user context is unknown to the agent apriori, based on the exploration strategy enhanced with the information provided by the prior causal bounds, LinUCB-PCB will not pull the arms that are sub-optimal implied by their upper causal bounds at each round, thus leading to a reduced exploration arm set and a lower value of L. □

We further define the total number of sub-optimal arms implied by prior causal bounds as

$$N^-_{pcb} = \sum_{a,\mathbf{c} \in \{\mathcal{A},\mathcal{C}\}} \mathbb{1}_{U_{a,\mathbf{c}} - \langle \boldsymbol{\theta}, \mathbf{x}_{a^*,\mathbf{c}} \rangle < 0}$$

Note that the value of N^-_{pcb} depends on the accuracy of the causal upper bound for each arm. This is because if the estimated causal bounds are more concentrated, that is, $U_{a,\mathbf{c}}$ is close to $\langle \boldsymbol{\theta}, \mathbf{x}_{a,\mathbf{c}} \rangle$ for each $a, \mathbf{c} \in \{\mathcal{A}, \mathcal{C}\}$, there will be more arms whose prior causal upper bound is less than the optimal mean reward, thus N^-_{pcb} will increase accordingly. A large N^-_{pcb} value implies less uncertainty regarding the sub-optimal arms implied by prior causal bounds. As a result there are in general less arms to be explored and the L value will decrease accordingly, leading to a more significant improvement by applying LinUCB-PCB algorithm.

7.4.2 OAM-PCB Algorithm

Recently, [113] developed one state-of-the-art contextual linear bandit algorithm based on the optimal allocation matching (OAM) policy. It alternates between exploration and exploitation based on whether or not all the arms have satisfied the approximated allocation rule. We investigate how to incorporate prior causal bounds in OAM and develop the new

OAM-PCB algorithm shown in the following algorithm.

Algorithm 11 Optimal Allocation Matching with Prior Causal Bounds

Input: Time horizon T, arm set $\mathcal{A}$, exploration parameter ϵ_t, exploration counter $s(d) = 0$, prior causal bounds $\{[L_{a,\mathbf{c}}, U_{a,\mathbf{c}}]\}_{a,\mathbf{c}\in\{\mathcal{A},\mathcal{C}\}}$.
for $t = 1$ **to** T **do**
  Solve the optimization problem in Equation 7.13 based on the estimated gap $\widehat{\Delta}(t-1)$.
  if $||a||^2_{G_{t-1}^{-1}} \leq \max\{\frac{\widehat{\Delta}^2_{min}(t-1)}{f_n}, \frac{(\widehat{\Delta}_a^{c_t}(t-1))^2}{f_n}\},\ \forall a \in \mathcal{A}$, **then**
    // Exploitation
    for Each arm $a \in \mathcal{A}$ **do**
      $\hat{\mu}_a(t-1) = \max\{\min\{U_{a,\mathbf{c}}, a^\top \hat{\boldsymbol{\theta}}_{t-1}\}, L_{a,\mathbf{c}}\}$
    end for
    Pull arm $a_t = \text{argmax}_{a\in\mathcal{A}}\, \hat{\mu}_a(t-1)$.
  else
    //Wasted (LinUCB) Exploration
    $s(t) = s(t-1) + 1$
    if $N_a(t-1) \geq \min(T_a(\widehat{\Delta}(t-1)), f_n/(\widehat{\Delta}_{min}(t-1)))^2,\ \forall a \in \mathcal{A}$, **then**
      Pull an arm following Equation 7.8.
    else
      Calculate b_1, b_2 following Equation 7.9.
      if $N_{b_2}(t-1) \leq \epsilon_t s(t-1)$ **then**
        // Forced Exploration
        Pull arm $a_t = b_2$.
      else
        // Unwasted Exploration
        Pull arm $a_t = b_1$.
      end if
    end if
  end if
  Observe reward and update $\widehat{\boldsymbol{\theta}}_t, \widehat{\Delta}_a^{c_t}(t), \widehat{\Delta}_{min}(t)$
end for

As shown in Algorithm 11, at each round in both exploitation and wasted exploration scenarios, we truncate the upper confidence bound for each arm with the upper causal bound to obtain a more accurate estimated upper bound:

$$\begin{aligned}\widehat{UCB}_a(t-1) &= \min\Big\{a^\top \hat{\theta}_{t-1} + \sqrt{f_{n,1/s(t))^2}}||a||_{G_{t-1}^{-1}}, U_{a,\mathbf{c}}\Big\} \\ a_t &= \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \widehat{UCB}_a(t-1)\end{aligned} \tag{7.8}$$

where $G_t = \sum_{s=1}^{t} \mathbf{X}_s \mathbf{X}_s^\top$. The algorithm then explores by computing two arms:

$$\begin{aligned}b_1 &= \underset{a\in\mathcal{A}}{\operatorname{argmin}}\, \frac{N_a(t-1)}{\min(T_a^{c_t}(\widehat{\Delta}(t-1)), f_n/\widehat{\Delta}^2_{min}(t-1))} \\ b_2 &= \underset{a\in\mathcal{A}}{\operatorname{argmin}}\, N_a(t-1)\end{aligned} \tag{7.9}$$

where $f_{n,\delta} = 2(1+1/log(n))log(1/\delta) + cdlog(dlog(n))$. c is a constant and we denote $f_n = f_{n,1/n}$ for simplicity. $N_a(T)$ denotes the number of pulls of arm a up to time T. For any $\tilde{\Delta} \in [0,\infty)^{|\cup_{a\in\mathcal{A}}|}$ that is an estimate of Δ, $T(\tilde{\Delta})$ could be treated as an approximated allocation rule in contrast to the optimal allocation rule, which is defined as a solution to the following optimization problem:

$$\min_{(T_a^c)_{a,c}\in[0,\infty]} \sum_{c=1}^{|\mathcal{C}|} \sum_{a\in\mathcal{A}} T_a^c \tilde{\Delta}_a^c \tag{7.10}$$

subject to

$$||x||^2_{H_T^{-1}} \leq \frac{\Delta_a^2}{f_n}, \forall a \in \mathcal{A}, c \in [|\mathcal{C}|] \tag{7.11}$$

and $H_T = \sum_{c=1}^{|\mathcal{C}|} \sum_{a\in\mathcal{A}} T_a^c a a^\top$ is invertible.

We next derive the asymptotic regret bound of OAM-PCB and show our theoretical results.

Theorem 16 (Regret of OAM-PCB)**.** Given causal bounds $\mathbb{E}[Y_{a,\mathbf{c}}] \in [L_{a,\mathbf{c}}, U_{a,\mathbf{c}}]$ over $a \in \mathcal{A}$, the asymptotic regret of optimal allocation matching policy augmented with prior causal

bounds is bounded by

$$R_{\pi_{\text{oam}}}(T) \leq log(T) \cdot \mathcal{V}(\boldsymbol{\theta}, \mathcal{A}) \tag{7.12}$$

where $\mathcal{V}(\boldsymbol{\theta}, \mathcal{A})$ denotes the optimal value of the optimization problem defined as:

$$\inf_{\alpha_{a,\mathbf{c}} \in [0,\infty]} \sum_{c=1}^{|\mathcal{C}|} \sum_{a:\mathbf{U}_{a,c} \geq \mu_c^*} \alpha_{a,c} \Delta_a^c \tag{7.13}$$

subject to the constraint that for any context $\mathbf{c}$ and suboptimal arm $a \in \mathcal{A}$,

$$a^\top \Bigg(\sum_{c=1}^{|\mathcal{C}|} \sum_{a:\mathbf{U}_{a,c} \geq \mu_c^*} \alpha_{a,c} a a^\top \Bigg)^{-1} a \leq \frac{(\Delta_a^c)^2}{2} \tag{7.14}$$

In Theorem 16, c indexes a domain value of the context vector $\mathbf{C}$, $\mu_c^* = \langle \boldsymbol{\theta}, a_c^* \rangle$ is the mean reward of the best arm given context c, $\Delta_a^{\mathbf{c}} = \langle \boldsymbol{\theta}, a_c^* - a \rangle$ is the suboptimality gap and $\Delta_{min} = \min_{\mathbf{c} \in [|\mathcal{C}|]} \min_{a \in \mathcal{A}, \Delta_a^{\mathbf{c}} > 0} \Delta_a^{\mathbf{c}}$. Please refer to Appendix for the proof.

7.4.3 Non-contextual Setting

Our prior causal bounds can also be incorporated into non-contextual bandits. We derive the UCB-PCB algorithm, a non-contextual upper confidence bound-based multi-arm bandit algorithm enhanced with prior causal bounds, and give its pseudo-code and regret analysis as follows:

Theorem 17 (Regret of UCB-PCB algorithm)**.** Suppose the noise term is 1-subgaussian distributed, let $\delta = 1/T^2$, the cumulative regret for k-arm bandit bounded by:

$$R(T) = 3 \sum_{a=1}^{k} \Delta_a + \sum_{a:\mathbf{U}_a \geq \mu^*} \frac{16 log(T)}{\Delta_a}$$

Algorithm 12 UCB Algorithm with Prior Causal Bounds (UCB-PCB)

Input: Time horizon T, arm set $\mathcal{A}$, causal Graph $\mathcal{G}$, prior causal bounds $\{[L_a, U_a]\}_{a\in\mathcal{A}}$.
Initialization: Values assigned to parent variables: $\hat{\mu}_a$, $T_a(0) = 0$.
for $t = 1, 2, 3, ..., T$ **do**
    for each arm $a \in \mathcal{A}$ **do**
        $UCB_a(t-1) = \hat{\mu}_a(t-1) + \sqrt{\frac{2log(1/\delta)}{N_a(t-1)}}$.
        $\widehat{UCB}_a(t-1) = \max\{\min\{U_a, UCB_a(t-1)\}, L_a\}$
    end for
    Pull arm $a_t = \text{argmax}_{a\in\mathcal{A}} \widehat{UCB}_a(t-1)$
    Observe Y_t and update upper confidence bounds accordingly.
end for

where Δ_a denotes the reward gap between arm a and the optimal arm a^*

Notice that the improvement could be significant if we obtain concentrated causal bounds from observational data and consequently exclude more arms whose causal upper bounds are less than μ^*. Please refer to Appendix for the proof.

7.5 Empirical Evaluation

In this section, we conduct experiments to validate our proposed methods. We use the synthetic data generated following the graph structure in Figure 7.2. We generate 30000 data points to simulate the confounded and selection biased setting shown in Table 7.1. After conducting the preferential exclusion indicated by the selection mechanism, there are approximately 15000 data points used for offline evaluation.

Table 7.1: Conditional probabilities for synthetic data.

Variables $V_i \in \mathbf{V}$	Distributions $P(V_i = 1\|Pa_{V_i})$
U_1	$P(U_1 = 1) = 0.4$
U_2	$P(U_2 = 1) = 0.6$
X_1	$P(X_1 = 1\|U_1 = u_1) = (\mathbb{1}_{\{u_1=1\}} + 0.5)/2$
X_2	$P(X_2 = 1\|U_2 = u_2) = (\mathbb{1}_{\{u_2=1\}} + 0.3)/2$
I_1	$P(I_1 = 1\|X_1 = x_1, C_1 = c_1) = 0.3 + (\mathbb{1}_{\{x_1=1\}} + \mathbb{1}_{\{c_1=1\}})/4$
Y	$P(Y = 1\|C_1 = c_1, U_1 = u_1, X_2 = x_2, I_1 = i_1) = (\mathbb{1}_{\{c_1=1\}} + \mathbb{1}_{\{u_1=1\}} + \mathbb{1}_{\{x_2=1\}} + \mathbb{1}_{\{i_1=1\}})/6 + 0.1$
C_1	$P(C_1 = 1) = 0.5$
S	$P(S = 1\|I_1 = i_1) = 0.8$ if $i_1 = 1$ $P(S = 1\|I_1 = i_1) = 0.1$ if $i_1 = 0$

Offline Evaluation We use our BCE algorithm to obtain the bound of each arm based on the input offline data and compare the causal bound derived by the algorithm with the estimated values from two baselines: an estimate that is derived based on Equation 3.2 which only takes into account confounding bias (Biased), and a naive conditional probability estimate derived without considering both confounding and selection biases (CP). Table 7.2 shows the comparison results in offline evaluation phase. *lb* and *ub* denote the lower bound and upper bound derived by our BCE algorithm for the conditional causal effect related to a value of the context vector. We also report the visualized comparison results on causal bound and the estimated values among 16 different values of the context vector in Figure 7.3. The comparison results show our BCE algorithm contains the ground-truth causal effect (denoted by the red lines in the figure) for each value of the context vector. On the contrary, the estimated values from CP and Biased baselines deviate from the true causal effect in the presence of compound biases. The experimental results reveal the fact that neglecting any bias will inevitably lead to an inaccurate estimation of the target causal effect.

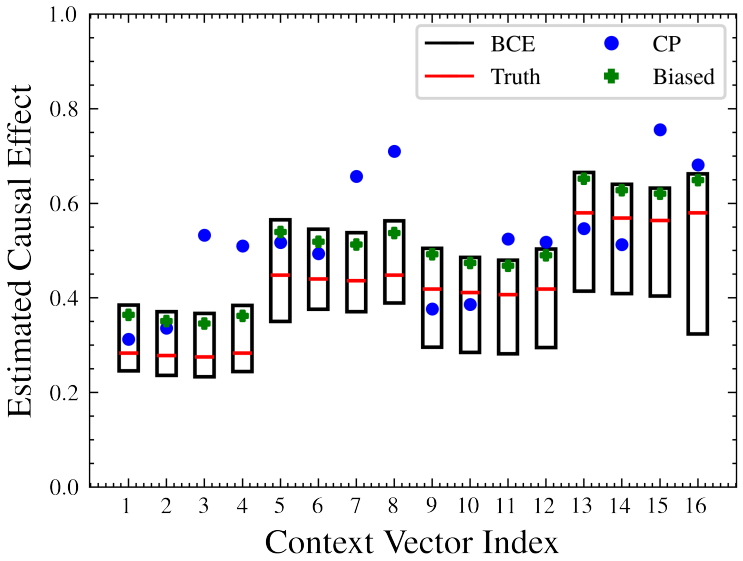


Figure 7.3: Comparison results for offline evaluation under confounding and selection biases.

Table 7.2: Reward estimation for the synthetic data.

Context Index	**CP**	**Biased**	**BCE**		**Truth**
			lb	*ub*	
1	0.313	0.364	0.246	0.385	0.283
2	0.336	0.351	0.236	0.371	0.278
3	0.533	0.346	0.233	0.367	0.275
4	0.510	0.362	0.244	0.384	0.283
5	0.517	0.539	0.350	0.565	0.448
6	0.494	0.519	0.376	0.545	0.440
7	0.657	0.513	0.371	0.538	0.436
8	0.710	0.537	0.389	0.563	0.448
9	0.377	0.492	0.296	0.505	0.419
10	0.386	0.474	0.285	0.486	0.411
11	0.525	0.468	0.282	0.480	0.407
12	0.518	0.490	0.295	0.503	0.419
13	0.547	0.652	0.414	0.665	0.580
14	0.513	0.628	0.409	0.640	0.569
15	0.756	0.620	0.404	0.632	0.564
16	0.681	0.649	0.324	0.662	0.580

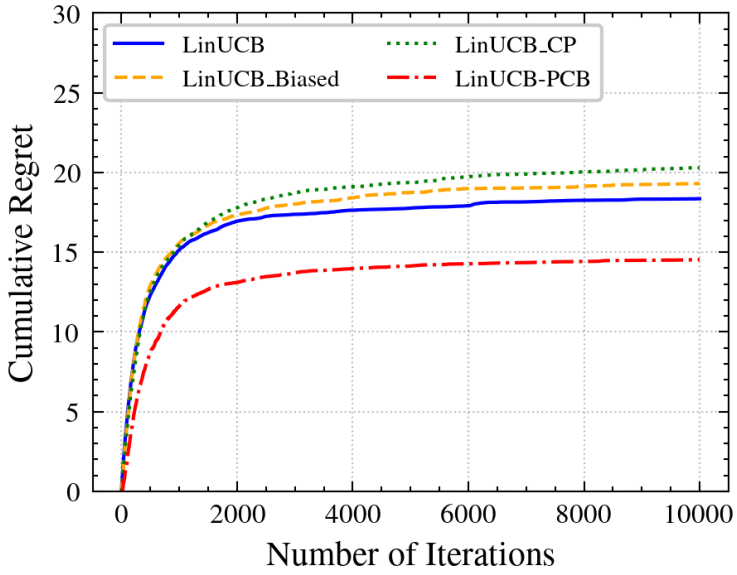


Figure 7.4: Comparison results for contextual linear bandit.

Online Bandit Learning We use 15000 samples from the generated data to simulate the online bandit learning process. In Figure 7.4, we compare the performance of our LinUCB-PCB

algorithm regarding cumulative regret with the following baselines: LinUCB, LinUCB_Biased and LinUCB_CP, where LinUCB_Biased and LinUCB_CP are LinUCB-based algorithms initialized with the estimated reward for each value of the context vector (arm) from the Biased and CP baselines in the offline evaluation phase. Each curve denotes the regret averaged over 100 simulations to approximate the true expected regret. We find that LinUCB-PCB achieves the lowest regret compared to the baselines. Moreover, both LinUCB_Biased and LinUCB_CP perform worse than the LinUCB baseline, which is consistent with the conclusion from our theoretical analysis that blindly utilizing biased estimates from offline data could negatively impact the performance of online bandit algorithms.

7.6 Summary

This chapter studied bounding conditional causal effects in the presence of confounding and sample selection biases using causal inference techniques and utilizes the derived bounds to robustly improve online bandit algorithms. We presented two novel causal-based techniques to derive a bound for conditional causal effects given offline data with compound biases. We developed contextual and non-contextual bandit algorithms that leverage the derived causal bounds and conduct their regret analysis. Theoretical analysis and empirical evaluation demonstrate the improved regrets of our algorithms.

8 Achieving Fairness through Multiple Causes Discrimination Analysis

8.1 Introduction

Discrimination or unfairness has been a paramount concern in many big data applications like employment, credit, and insurance. How to strike a balance between accurate predictions and fairness is receiving increasing attention in the machine learning field. Causal modeling based fair learning models [17, 18, 6, 19, 20, 9, 16], which are based on Pearl's (probabilistic) causal model [21], have been developed to capture and quantify different fairness measures (e.g., direct/indirect discrimination, counterfactual fairness) through counterfactual inference along specific paths in causal graphs. However, most existing causal modeling based fair learning research focuses on single cause effect of one protected attribute on decision.

In this chapter, we focus on discrimination discovery when multiple protected attributes and redlining attributes are present in addition to other covariates. Protected attributes refer to certain characteristics that are the subject of discrimination analysis, such as race, gender, marital status, whereas redlining attributes (e.g., zipcode in loan application) are a set of attributes that cannot be legally justified if used in decision-making. We are interested in evaluating the causal effects of those protected and redlining attributes on the decision (the outcome variable). We regard those protected and redlining attributes as multiple causes of the outcome variable.

One big challenge for causal modeling is to deal with hidden variables. Most previous works [17, 18, 6] based on Pearl's structural causal modeling make the Markovian assump-

tion (i.e., there is no hidden variable that affects both protected attribute and decision) to facilitate the causal inference. In open world scenario, the existence of the hidden variable mentioned above (also known as hidden confounder) is an inescapable fact. Simply ignoring the presence of these variables in a causal model can lead to erroneous conclusions about the causal relationship among endogenous variables. Furthermore, causal effects are not computable from observational data in some situations known as the unidentifiable situations. Those methods have to make simplified assumptions to avoid the unidentifiable situations, but the validity issue of the assumptions imposes uncertainty on the performance and reliability of these methods.

To deal with hidden confounders, we adopt the potential outcome framework [114] and leverage the state-of-the-art *deconfounder* algorithm [115] to do causal inference under multiple causes. The potential outcome framework focuses on the causal relationship between a treatment and its effect given other covariates. Potential outcomes are expressed in the form of counterfactual conditional statements of the case conditional on a prior event occurring. For each instance, only one potential outcome can be observed. The *deconfounder* algorithm combines unsupervised machine learning and predictive model checking to perform causal inference in multiple-cause settings. Its main idea is to infer a latent variable as a substitute for unobserved confounders and then use that substitute to perform causal inference. Combining them, we are able to relax the Markovian assumption and avoid the unidentifiability issue in structural causal modeling approaches. We compare our approach with the widely adopted structural casual modeling approach [21] in our empirical evaluation on both synthetic data and real data. Empirical evaluation results demonstrate the effectiveness of the proposed approach.

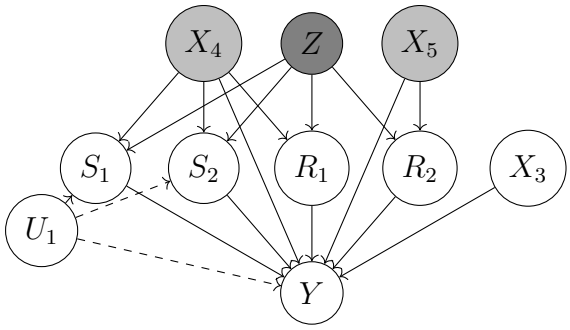


Figure 8.1: Graph structure under multiple treatments setting.

8.2 Modeling Multi-cause Discrimination

8.2.1 Problem Formulation

Assume that there is a population over the space $\mathbf{S} \times \mathbf{X} \times Y$ where $\mathbf{S}$ are the protected attributes, Y is the decision attribute, and $\mathbf{X}$ are the covariates. The underlying mechanism that determines the values of all the attributes is represented by a causal model. In practice the causal model is unknown, but we can observe a training dataset $\mathcal{D} = \{(\mathbf{s}_i, \mathbf{x}_i, y_i); i = 1, \cdots, n\}$ drawn from the population and consider it as in fact generated by the causal model. Among covariates $\mathbf{X}$ there is a set of attributes that cannot be legally justified if used in the decision making process, referred to as the redlining attributes denoted by $\mathbf{R}$. In discrimination discovery, we are interested in the causal effects of protected attributes $\mathbf{S}$ and redlining attributes $\mathbf{R}$ on the decision Y. Traditional causal inference usually uses an upper letter A to represent a single treatment variable. In this chapter we generalize the notation A to a bold letter $\mathbf{A} = \{A_1, ..., A_m\} = \mathbf{S} \cup \mathbf{R}$ to represent m possible multiple treatments and aim to estimate their causal effect on Y. In addition to the observed attributes ($\mathbf{S}$, $\mathbf{X}$, Y), there may exist a set $\mathbf{U}$ containing unobserved hidden confounders.

Figure 8.1 shows an illustrative example under the multiple treatments setting. In Figure 8.1, the shaded node Z denotes the substitute variable from the deconfounder algorithm.

Y is the outcome variable. $\mathbf{S} = \{S_1, S_2\}$ is the protected attributes set. $\mathbf{X} = \{R_1, R_2, X_3, X_4, X_5\}$ is the covariate set, among which lies the redlining attribute set $\mathbf{R} = \{R_1, R_2\}$. $\mathbf{A} = \{S_1, S_2, R_1, R_2\}$ represents the multiple treatments set. X_3 is a pre-treatment covariate, X_4 is a multiple-cause confounder that affects more than two causes and the outcome variable simultaneously and X_5 is a single-cause confounder that affects exactly one cause and the outcome variable. $\mathbf{U} = \{U_1\}$ is hidden variable set and U_1 is also a multi-cause hidden confounder.

The ATE of $\mathbf{A}$ on Y under multiple treatments scenario can then be expressed as $\mathbb{E}[Y_i(\mathbf{a})] - \mathbb{E}[Y_i(\mathbf{a}')]$ where $\mathbf{a}$ and $\mathbf{a}'$ are two treatment configurations. However, the presence of hidden variables (especially hidden confounders such as U_1 in Figure 8.1) can make the estimate of ATE inaccurate. In this chapter, we develop a Multi-Cause Discrimination Analysis (MCDA) algorithm to derive the ATE of $\mathbf{A}$ on Y with the presence of hidden confounders. Our MCDA involves two phases. In phase 1, we apply the deconfounder algorithm [115] to infer a latent variable as a substitute for unobserved confounders and then use that substitute to perform causal inference. The shaded node Z in Figure 8.1 is the substitute variable derived from the deconfounder algorithm. Note that the deconfounder algorithm relaxes the assumption of no hidden confounders to that of no single-cause confounders, which significantly improves the applicability of causal inference. In phase 2, we apply the propensity score approach, in particular, the inverse probability of treatment weighting method, to estimate the causal effect. We emphasize that the deconfounder provides a checkable approach to estimating closer-to-truth causal effects as its weakened assumption is more likely held in practice. The causal inference based on the combination of the deconfounder and the propensity score approach is more appropriate for analyzing the simultaneous effects of multiple protected and redlining attributes on the decision in discrimination discovery and

fair learning.

8.2.2 The Deconfounder Algorithm

In [115], the authors proposed the deconfounder algorithm to conduct causal inference under multiple treatments setting. The algorithm relaxes the strong ignorability assumption to single ignorability assumption. The single ignorability assumes that there are no unobserved single-cause confounders. Roughly speaking, single ignorability implies that we observe all the confounders that affect exactly one of the causes and the outcome variable. The assumption is much weaker than the strong ignorability that requires all confounders are observed. For those application problems which may involve multiple causes in the model, confounders are unlikely to have effect on only one cause. Hence, single ignorability is more likely to be satisfied in practice. The deconfounder algorithm can be divided into two parts: the assignment model and the outcome model.

8.2.2.0.1 Assignment model The assignment model is basically a factor model of the assigned causes. The main point is that if we can infer a reasonable latent variable Z (shown in Figure 8.1) such that each cause is conditionally independent given Z, then Z could be regarded as a substitute confounder. This is because if there exists any other multiple-cause confounder, it will break the conditional independence between treatments. Through the use of substitute confounders, we can get rid of the barrier of unobserved confounders when the single ignorability assumption is satisfied.

To implement the deconfounder algorithm we firstly define and fit a probabilistic factor model to capture the joint distribution of causes $p(a_1, \cdots, a_m)$. The factor model posits per-instance latent variables Z_i and uses them to model the assigned causes. The

model can be represented as:

$$\begin{aligned} Z_i &\sim p(\cdot|\alpha) \quad i = 1, \cdots, n \\ A_{ij}|Z_i &\sim p(\cdot|(z_i, \theta_j) \quad j = 1, \cdots, m \end{aligned} \tag{8.1}$$

where α parameterizes the distribution of Z_i and θ_j parameterizes the per-cause distribution of A_{ij}. Generally Z_i can be multi-dimensional and factor models include many methods from Bayesian statistics and probabilistic machine learning. In our chapter, we use probabilistic principal component analysis proposed by [116] as a factor model. Its structure can be expressed as follows:

$$\begin{aligned} Z_{ik} &\sim \mathcal{N}(0, \lambda^2) \quad k = 1, \cdots, K \\ A_{ij}|Z_i &\sim \mathcal{N}(z_i^T \theta_j, \sigma^2) \quad j = 1, \cdots, m \end{aligned} \tag{8.2}$$

In Equation 8.2 both z_i and θ_j are real-valued K-dimension vectors, λ and σ are hyper parameters. Since the deconfounder rests on finding a good factor model to capture the dependent relationship of all assigned causes, posterior predictive checks are used to assess the fidelity of the model. We use the fitted factor model to calculate the posterior distribution $p(Z_i|\mathbf{A}_i)$ by applying Bayes' theorem and then derive the conditional expectation $\hat{Z}_i = \mathbb{E}[Z_i|\mathbf{A}_i]$ as the approximation of Z_i. Since the factor model captures the population distribution of assigned causes, we have essentially discovered a variable (set) that captures all multiple-cause confounders.

8.2.2.0.2 Outcome model The outcome model aims to estimate causal effects given the information from the augmented dataset $\{\mathbf{A}, Z\}$, where Z is the substitute confounder inferred in assignment model. It can be formulated as the function $f(\mathbf{a}, z) = \mathbb{E}[Y_i(\mathbf{A}_i)|\mathbf{A}_i =$

$\mathbf{a}, Z_i = z]$. From the equation above, we can see the outcome modal could have various expression form. Any reasonable model can be fitted here if it passes the model checking and has a good performance on approximating the causal effects. For example, we can simply fit a linear regression model. If the outcome variable is binary, we can apply logistic regression and each coefficient of the logistic regression corresponds to the causal odds ratio of a certain covariate.

We emphasize that causal effects can be accurately estimated since strong ignorability is guaranteed with the help of substitute confounder derived from phase 1. In the illustrative example shown in Figure 8.1, after we infer a reasonable substitute confounder Z from the assignment model, both observed and unobserved multiple confounders (such as X_4 and U_1) can be well represented by the substitute confounder Z. Hence we can apply classical causal inference methods (e.g., matching and weighting methods) without worrying any biases caused by those hidden confounders. In this chapter we focus on inverse probability of treatment weighting method as one legitimate choice of the outcome model.

8.2.3 Inverse Probability of Treatment Weighting

Definition 9 (Propensity Score)**.** Propensity score, $e(x) = Pr(A = 1|\mathbf{X} = \mathbf{x})$, is the conditional probability of receiving treatment A given the pretreatment variables $\mathbf{X}$.

With the help of the propensity score, we can divide individuals with the same propensity score into one stratum and treat those strata as randomized controlled trial. Treatment effect is then automatically identified within each stratum and simple methods can be applied to obtain unbiased estimation of the average treatment effect. In statistics, the propensity score is usually estimated using regression models. There are four predominating propensity score methods used for removing the confounding bias when estimating causal effects: propen-

sity score matching (PSM), propensity score stratification, inverse probability of treatment weighting (IPTW) and covariate adjustment using the propensity score.

In this chapter, we mainly focus on the IPTW method. The main idea is to use inverse propensity score as the weight for each individual. By multiplying such a weight to all the data points we can create a pseudo-population to synthesize randomized controlled experiments and estimate the average treatment effect unbiasedly. Specifically, the weight for individual i can be defined as

$$\omega_i = \frac{I(A_i)}{e(x_i)} + \frac{1 - I(A_i)}{1 - e(x_i)} \tag{8.3}$$

where $e(x_i)$ is the propensity score for i-th individual, $I(A_i)$ is an indicator variable denoting whether i-th individual received treatment.

After weighting procedures we generate a balanced dataset such that each individual has the same chance to receive the treatment. The average treatment effect can thus be estimated by a naive estimator shown in Equation 8.4 and Equation 8.5.

$$\widehat{ATE} = \frac{1}{N_{A=1}} \sum_{i:A_i=1} w_i Y_i - \frac{1}{N_{A=0}} \sum_{i:A_i=0} w_i Y_i \tag{8.4}$$

$$N_{A_i=1} = \sum \frac{A_i}{e(x_i)} \qquad N_{A_i=0} = \sum \frac{1 - A_i}{1 - e(x_i)} \tag{8.5}$$

$N_{A_i=1}$ and $N_{A_i=0}$ denote the number of instances under treatment and control. In our chapter, we generalize the treatment variable A to multiple treatments set $\mathbf{A}$ and use $\mathbf{A} = \mathbf{a}$ to represent a certain treatments configuration. The selection of covariates to condition on will also influence the estimated propensity scores, and IPTW method may be sensitive

to whether the propensity score has been accurately estimated. However, our multi-cause discrimination analysis combines the potential outcome framework and the deconfounder algorithm to deal with the hidden confounders.

8.3 Empirical Evaluation

We implement our MCDA algorithm and compare with the traditional IPTW method to evaluate how the deconfounder can improve the estimation of causal effects. Furthermore, we compare with the structural casual model [21], denoted as SCM. We use Tetrad [105] to learn the causal graph and then calculate the causal effect using the truncated factorization.

8.3.1 Synthetic Data

We generate $10,000$ data points following the data generating process:

$$H_1, H_2 \overset{iid}{\sim} \mathcal{U}(0,1)$$

$$A_1, A_2 \overset{iid}{\sim} \mathcal{B}(0, f(H_1, H_2))$$

$$Y \sim \mathcal{B}(0, g(H_1, H_2, A_1, A_2)) \tag{8.6}$$

Here $\mathcal{U}$ refers to uniform distribution and $\mathcal{B}$ refers to binomial distribution. f and g are functions with linear form. We consider A_1,A_2 as treatment variables, Y as outcome variable, and H_1,H_2 as confounders. As we have the full knowledge and the ground truth here, we can compare the performance of different methods under multiple scenarios (based on whether the confounders are observed or hidden) against the ground truth.

We observe that the graph structure of the synthetic data satisfies the back-door criterion. Hence we apply the structural equation model and use the truncated factorization

Table 8.1: Causal effects for synthetic data where the most accurate estimates are highlighted.

Causal effect	H_1,H_2 observed	H_1 hidden	H_1,H_2 hidden
Ground Truth	0.3982	0.3982	0.3982
SCM	**0.3982**	0.7241	0.7729
IPTW	0.4063	0.7732	0.6978
MCDA	0.5763	**0.5763**	**0.5763**

to get the true causal effect of X_1 and X_2 on Y. We measure the causal effect between two different treatment configurations: $(X_1, X_2) = (1, 1)$ and $(X_1, X_2) = (0, 0)$. The ground truth ATE is 0.3982. Our evaluation focuses on the scenario with H_1 and H_2 hidden. We see this scenario has two treatments and no single-cause confounder. As shown in the last column of Table 8.1, our MCDA algorithm achieves more accurate estimate (0.5763) than the IPTW (0.6978) and SCM (0.7729) compared with the ground truth (0.3982). We also conduct comparisons under the scenario with only H_1 hidden. As shown in the third column of Table 8.1, MCDA achieves the best estimate. We also show the scenario with both H_1 and H_2 observed. Note that in this scenario, there are no hidden variables. It is not surprising that SCM and IPTW outperform MCDA because of no hidden variables in this scenario.

8.3.2 Adult Dataset

We also use the adult dataset [86] that contains $65,123$ records with 11 attributes. We assume there are no unknown variables associated with this dataset. Under this assumption, we also have the ground truth. We binarize the categorical variables due to the data sparsity issue. We then apply the PC algorithm in Tetrad to build the causal graph. Since *native country*, *sex*, *age*, *race* are unlikely to be caused by other covariates, we set them in the first tier. The built causal graph is shown in Figure 8.2. We take *income* as the outcome variable, *sex* as the protected attribute, *workclass* and *relationship* as two

redlining attributes, *education*, *occupation*, *hours*, *marital-status* as observed covariates, and *native country* and *age* as two unobserved covariates. The causal effect can be calculated accurately if we know the whole knowledge of the graph. We may still estimate the causal effects using structural equation model when there exits hidden confounders, which is usually the case in observational study (SCM). From the graph structure we can see that there are no single-cause confounders. Hence our MCDA algorithm is applicable.

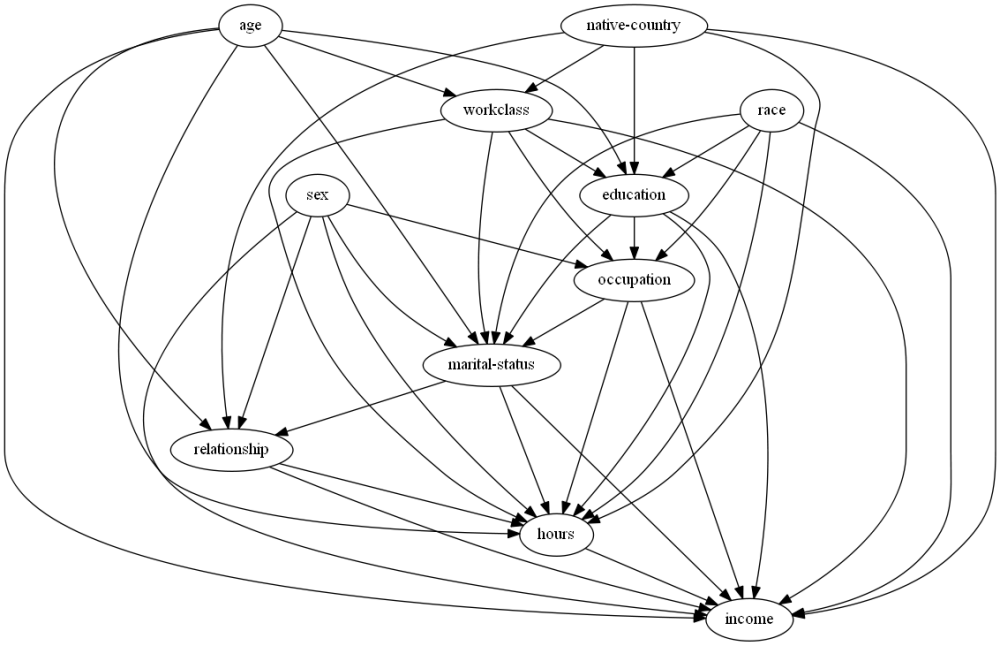


Figure 8.2: Causal graph for Adult Dataset.

Table 8.2: Comparison result from adult dataset, where A_1, A_2, and A_3 correspond to *workclass*, *relationship*, and *sex*.

Treatment Configuration	MCDA	SCM	Ground Truth
$(A_1, A_2, A_3) = (0, 0, 0)$	0.277	0.464	0.275
$(A_1, A_2, A_3) = (0, 0, 1)$	0.526	0.397	0.610
$(A_1, A_2, A_3) = (0, 1, 0)$	0.185	0.333	0.201
$(A_1, A_2, A_3) = (0, 1, 1)$	0.434	0.250	0.288
$(A_1, A_2, A_3) = (1, 0, 0)$	0.446	0.493	0.400
$(A_1, A_2, A_3) = (1, 0, 1)$	0.695	0.481	0.673
$(A_1, A_2, A_3) = (1, 1, 0)$	0.354	0.352	0.250
$(A_1, A_2, A_3) = (1, 1, 1)$	0.603	0.088	0.362

We focus on estimating the causal effects with two redlining attributes *workclass*

and *relationship* and the sensitive attribute *sex*. We emphasize our MCDA algorithm can analyze multi-cause effects simultaneously. Table 8.2 shows the expected potential outcome value for each treatment configuration of *workclass*, *relationship*, and *sex*. By applying average treatment effect formula, we can calculate the average treatment effect between any pair of two configurations. For example, the causal effect between two different treatment configurations $(A_1, A_2, A_3) = (1, 0, 1)$ and $(A_1, A_2, A_3) = (0, 0, 0)$ is 0.418 $(0.695 - 0.277)$, which is also close to the ground truth 0.398. The result shows that MCDA significantly outperforms the SCM and is more robust to unmeasured confounding. In other words, we can easily conduct counterfactual analysis and answer "what-if" questions in causal inference, which is imperative for exploration based fair learning.

8.4 Summary

In this chapter, we developed one approach based on the potential outcome framework to analyze the discrimination effects of protected and redlining attributes on the decision. The developed approach is based on the potential outcome framework and combines the deconfounder and inverse probability of treatment weighting. It can better handle the presence of hidden confounders and can lead to a more robust estimate of causal effects. We have empirically compared our approach with the structural causal modeling based approach and experimental results demonstrated the advantages of the proposed approach. The early version of this work is published at SBP-BRiMS 2020 [117].

9 Achieving Fairness through Equality of Effort

9.1 Introduction

Fair machine learning is receiving an increasing attention in machine learning fields. Discrimination is unfair treatment towards individuals based on the group to which they are perceived to belong. The first endeavor of the research community to achieve fairness is developing correlation or association-based measures, including demographic disparity (e.g., risk difference), mistreatment disparity, calibration, etc. [1, 2, 3, 4, 5], which mainly focus on discovering the disparity of certain statistical metrics between two groups of individuals. However, as paid increasing attention recently [6, 7, 8, 9, 10, 11, 12, 13, 14], unlawful discrimination is a causal connection between the challenged decision and a protected characteristic, which cannot be captured by simple correlation or association concepts. To address this limitation, causal-based fairness measures have been proposed, including total effect [15], direct and indirect discrimination [6, 15, 16], counterfactual fairness [17, 18, 9], and path-specific counterfactual fairness [19]. Fairness notions have also been extended to considering both decisions in the training data and decisions made by predictive models, such as equality of opportunity and equalized odds [31, 32], and counterfactual direct and indirect error rates [118].

In this chapter, we develop a new causal-based fairness notation, called equality of effort. Consider a dataset with N individuals with attributes $(S, T, \mathbf{X}, Y)$ where S denotes a protected attribute such as *gender* with domain values $\{s^+, s^-\}$, Y denotes a decision attribute such as *loan* with domain values $\{y^+, y^-\}$, T denotes a legitimate attribute such

as *credit score*, and $\mathbf{X}$ denotes a set of covariates. For a particular applicant i in the dataset with profile $(S_i = s^-, T_i = t, \mathbf{X}_i = \mathbf{x}, Y_i = y^-)$, she may ask the counterfactual question, how much her credit score she should improve such that the probability of her loan application approval is above a threshold γ (e.g., 80%). Informally speaking, our proposed equality of effort notation addresses her concern on whether her future effort (the increase of her credit score) has no difference from male applicants with similar profile $\mathbf{x}$.

Following Rubin's causal modeling notations, we use $Y_i(t)$ to represent the potential outcome for individual i given a new treatment $T = t$, $\mathbb{E}[Y_i(t)]$ to denote the individual-level expectation of outcome variable. If $\mathbb{E}[Y_i(t)] \geq \gamma$, we say applicant i tends to receive loan approval with at least probability γ. We can then calculate or estimate the minimum value of the treatment variable to achieve γ-level outcome for individual i. If the minimum value of individual i is significantly higher than her counterparts (i.e., males with similar characteristics), discrimination exists in terms of effort discrepancy.

Our fairness notation, equality of effort, is different from existing fairness notions, e.g., statistical disparity, path-specific effects, which mainly focus on the the effect of the sensitive attribute S on the decision attribute Y. Our proposed equality of effort instead focuses on to what extend the treatment variable T should change to make the individual achieve a certain outcome level. This notation addresses the concerns whether the efforts that would need to make to achieve the same outcome level for individuals from the protected group and the efforts from the unprotected group are different. We develop algorithms for determining whether an individual or a group of individuals are discriminated in terms of equality of effort based on three widely used techniques for causal inference, outcome regression, propensity score weighting, and structural causal modeling. We also develop an optimization-based method for removing discriminatory efforts from biased datasets. We

conduct empirical evaluations to compare the equality of effort and existing fairness notions and evaluation results show the effectiveness of our proposed algorithms.

9.2 Fairness Through Equal Effort

For the sake of simplicity, we assume there is only one binary protected attribute, one binary decision attribute, and one ordered multi-categorical legitimate attribute. Our formulation and methods are readily to extend to general cases where there are multiple protected/decision/legitimate attributes. In this chapter, we simply use the change of T as the effort needed to achieve a certain level of outcome and do not consider the real monetary or resource cost behind that change.

9.2.1 Equality of Effort at the Individual Level

For an individual i in the dataset with profile $(s_i, t_i, \mathbf{x}_i, y_i)$, we want to figure out what is the minimal change on treatment variable T to achieve a certain outcome level based on observational data. If the minimal change for individual i has no difference from that of counterparts (individuals with similar profiles except the sensitive attribute), we say individual i achieves fairness in terms of equality of effort.

Formally, we use $Y_i(t)$ to represent the potential outcome for individual i given a new or counterfactual treatment $T = t$. We use $\mathbb{E}[Y_i(t)]$ to denote the individual-level expectation of outcome variable where $\mathbb{E}[\cdot]$ is the expectation operator from probability theory. When $\mathbb{E}[Y_i(t)]$ is larger than a predefined threshold γ, we say individual i would receive a positive decision with probability γ.

Definition 10 (γ-Minimum Effort)**.** For individual i with value $(s_i, t_i, \mathbf{x}_i, y_i)$, the minimum

value of the treatment variable to achieve γ-level outcome is defined as:

$$\Psi_i(\gamma) = \underset{t \in T}{\operatorname{argmin}} \left\{ \mathbb{E}[Y_i(t)] \geq \gamma) \right\}$$

and the minimum effort to achieve γ-level outcome is $\Psi_i(\gamma) - t_i$.

However $Y_i(t)$ cannot be directly observed and we have to derive its estimate from samples with similar characteristics. We design an estimation procedure based on the idea of situation testing [2], which is one normal practice of determining whether an individual is discriminated. How to select variables for finding similar individuals has been studied in situation testing based individual discrimination discovery [119]. The proposed idea there was to first construct a causal graph for all variables and then select variables that are the parents of the decision. Their work is also applicable to our equal effort definition. We first find a subset of users, denoted as I, each of whom has the same (or similar) characteristics ($\mathbf{x}$ and t) as individual i. We denote I^+ (I^-) the subgroup of users in I with the sensitive attribute value s^+ (s^-). Similarly, $\mathbb{E}[Y_{I^+}(t)]$ denotes the expected outcome under treatment t for the subgroup I^+. The minimal effort needed to achieve γ level of outcome variable within the subgroup I^+ is then defined as:

$$\Psi_{I^+}(\gamma) = \underset{t \in T}{\operatorname{argmin}} \{ \mathbb{E}[Y_{I^+}(t)] \geq \gamma \}.$$

Definition 11 (γ-Equal Effort Fairness at the Individual Level)**.** For a certain outcome level γ, we define equality of effort for individual i if

$$\Psi_{I^+}(\gamma) = \Psi_{I^-}(\gamma).$$

The difference $\delta_i(\gamma) = \Psi_{I^+}(\gamma) - \Psi_{I^-}(\gamma)$ measures the effort discrepancy at the individual level.

9.2.2 Equality of Effort at the Group or System Level

In addition to the task of checking individual level discrimination, we also want to check whether discrimination exists at the group or system level. System-level discrimination deals with the average discrimination across the whole system, e.g., all applicants to a university, and group-level discrimination deals with discrimination that occurs in one particular subgroup, e.g., the applicants applying for a particular major. Existing works [3, 6] apply demographic disparity metrics (e.g., risk difference) or causal effect (e.g., direct and indirect causal discrimination) on the whole dataset (the subset of data) to determine the system-level (group-level) discrimination. Similarly, we may want to check whether there are effort discrepancies at the group or system level.

We denote D as the whole dataset, and D^+ (D^-) as the subset with the sensitive attribute value s^+ (s^-). We define the minimum value of treatment variable to achieve a certain outcome level γ for D^* as:

$$\Psi_{D^*}(\gamma) = \underset{t \in T}{\operatorname{argmin}} \left\{ \mathbb{E}[Y_{D^*}(t)] \geq \gamma \right\}.$$

Definition 12 (γ-Equality of Effort at the System Level)**.** For a certain outcome level γ, equality of effort between two sensitive attributes s^+ and s^- is achieved if

$$\Psi_{D^+}(\gamma) = \Psi_{D^-}(\gamma).$$

The difference $\delta_D(\gamma) = \Psi_{D^+}(\gamma) - \Psi_{D^-}(\gamma)$ measures the effort discrepancy at the system

Table 9.1: Formula of previous fairness notions.

Notation	References	Formula
Demographic parity	[120]	$P(y^+\|s^+) - P(y^+\|s^-)$
Conditional parity	[120]	$P(y^+\|s^+, \mathbf{o}) - P(y^+\|s^-, \mathbf{o})$
Total causal discrimination	[6, 15]	$\mathbb{E}[Y(s^+)] - \mathbb{E}[Y(s^-)]$
Path-specific causal discrimination	[6, 8]	$\mathbb{E}[Y(s^+)\|\pi] - \mathbb{E}[Y(s^-)\|\pi]$
Counterfactual fairness	[17]	$\mathbb{E}[Y_{\mathbf{o}}(s^+)] - \mathbb{E}[Y_{\mathbf{o}}(s^-)]$
Path-specific counterfactual fairness	[19]	$\mathbb{E}[Y_{\mathbf{o}}(s^+)\|\pi] - \mathbb{E}[Y_{\mathbf{o}}(s^-)\|\pi]$
Equality of opportunity	[31, 32]	$P(\hat{Y} = y^+\|s^+, y^+) - P(\hat{Y} = y^+\|s^-, y^+)$
Calibration	[31, 32]	$P(y^+\|s^+, \hat{Y} = y^+) - P(y^+\|s^-, \hat{Y} = y^+)$

level.

Definition 12 can be straightforwardly adapted to the group level. Given two compared groups, their distributions in terms of certain attributes (e.g., outstanding debt) could be different. The simple use of our group equal-effort fairness may not be appropriate. In this case, we could apply the path-specific effect/mediator analysis [6, 8] to separate and measure different causal effects e.g., direct discrimination, indirect discrimination, and explainable effects.

9.2.3 Comparison with Other Fairness Metrics

Many different fairness metrics have been proposed to measure fairness of data and machine learning algorithms. Classic metrics include individual fairness, demographic parity, equality of opportunity, calibration, causal fairness, and counterfactual causal fairness. Refer to a recent survey [120]. We show in Table 9.1 the formula of previous representative fairness metrics to compare with our equality of effort notion. For example, demographic imparity requires that $P(y^+|s^+) = P(y^+|s^-)$ and similarly conditional demographic imparity requires $P(y^+|s^+, \mathbf{o}) = P(y^+|s^-, \mathbf{o})$ where $\mathbf{o}$ is the values of a specified variable set $\mathbf{O}$. Basically they require that a decision be independent of the protected attribute conditional or unconditional

on some other variables. For causal based fairness notions, the total causal discrimination is based on the average causal effect of S on Y and is defined as $\mathbb{E}[Y(s^+)] - \mathbb{E}[Y(s^-)]$, which represents the expected change of outcome Y when S of all individuals changes from s^- to s^+. Different from the total causal discrimination that measures the causal effect transmitted along all the causal paths from S to Y in the causal graph, the path-specific causal discrimination is based on the causal effect that is transmitted along some specific paths π from S to Y, e.g., direct causal discrimination when π is the direct path from S to Y, and indirect causal discrimination when π is all paths from S to Y through redlining attribute T. Counterfactual fairness requires $\mathbb{E}[Y_{\mathbf{o}}(s^+)] - \mathbb{E}[Y_{\mathbf{o}}(s^-)]$, which means that a decision is fair towards an individual if it is the same in the actual world and a counterfactual world where the individual belonged to a different demographic group. Most recently, [19] developed a unified definition, path-specific counterfactual fairness (PC Fairness), that covers previous causality-based fairness notations. Different from demographic parity and causal based fairness notions, our proposed equality of effort considers to what extend the legitimate variable T should change to achieve a certain outcome level and whether the minimum effort made for individuals from the protected group and that from the unprotected group are the same.

When considering discrimination from the perspective of supervised learning, the equality of opportunity is based on the actual outcome Y and the predicted outcome $\hat{Y}$, requiring $P(\hat{Y} = y^+|s^+, y^+) = P(\hat{Y} = y^+|s^-, y^+)$. Basically it means the decision model should not mistakenly predict examples with y^+ as $\hat{Y} = y^-$ at a higher rate for one group than another. In other words, a predictor $\hat{Y}$ satisfies equalized opportunity with respect to protected attribute S and outcome Y if $\hat{Y}$ and S are independent conditional on Y. Similarly the calibration considers the fraction of correct positive predictions and requires

$P(y^+|s^+, \hat{Y} = y^+) = P(y^+|s^-, \hat{Y} = y^+)$. Different from the previous methods that focuses on prediction results, our proposed equality of opportunity focuses on the effort, i.e., the minimum change of T to achieve a certain outcome level Y, based on the causal framework.

We noticed a parallel work [121] that developed an effort-based measure of fairness and formulated effort unfairness as the inequality in the amount of effort required for members from disadvantage group and advantaged group. However, their work focused on characterizing the long-term impact of algorithmic policies on reshaping the underlying population based on the psychological literature on social learning and the economic literature on equality of opportunity. Our work is based on counterfactual causal inference and develops an optimization-based framework for removing discriminatory effort unfairness from the static data if discrimination is detected.

9.3 Calculating Average Effort Discrepancy

In real-world applications, we often have multiple values of γ used in decision making. We use the average effort discrepancy over all values of γ as the measure of equality of effort in this scenario. If γ has a set of discrete values, then the average is computed by the mean of all effort discrepancies. If γ is a continuous variable, then the average is defined as the integration over the range of γ.

Definition 13 (Average Effort Discrepancy (AED))**.** If $\gamma \in \Gamma$ where Γ denotes the effort level value set of the expectation of outcome variable, then the average effort discrepancy is defined as

$$AED = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \delta(\gamma). \tag{9.1}$$

If γ is a continuous variable in a range $[\gamma_1, \gamma_2]$, then the average effort discrepancy is defined

as

$$AED = \frac{1}{\gamma_2 - \gamma_1} \int_{\gamma_1}^{\gamma_2} \delta(\gamma) d\gamma. \tag{9.2}$$

To calculate the AED, we need to first compute the expected outcome $\mathbb{E}[Y_{I^*}(t)]$ or $\mathbb{E}[Y_{D^*}(t)]$, and then compute the minimum effort. In the following, we develop a general calculating method assuming the monotonicity and invertibility for $\mathbb{E}[Y_{D^*}(t)]$. Then, we consider three widely used techniques for causal inference: outcome regression and propensity score weighting from Rubin's framework, and structural causal analysis from Pearl's framework. We compute the AED for each of the techniques.

Algorithm 13 Discrimination detection through equal effort.

Ensure: Discrimination detection result

For each subset $D^* \in \{D^+, D^-\}$, identify expected outcome $f_{D^*}(t) = \mathbb{E}[Y_{D^*}(t)]$
if $f_{D^*}(t)$ is continuous, monotonous and invertible **then**
    Calculate AED according to Eq. (9.3)
else
    Identify inverse function $f^{-1}_{D^*}(\gamma)$
    if $f^{-1}_{D^*}(\gamma)$ has a closed form **then**
        for each γ **do**
            Find the minimum value of t such that $t \geq f^{-1}_{D^*}(\gamma)$
            Calculate effort discrepancy $\delta_D(\gamma)$
        end for
    else
        for each treatment level t **do**
            Use appropriate causal inference method to estimate z
        end for
        for each γ **do**
            Numerically find the minimum value of t such that $\hat{\mathbb{E}}[Y_{D^*}(t)] \geq \gamma$
            Calculate effort discrepancy $\delta_D(\gamma)$
        end for
        Calculate AED following Definition 13
    end if
end if
if $|AED| \geq \tau$ **then**
    Result = True
else
    Result = False
end if

Algorithm 13 shows the pseudocode of our algorithm for computing the AED and making the judge of discrimination through equal effort. Lines 2-3 deal with the situation where $f_{D^*}(t) = \mathbb{E}[Y_{D^*}(t)]$ is a continuous, monotonous and invertible function of t, and AED can be directly computed through an integration over $f_{D^*}(t)$ given in the next subsection. If the assumptions are not satisfied, lines 6-10 handle the situation where the closed-form of inverse function $f^{-1}_{D^*}(\gamma)$ can be derived; and lines 12-19 handle the situation otherwise.

9.3.1 General Method under Monotonicity and Invertibility Assumption

As discussed in the previous section, $\mathbb{E}[Y_{D^+}(t)]$ and $\mathbb{E}[Y_{D^-}(t)]$ denote the expectations of outcome variable for groups D^+ and D^-. We can treat them as functions of t, denoted as $f_{D^+}(t)$ and $f_{D^-}(t)$. Under the assumptions of being monotonically increasing and invertible, inequality $\mathbb{E}[Y_{D^+}(t)] \geq \gamma$ can be expressed as $f_{D^+}(t) \geq \gamma$, which leads to $t \geq f^{-1}_{D^+}(\gamma)$, where $f^{-1}_{D^+}(\cdot)$ is the inverse function of $f_{D^+}(\cdot)$. As a result, we directly obtain that $\Psi_{D^+}(\gamma) = f^{-1}_{D^+}(\gamma)$, and similarly $\Psi_{D^-}(\gamma) = f^{-1}_{D^-}(\gamma)$.

If the closed forms of $f^{-1}_{D^+}(\cdot)$ and $f^{-1}_{D^-}(\cdot)$ can be derived, then the AED can be easily computed; otherwise its calculation is not straightforward. However, when γ is a continuous variable, then we don't need to derive the closed form of the inverse functions to compute the AED, but only require the integration of $f_{D^+}(\cdot)$ and $f_{D^-}(\cdot)$ to be tractable. This is because based on the Laisant's theorem we have

$$\int_{\gamma_1}^{\gamma_2} f^{-1}_{D^+}(\gamma)d\gamma = \gamma_2 t_2^+ - \gamma_1 t_1^+ - \int_{t_1^+}^{t_2^+} f_{D^+}(\gamma)d\gamma,$$

where $t_1^+ = f^{-1}_{D^+}(\gamma_1)$ and $t_2^+ = f^{-1}_{D^+}(\gamma_2)$. In practice, t_1^+ and t_2^+ can be estimated using

numerical methods. As a result, the AED is given by

$$(t_2^+ - t_2^-)\gamma_2 - (t_1^+ - t_1^-)\gamma_1 - \Big(\int_{t_1^+}^{t_2^+} f_{D^+}(\gamma)d\gamma - \int_{t_1^-}^{t_2^-} f_{D^-}(\gamma)d\gamma\Big). \tag{9.3}$$

9.3.2 Outcome Regression

Outcome regression is one straightforward method to conduct causal inference. In this approach, a model is posited for the outcome variable as a function of the treatment variable and the covariates. The basic outcome regression model is the linear regression of the form:

$$\mathbb{E}[Y|T, \mathbf{X}] = \beta_0 + \beta_1 T + \boldsymbol{\beta_2}\boldsymbol{X} + \boldsymbol{\beta_3}\boldsymbol{X}T,$$

where β_0, β_1 are regression coefficients, $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ are the coefficient vectors with the same length as $\boldsymbol{X}$. All the parameters can be estimated by least squares method.

One advantage of outcome regression is it can help us directly calculate the relative treatment value given a certain expected outcome level. Suppose the regression model is correctly specified, the expected outcome of any subset D^* is given by

$$\mathbb{E}[Y_{D^*}(t)] = \frac{1}{|D^*|}\sum_{i\in D^*}(\beta_0 + \beta_1 t + \boldsymbol{\beta_2}\boldsymbol{x}_i + \boldsymbol{\beta_3}\boldsymbol{x}_i t).$$

Thus, the minimum value of the treatment variable to achieve γ-level outcome, i.e., $\Psi_{D^*}(\gamma)$, can be expressed as:

$$\underset{t\in T}{\operatorname{argmin}}\left\{\mathbb{E}[Y_{D^*}(t)] \geq \gamma\right) = \frac{\gamma - \frac{1}{|D^*|}\sum_{i\in D^*}(\beta_0 + \boldsymbol{\beta_2})}{\frac{1}{|D^*|}\sum_{i\in D^*}(\beta_1 + \boldsymbol{\beta_3})}. \tag{9.4}$$

9.3.3 Propensity Score Weighting

Another widely used branch of causal inference is based on weighting and one typical method is the inverse propensity score weighting. In our context, the treatment variable is a multiple valued ordinal variable, we apply generalized propensity score [122] to estimate the weights.

Definition 14 (Generalized Propensity Score)**.** The generalized propensity score for individual i is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables:

$$r(t, \mathbf{x_i}) = Pr(T = t|\mathbf{X_i} = \mathbf{x_i}).$$

The weighted mean of the potential outcomes for those who received the treatment t had they received another treatment t' can be consistently estimated by

$$\hat{\mathbb{E}}[Y(t')|t] = \frac{\sum_{i \in N} \mathbb{1}_{T_i = t'} Y_i \omega_i(t, t')}{\sum_{i \in N} \mathbb{1}_{T_i = t'} \omega_i(t, t')},$$

where

$$\omega_i(t, t') = \frac{r(t, \mathbf{x_i})}{r(t', \mathbf{x_i})}.$$

Following the above method, we can get a table showing estimation values of the expected outcome under all treatment pair combinations (t, t'). Thus, the minimum treatment value to achieve $\hat{\mathbb{E}}[Y(t')|t] \geq \gamma$ can be determined by comparing the results in that table.

9.3.4 Structural Causal Model

The structural causal model describes the causal mechanisms of a system as a set of structural equations. For ease of representation, each causal model can be illustrated by a directed acyclic graph called the causal graph, where each node represents a variable and each edge represents the direct causal relationship specified by the causal model. In addition, each node V is associated with a conditional probability distribution $P(v|\mathbf{pa}V)$ where $\mathbf{pa}V$ is the realization of a set of variables $\mathbf{Pa}V$ called the parents of V. The treatment is modeled using the intervention, which forces the treatment variable T to take certain value t, formally denoted by $do(T = t)$ or $do(t)$. The potential outcome of variable Y under intervention $do(t)$ is denoted as Y_t. The distribution of Y_t, also referred to as the post-intervention distribution of Y under $do(t)$, is denoted as $P(Y_t)$. Facilitated by the intervention, the expected outcome $\mathbb{E}[Y_{D^*}(t)]$ can be measured by the counterfactual quantity $\mathbb{E}[Y_t|\mathbf{z}^*]$, where $\mathbf{z}^*$ represents attribute values that form the subgroup D^*. The counterfactual quantity measures the expected outcome of Y assuming that the intervention is performed on the subgroup of individuals only. According to [21], if attributes $\mathbf{Z}$ are non-descendant of T in the causal graph, then $P(Y_t|\mathbf{z}^*)$ can be computed from observational data as

$$\frac{\sum_{\mathbf{X}\setminus\mathbf{Z}} \prod_{V\in\{Y,S,\mathbf{X}\}} P(v|\mathbf{pa}V)_{\delta_{T=t}}}{P(\mathbf{z}^*)},$$

where $\delta_{T=t}$ means assigning T involved in all probabilities with the corresponding value t.

If the inverse function of $\mathbb{E}[Y_t|\mathbf{z}^*]$ can be derived, then we follow lines 6-10 in Algorithm 13 to compute AED; otherwise, we follow lines 12-19 to compute AED.

Table 9.2: Preprocessing *education*.

Category	Original Values
0	Preschool, 1st-4th, 5th-6th
1	7th-8th, 9th, 10th, 11th
2	12th, HS-grad, Some-college, Assoc-voc
3	Assoc-acdm, Bachelors, Masters, Prof-school
4	Doctorate

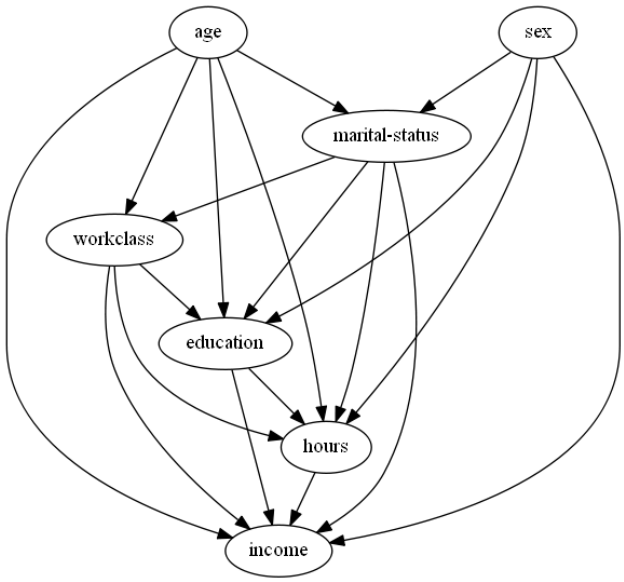


Figure 9.1: Constructed causal graph for Adult Dataset.

Table 9.3: Expectation of the potential outcome for males and females in Adult dataset.

education	*sex=male*			*sex=female*		
	Weighting	*Regression*	*SCM*	*Weighting*	*Regression*	*SCM*
0	0.196	0.086	0.164	0.048	0.026	0.057
1	0.269	0.214	0.239	0.066	0.051	0.075
2	0.513	0.491	0.498	0.211	0.190	0.221
3	0.736	0.781	0.741	0.416	0.497	0.469
4	0.842	0.933	0.859	0.485	0.807	0.706

9.4 Achieving Equal Effort

When our discrimination detection algorithm shows that a dataset does not satisfy the equal effort requirement, then we may want to remove the discriminatory effects from the dataset before it is used for any predictive analysis, i.e., training a decision model. In this section, we develop a method for generating a new dataset which is close to the original dataset and also satisfies equal effort. Our removal method is based on the use of outcome regression to estimate the potential outcome, but it can be easily extended to any method where the closed form of $\Psi(\gamma)$ can be derived. The general idea is to derive a new outcome regression model satisfying the equal effort constraints. Then, for each individual in the original dataset, we randomly generate a new value $\tilde{Y}$ based on the expectation computed from the fair outcome regression model.

Specifically, we consider two outcome regression models for subsets D^+ and D^- respectively, given by

$$\mathbb{E}[Y_{D^+}|T, \mathbf{X}] = \beta_0^+ + \beta_1^+ T + \boldsymbol{\beta}_2^+ \boldsymbol{X} + \boldsymbol{\beta}_3^+ \boldsymbol{X} T,$$

$$\mathbb{E}[Y_{D^-}|T, \mathbf{X}] = \beta_0^- + \beta_1^- T + \boldsymbol{\beta}_2^- \boldsymbol{X} + \boldsymbol{\beta}_3^- \boldsymbol{X} T.$$

Then, as shown by Eq. (9.4), the minimum effort for subgroup D^+ (and similarly for subgroup D^-) is given by

$$\Psi_{D^+}(\gamma) = \frac{\gamma - \frac{1}{|D^+|}\sum_{i \in D^+}(\beta_0^+ + \boldsymbol{\beta}_2^+)}{\frac{1}{|D^+|}\sum_{i \in D^+}(\beta_1^+ + \boldsymbol{\beta}_3^+)}.$$

Table 9.4: Expectation of the potential outcome for males and females with the original *education*=0.

education	*sex=male*			*sex=female*		
	Weighting	*Regression*	*SCM*	*Weighting*	*Regression*	*SCM*
1	0.225	0.232	0.227	0.071	0.084	0.081
2	0.457	0.462	0.467	0.205	0.205	0.224
3	0.692	0.694	0.719	0.418	0.411	0.497
4	0.810	0.870	0.842	0.497	0.693	0.754

As a result, the AED according to either Eq. (9.1) or (9.2) is given by

$$\frac{\bar{\gamma} - \frac{1}{|D^+|}\sum_{i\in D^+}(\beta_0^+ + \boldsymbol{\beta}_2^+)}{\frac{1}{|D^+|}\sum_{i\in D^+}(\beta_1^+ + \boldsymbol{\beta}_3^+)} - \frac{\bar{\gamma} - \frac{1}{|D^-|}\sum_{i\in D^-}(\beta_0^- + \boldsymbol{\beta}_2^-)}{\frac{1}{|D^-|}\sum_{i\in D^-}(\beta_1^- + \boldsymbol{\beta}_3^-)},$$

where $\bar{\gamma}$ equals $\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\gamma$ if discrete and $\frac{\gamma_2^2-\gamma_1^2}{2}$ if continuous. We want the AED to approach zero. After adding the penalty term for the AED, the objective function becomes

$$\underset{\beta}{\operatorname{argmin}} \sum_{i\in D^+, D^-} \left(\mathbb{E}[Y_{D^*}|t_i, \mathbf{x}_i] - y_i\right)^2 + \lambda \cdot AED^2$$

where $D^* = D^+$ or D^- and λ is the parameter for balancing the two objectives.

Finally, for each individual i in the dataset with profile $(s_i, t_i, \mathbf{x}_i, y_i)$, we first compute his expected value of Y using the fair outcome regression model, i.e., $\mathbb{E}[Y_{D^*}|t_i, \mathbf{x}_i]$, where $D^* = D^+$ or D^- depending on the value of s_i. Then, we randomly assign 0 or 1 to the new value $\tilde{y}_i$ based on the probability given by $\mathbb{E}[Y_{D^*}|t_i, \mathbf{x}_i]$. The generated data then satisfies the equal effort requirement.

9.5 Experiments

We evaluate our discrimination detection and removal algorithms based on the proposed equality of effort on the UCI Adult dataset [123]. The Adult dataset contains 65,123

Table 9.5: Expectation of the potential outcome for three randomly chosen individuals.

education	*User 1*		*User 2*		*User 3*	
	sex=male	*sex=female*	*sex=male*	*sex=female*	*sex=male*	*sex=female*
0	-	-	-	-	0.012	0.006
1	0.022	0.007	0.058	0.030	0.051	0.024
2	0.085	0.036	0.206	0.134	0.188	0.096
3	0.282	0.159	0.523	0.438	0.501	0.317
4	0.624	0.487	0.823	0.796	0.813	0.669

records with 14 attributes. We select 7 attributes, *sex*, *age*, *marital status*, *workclass*, *education*, *hours*, and *income* in our experiments. We consider *income* as the outcome, *education* as the treatment attribute, and *sex* as the protected attribute. Due to the sparse data issue, we binarize the domain of *age*, *marital status*, *workclass*, and *hours* into two classes. We also categorize 16 values of *education* into five levels, as shown in Table 9.2.

In our experiments, we calculate the minimum effort based on three methods, outcome regression (*Regression*), propensity score weighting (*Weighting*), and structural causal model inference (*SCM*). For *Weighting*, we implement the propensity score weighting for multiple treatments by following the work of [124] and [76]. For *SCM*, we follow the settings of [6] and use three tiers for causal graph learning: *sex*, *age* in Tier 1, *marital-status*, *education*, *workclass*, and *hours* in Tier 2, and *income* in Tier 3. The causal graph is constructed and presented by utilizing the open-source software TETRAD [105]. We employ the original PC algorithm [78] and set the significance threshold 0.01 for conditional independence setting in causal graph construction. Figure 9.1 shows the built causal graph. We apply the nonparametric inference of the structural causal model by following the work of [125]. In discrimination removal, the quadratic programming is solved using PyTorch [126].

9.5.1 Discrimination Discovery

9.5.1.1 Checking equal effort at the system level

Table 9.3 shows the comparison results of the expectations of the potential outcome for males ($\mathbb{E}[Y_{D^+}(t)]$) and that for females ($\mathbb{E}[Y_{D^-}(t)]$) in Adult. We calculate the expectation of the potential outcomes using three methods, *Weighting*, *Regression*, and *SCM*, and vary the treatment variable *education* from 0 to 4. As shown in Table 9.3, the expectations of potential outcome for males are significantly higher than the corresponding values for females, indicating large effort discrepancy exists in Adult. For example, $\mathbb{E}[Y_{D^+}(t)] = 0.498$ and $\mathbb{E}[Y_{D^-}(t)] = 0.221$ when $t = 2$ based on *SCM*. If we set $\gamma = 0.7$, the minimum values of treatment variable (*education*) to achieve γ-level outcome are 3 for males (with the expectation of the potential outcome 0.741) and 4 for females (with the expectation of the potential outcome 0.706). The effort discrepancy between females and males is 1, which indicates the existence of significant discrimination in terms of equal effort fairness. We would like to point out that the expectations of potential outcome calculated from three methods are generally consistent as shown in Table 9.3. However, each calculation method has its own applicable assumptions and may not achieve reliable results when those assumptions are not met. There are extensive researches on the applicability of those causal inference methods (e.g., refer to [21]), which are out of the scope of this work.

9.5.1.2 Checking equal effort at the group level

For the group level equality of effort, we split the Adult dataset into five groups by *education*. Individuals with the same education value form one group. For each group, we calculate the expectations of potential outcome for males ($\mathbb{E}[Y_{D^+}(t)]$) and females ($\mathbb{E}[Y_{D^-}(t)]$).

We report in Table 9.4 the expectations of the potential outcome variable for group one with *education=0*. Each expectation is calculated using three methods. We can see the significant discrepancy between males and females in this group. We also observe the similar phenomena in other four groups. When considering $\gamma = 0.5$, the minimum education value to achieve the outcome for males in this group is 3 (with all expectation values from three methods close to 0.7) whereas the minimum education level for females is 4.

9.5.1.3 Checking equal effort at the individual level

To detect effort discrepancy at the individual level, we need to first identify a subset of users I with the same characteristics of the given individual and then split them into the male group (I^+) and female group (I^-). We then calculate the expectations of potential outcome for the male group ($\mathbb{E}[Y_{I^+}(t)]$) and female group ($\mathbb{E}[Y_{I^+}(t)]$) with each treatment level t. We report in Table 9.5 the results of three randomly chosen female users whose index numbers are 425, 9569, and 46437. Both users 1 and 2 have the original education value 1 and user 3 has education value 0. As shown in Table 9.5, the expectations of outcome for I^+ are consistently higher than I^-, indicating the existence of discrimination in terms of equal effort for these three individuals. For example, results of user 3 show that the minimum effort for her to achieve 0.5-level outcome is education $t = 4$ whereas the corresponding minimum effort to achieve the same level outcome is $t = 3$ had she been a male.

9.5.2 Discrimination Removal

We run our removal algorithm to remove discrimination in terms of equality of effort from the Adult dataset, and then run the discovery algorithm to further examine whether discrimination is truly removed in the modified dataset. For comparison, we include the re-

moval algorithm (Denoted by DI) of [5], which removes discrimination from the demographic parity perspective. Basically, DI tries to modify X such that the modified $\hat{X}$ cannot be used to predict S. The results show that, after executing our removal method (with $\lambda = 5$), the average difference between $\mathbb{E}[Y_{D^+}(t)]$ and $\mathbb{E}[Y_{D^-}(t)]$ for all ts is -0.0136, indicating all effort discrepancy has been removed. However, the average difference for the DI algorithm is 0.2628, showing that DI does not remove effort discrepancy. Regarding data utility loss in terms of χ^2, our method also outperforms the DI algorithm in that the utility loss of our method is 34778, while the utility loss of the DI algorithm is 37997.

9.6 Summary

In this chapter, we proposed a new causality-based fairness notion called the equality of effort. Although previous notions can be used to judge discrimination from various perspectives (e.g., demographic parity, equal opportunity), they cannot quantify the (difference in) efforts that individuals need to make in order to achieve certain outcome levels. Our proposed notion, on the other hand, can help answer counterfactual questions like "how much credit score an applicant should improve such that the probability of her loan application approval is above a threshold", and judge discrimination from the equal-effort perspective. To quantify the average effort discrepancy, we developed a general method under certain assumptions and specific methods based on three common causal inference techniques. When equality of effort is not achieved in a dataset, we developed an optimization method to remove discrimination. In the experiments, we show that the Adult dataset does contain effort discrepancy at system, group, and also individual levels, and our removal method can ensure the newly generated dataset satisfies equality of effort. The early version of this work is published at WWW Workshop: FATES 2020 [20].

10 Conclusion and Future Work

In this chapter, we summarize our works and, based on the observed results and performance analysis, propose several potential research directions associated with causal fairness in recommendation.

10.1 Conclusion

Around achieving causal fairness in recommendation, this dissertation aims to explore and address the following problems:

1. How to achieve user-side group level fairness in bandit-based recommendation;
2. How to achieve user-side counterfactual fairness at individual level in bandit-based recommendation;
3. How to deal with biases from various sources simultaneously in recommender systems;
4. How to robustly improve bandit-based recommendation algorithms by leveraging offline data under compound biases;
5. How to achieve causal fairness under hidden confounding with the benefit of multiple treatment/redlining variables;
6. How to discover discrimination and achieve causal fairness in terms of equality of efforts.

In Chapter 4, we study how to achieve user-side fairness in personalized recommendation. We formulate our fair personalized recommendation as a modified contextual bandit

and focus on achieving fairness on the individual whom is being recommended an item as opposed to achieving fairness on the items that are being recommended. We introduce and define a metric that captures the fairness in terms of rewards received for both the privileged and protected groups. We develop a fair contextual bandit algorithm, Fair-LinUCB, that improves upon the traditional LinUCB algorithm to achieve group-level fairness of users. Our algorithm detects and monitors unfairness while it learns to recommend personalized videos to students to achieve high efficiency. We provide a theoretical regret analysis and show that our algorithm has a slightly higher regret bound than LinUCB. We conduct numerous experimental evaluations to compare the performances of our fair contextual bandit to that of LinUCB and show that our approach achieves group-level fairness while maintaining a high utility.

In Chapter 5, we study how to recommend an item at each step to maximize the expected reward while achieving user-side fairness for customers, i.e., customers who share similar profiles will receive a similar reward regardless of their sensitive attributes and items being recommended. By incorporating causal inference into bandits and adopting soft intervention to model the arm selection strategy, we first propose the d-separation based UCB algorithm (D-UCB) to explore the utilization of the d-separation set in reducing the amount of exploration needed to achieve low cumulative regret. Based on that, we then propose the fair causal bandit (F-UCB) for achieving the counterfactual individual fairness. Both theoretical analysis and empirical evaluation demonstrate effectiveness of our algorithms.

In Chapter 6, we formulate the causal personalized recommendation problem based on the structural causal model (SCM) and a generalization of the notion of backdoor adjustment to account for both biases. Our approach leverages external data of some variables that are also measured without selection bias and uses an adjustment pair based on the de-

rived graphical conditions for identifying conditional causal effects. We present a statistical estimation procedure based on inverse probability weighting to calculate conditional causal effects when training samples are limited. Under the presence of confounding and selection biases, we also show how to derive path-specific effects and counterfactual effects, both of which are needed in recommendation analysis. We show the effectiveness of our approach through experimental analysis.

In Chapter 7, we investigate bounding conditional causal effects in the presence of confounding and sample selection biases using causal inference techniques and utilizes the derived bounds to robustly improve online bandit algorithms. We present two novel causal-based techniques to derive a bound for conditional causal effects given offline data with compound biases. We develop contextual and non-contextual bandit algorithms that leverage the derived causal bounds and conduct their regret analysis. Theoretical analysis and empirical evaluation demonstrate the improved regrets of our algorithms.

In Chapter 8, we focus on discrimination discovery when multiple protected attributes and redlining attributes are present in addition to other covariates. We regard those protected and redlining attributes as multiple causes of the outcome variable. To deal with unobserved variables, especially hidden confounders, we adopt the potential outcome framework and leverage the state-of-the-art *deconfounder* algorithm to do causal inference under multiple causes. The deconfounder algorithm infers a latent variable as a substitute for unobserved confounders and then uses that substitute to perform causal inference. Our approach is more appropriate for discrimination discovery as it is able to relax the Markovian assumption and avoid the unidentifiability issue in structural causal modeling approaches. We conduct empirical evaluation on both synthetic data and real data. Empirical evaluation results demonstrate the effectiveness of our proposed approach.

In Chapter 9, we develop a new causal-based fairness notation, called equality of effort. Different from existing fairness notions which mainly focus on discovering the disparity of decisions between two groups of individuals, the proposed equality of effort notation helps answer questions like to what extend a legitimate variable should change to make a particular individual achieve a certain outcome level and addresses the concerns whether the efforts made to achieve the same outcome level for individuals from the protected group and that from the unprotected group are different. We develop algorithms for determining whether an individual or a group of individuals is discriminated in terms of equality of effort. We also develop an optimization-based method for removing discriminatory effects from the data if discrimination is detected. We conduct empirical evaluations to compare the equality of effort and existing fairness notion and show the effectiveness of our proposed algorithms.

10.2 Future Work

In this section, based on the works we have done, we propose several interesting directions that deserve further exploration and investigation.

Generally speaking, in this dissertation we studied the problems of achieving causal fairness in recommendation. In future work, we will continue along this general direction and explore new challenging research problems. First, in causal inference literature we usually categorize compound biases into confounding bias and selection bias due to the orthogonality of these two kinds of biases. In recommendation research field, how to disentangle biases from different sources based on the causal graph structure and causal representation learning techniques requires further exploration. Second, further research should be undertaken to investigate how to combine the abstract causal graph with features that have semantic meanings to achieve better performance in recommendation. More broadly, research is also

needed to build a general causal analysis framework for recommendation. It is also imperative to unify current causal based prediction and debiasing methods under the scope of structure causal model framework.

There are also potential research problems related to specific chapters. For example, in Chapter 4 we made a linear assumption on the reward function. In the future work, we plan to extend the user-level fairness to more general cases and make it easier to be implemented in multifarious reward functions. We plan to develop heuristics to determine the appropriate value for the fairness-accuracy trade off parameter γ. We also plan to study user-side fairness in the multiple choice linear bandits, e.g., recommending multiple videos to a student at each round. Finally, we plan to study how to achieve individual fairness in bandits algorithms. In Chapter 5 we assume the causal graph is a faithful representation of the ground truth causal mechanism and could be learned from logged data. In real world it is usually hard to obtain the accurate causal graph. How to derive causal bandit algorithms with unknown causal structure is still an interesting yet challenging problem. In Chapter 6, in future work, we will conduct empirical comparisons of our debiased recommendation algorithm with existing causal recommendation methods based on user/item embeddings and abstract causal graphs. In Chapter 7, in future work, we will study incorporating causal bounds into advanced bandit algorithms such as contextual bandits under non-linearity assumption and bandits with adversarial feedback. In Chapter 8, the multi-cause discrimination analysis framework is proposed based on the potential outcome framework. How to derive the equivalent or similar graphical assumptions and analysis framework based on structural causal models could be a potential direction in our future exploration. In Chapter 9, We made several assumptions including the no-hidden-confounder assumption, monotonicity of the expectation of outcome variable, and invertibility of outcome function. We also assumed

one binary protected attribute and one binary decision for simplicity's sake. The no-hidden-confounder assumption is a common assumption for causal inference [21] and widely adopted by causal inference based fair learning. The monotonicity assumption reflects the real world phenomena (the more effort, the better outcome). The invertibility assumption is used in our general method of calculating the average effort discrepancy without deriving the closed form of the inverse function. When this invertibility assumption is not held, we have presented in our algorithm several inference methods that could also have their limitations. Moreover, we implicitly assumed that the discrimination detection algorithm knows the same information as the decision-maker, i.e., there are no omitted variables used in decision making but invisible to the discrimination detection. In our future work, we will study how to achieve equal effort fairness when some of those assumptions are not met in practice.

Bibliography

[1] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 05, pp. 582–638, Nov. 2014.

[2] B. T. Luong, S. Ruggieri, and F. Turini, "k-nn as an implementation of situation testing for discrimination discovery and prevention," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 502–510.

[3] I. Žliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 992–1001.

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. New York, New York, USA: ACM Press, 2012, pp. 214–226.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and Removing Disparate Impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. ACM Press, 2015, pp. 259–268.

[6] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017, pp. 3929–3935.

[7] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding Discrimination through Causal Reasoning," *Neural Information Processing Systems*, Jun. 2017.

[8] R. Nabi and I. Shpitser, "Fair Inference On Outcomes," *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 1931–1940, 2018.

[9] Y. Wu, L. Zhang, and X. Wu, "Counterfactual fairness: Unidentification, bound and algorithm," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019, pp. 1438–1444.

[10] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality," p. 7, 2019.

[11] J. Li, J. Liu, L. Liu, T. D. Le, S. Ma, and Y. Han, "Discrimination detection by causal effect estimation," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2017, pp. 1087–1094.

[12] L. Zhang, Y. Wu, and X. Wu, "On discrimination discovery using causal networks," in *Proceedings of SBP-BRiMS 2016*, 2016.

[13] L. Zhang and X. Wu, "Anti-discrimination learning: A causal modeling-based framework," *International Journal of Data Science and Analytics*, vol. 4, no. 1, pp. 1–16, Aug. 2017.

[14] L. Zhang, Y. Wu, and X. Wu, "Causal modeling-based discrimination discovery and removal: criteria, bounds, and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[15] J. Zhang and E. Bareinboim, "Fairness in Decision-Making – The Causal Explanation Formula," *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[16] S. Chiappa, "Path-specific counterfactual fairness," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 7801–7808.

[17] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," *Neural Information Processing Systems*, 2017.

[18] C. Russell, M. J. Kusner, J. Loftus, and R. Silva, "When worlds collide: integrating different counterfactual assumptions in fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 6414–6423.

[19] Y. Wu, L. Zhang, X. Wu, and H. Tong, "PC-Fairness: A Unified Framework for Measuring Causality-based Fairness," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, December 8-14, 2019, Vancouver, Canada, 2019*. Curran Associates, Inc., Dec. 2019, pp. 3399–3409.

[20] W. Huang, Y. Wu, L. Zhang, and X. Wu, "Fairness through equality of effort," in *Companion of The 2020 Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*. ACM / IW3C2, 2020, pp. 743–751.

[21] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.

[22] D. Bouneffouf, I. Rish, and C. C. Aggarwal, "Survey on applications of multi-armed and contextual bandits," in *IEEE Congress on Evolutionary Computation, CEC 2020, Glasgow, United Kingdom, July 19-24, 2020*. IEEE, 2020, pp. 1–8. [Online]. Available: https://doi.org/10.1109/CEC48606.2020.9185782

[23] R. Epstein and R. E. Robertson, "The search engine manipulation effect (seme) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, pp. E4512–E4521, 2015.

[24] A. Farahat and M. C. Bailey, "How effective is targeted advertising?" in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 111–120.

[25] L. E. Celis, S. Kapoor, F. Salehi, and N. K. Vishnoi, "An algorithmic framework to control bias in bandit-based personalization," *arXiv preprint arXiv:1802.08674*, 2018.

[26] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes, "Calibrated fairness in bandits," *arXiv preprint arXiv:1707.01875*, 2017.

[27] Z. Zhu, X. Hu, and J. Caverlee, "Fairness-aware tensor-based recommendation," in *CIKM'18*, 2018, pp. 1153–1162.

[28] Y. Zhang, F. Feng, X. He, T. Wei, C. Song, G. Ling, and Y. Zhang, "Causal intervention for leveraging popularity bias in recommendation," *arXiv preprint arXiv:2105.06067*, 2021.

[29] W. Wang, F. Feng, X. He, X. Wang, and T.-S. Chua, "Deconfounded recommendation for alleviating bias amplification," *arXiv preprint arXiv:2105.10648*, 2021.

[30] Z. Zhu, Y. He, X. Zhao, and J. Caverlee, "Popularity bias in dynamic recommendation," 2021.

[31] M. Hardt, E. Price, None, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016, pp. 3315–3323.

[32] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact," in *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*. New York, New York, USA: ACM Press, 2017, pp. 1171–1180, code: https://github.com/mbilalzafar/fair-classification.

[33] A. N. Carey and X. Wu, "The statistical fairness field guide: perspectives from social and formal sciences," *AI and Ethics*, vol. 3, no. 1, pp. 1–23, 2023.

[34] ——, "The causal fairness field guide: Perspectives from social and formal sciences," *Frontiers in Big Data*, vol. 5, 2022.

[35] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, "Delayed impact of fair machine learning," in *ICML'18*, 2018.

[36] A. Bower, S. N. Kitchen, L. Niss, M. J. Strauss, A. Vargas, and S. Venkatasubramanian, "Fair pipelines," *CoRR*, vol. abs/1707.00391, 2017.

[37] V. Emelianov, G. Arvanitakis, N. Gast, K. P. Gummadi, and P. Loiseau, "The price of local fairness in multistage selection," in *IJCAI'19*, 2019, pp. 5836–5842.

[38] C. Dwork and C. Ilvento, "Fairness under composition," in *ITCS'19*, ser. LIPIcs, vol. 124, 2019, pp. 33:1–33:20.

[39] C. Dwork, C. Ilvento, and M. Jagadeesan, "Individual fairness in pipelines," in *FORC'20*, 2020, pp. 7:1–7:22.

[40] M. Joseph, M. J. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," in *NeurIPS*, 2016.

[41] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *arXiv preprint arXiv:2010.03240*, 2020.

[42] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*. ACM, 2017, pp. 781–789.

[43] Y. Zheng, C. Gao, X. Li, X. He, Y. Li, and D. Jin, "Disentangling user interest and conformity for recommendation with causal embedding," in *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*. ACM / IW3C2, 2021, pp. 2980–2991.

[44] E. Bareinboim, J. Tian, and J. Pearl, "Recovering from selection bias in causal and statistical inference," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[45] J. Correa and E. Bareinboim, "Causal effect identification by adjustment under confounding and selection biases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[46] J. Correa, J. Tian, and E. Bareinboim, "Generalized adjustment under confounding and selection biases," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[47] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on applications of multi-armed and contextual bandits," in *2020 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2020, pp. 1–8.

[48] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proceedings of the 20th International Conference on Neural Information Processing Systems*. Citeseer, 2007, pp. 817–824.

[49] M. N. Katehakis and A. F. Veinott Jr, "The multi-armed bandit problem: decomposition and computation," *Mathematics of Operations Research*, vol. 12, no. 2, pp. 262–268, 1987.

[50] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems*, 2011, pp. 2312–2320.

[51] G. Ghalme, S. Jain, S. Gujar, and Y. Narahari, "Thompson sampling based mechanisms for stochastic multi-armed bandit problems," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*. ACM, 2017, pp. 87–95.

[52] V. Syrgkanis, A. Krishnamurthy, and R. E. Schapire, "Efficient algorithms for adversarial contextual learning," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, ser. JMLR Workshop and Conference Proceedings, vol. 48. JMLR.org, 2016, pp. 2159–2168.

[53] Y. Gur, A. J. Zeevi, and O. Besbes, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014, pp. 199–207.

[54] M. Joseph, M. J. Kearns, J. Morgenstern, S. Neel, and A. Roth, "Meritocratic fairness for infinite and contextual bandits," in *AIES'18*. ACM, 2018, pp. 158–163.

[55] S. Jabbari, M. Joseph, M. J. Kearns, J. Morgenstern, and A. Roth, "Fairness in reinforcement learning," in *ICML'17*, 2017.

[56] R. Burke, "Multisided Fairness for Recommendation," *FAT ML Workshop*, vol. 5, no. 17, Jul. 2017.

[57] R. Burke, N. Sonboli, and A. Ordoñez-Gauger, "Balanced Neighborhoods for Multisided Fairness in Recommendation," in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, vol. 81, 2018, pp. 202–214.

[58] M. D. Ekstrand and D. Kluver, "Exploring author gender in book rating and recommendation," *User modeling and user-adapted interaction*, vol. 31, no. 3, pp. 377–420, 2021.

[59] Q. Hu and H. Rangwala, "Metric-free individual fairness with cooperative contextual bandits," *CoRR*, vol. abs/2011.06738, 2020. [Online]. Available: https://arxiv.org/abs/2011.06738

[60] V. Patil, G. Ghalme, V. Nair, and Y. Narahari, "Achieving fairness in the stochastic multi-armed bandit problem," *CoRR*, vol. abs/1907.10516, 2019.

[61] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," in *2019 IEEE Conference on Computer Communications, INFOCOM 2019, Paris, France, April 29 - May 2, 2019*. IEEE, 2019, pp. 1702–1710.

[62] S. Gillen, C. Jung, M. J. Kearns, and A. Roth, "Online learning with an unknown fairness metric," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 2018, pp. 2605–2614.

[63] H. Heidari and A. Krause, "Preventing disparate treatment in sequential decision making," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. ijcai.org, 2018, pp. 2248–2254.

[64] Y. Sun, I. Ramírez, A. Cuesta-Infante, and K. Veeramachaneni, "Learning fair classifiers in online stochastic settings," *CoRR*, vol. abs/1908.07009, 2019.

[65] B. Metevier, S. Giguere, S. Brockman, A. Kobren, Y. Brun, E. Brunskill, and P. S. Thomas, "Offline contextual bandits with high probability fairness guarantees," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 14 893–14 904.

[66] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," in *Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 2921–2930.

[67] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness," in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, ser. Proceedings of Machine Learning Research, vol. 81. PMLR, 2018, pp. 172–186.

[68] E. Bareinboim, A. Forney, and J. Pearl, "Bandits with unobserved confounders: A causal approach," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[69] G. Tennenholtz, U. Shalit, S. Mannor, and Y. Efroni, "Bandits with partially observable confounded data," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 430–439.

[70] J. Zhang and E. Bareinboim, "Transfer learning in multi-armed bandit: a causal approach," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 2017, pp. 1778–1780.

[71] ——, "Bounding causal effects on continuous outcome," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021.

[72] N. Sharma, S. Basu, K. Shanmugam, and S. Shakkottai, "Warm starting bandits with side information from confounded data," *arXiv preprint arXiv:2002.08405*, 2020.

[73] Y. Li, H. Xie, Y. Lin, and J. C. Lui, "Unifying offline causal inference and online bandit learning for data driven decision," in *Proceedings of the Web Conference 2021*, 2021, pp. 2291–2303.

[74] Q. Tang and H. Xie, "A robust algorithm to unifying offline causal inference and online multi-armed bandit learning," in *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2021, pp. 599–608.

[75] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[76] L. Burgette, B. A. Griffin, and D. McCaffrey, "Propensity scores for multiple treatments: A tutorial for the mnps function in the twang package," *R package. Rand Corporation*, 2017.

[77] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[78] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000.

[79] I. Shpitser, T. VanderWeele, and J. M. Robins, "On the validity of covariate adjustment for estimating causal effects," *arXiv preprint arXiv:1203.3515*, 2012.

[80] J. D. Correa, J. Tian, and E. Bareinboim, "Identification of causal effects in the presence of selection bias," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2744–2751.

[81] J. Tian and J. Pearl, "A general identification condition for causal effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada*, R. Dechter, M. J. Kearns, and R. S. Sutton, Eds., pp. 567–573.

[82] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 208–214.

[83] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 661–670.

[84] Q. Wu, H. Wang, Q. Gu, and H. Wang, "Contextual bandits in a collaborative environment," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 529–538.

[85] T. Lattimore and C. Szepesvári, "Bandit algorithms," *preprint*, p. 28, 2018.

[86] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[87] W. Huang, K. Labille, X. Wu, D. Lee, and N. Heffernan, "Fairness-aware bandit-based recommendation," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 1273–1278.

[88] ——, "Achieving user-side fairness in contextual bandits," *Human-Centric Intelligent Systems*, pp. 1–14, 2022.

[89] F. Lattimore, T. Lattimore, and M. D. Reid, "Causal bandits: Learning good interventions via causal inference," in *NeurIPS'16*, 2016.

[90] R. Sen, K. Shanmugam, A. G. Dimakis, and S. Shakkottai, "Identifying best interventions through online importance sampling," in *ICML'17*, 2017, pp. 3057–3066.

[91] S. Lee and E. Bareinboim, "Structural causal bandits: Where to intervene?" in *NeurIPS'18*, 2018, pp. 2573–2583.

[92] ——, "Structural causal bandits with non-manipulable variables," in *AAAI'19*, 2019.

[93] Y. Lu, A. Meisami, A. Tewari, and W. Yan, "Regret analysis of bandit problems with causal background knowledge," in *UAI'20*, 2020, pp. 141–150.

[94] J. D. Correa and E. Bareinboim, "A calculus for stochastic interventions: Causal effect identification and surrogate experiments," in *AAAI'20*, 2020, pp. 10 093–10 100.

[95] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[96] D. Geiger, T. Verma, and J. Pearl, "d-separation: From theorems to algorithms," in *Machine Intelligence and Pattern Recognition*, 1990.

[97] J. Tian, A. Paz, and J. Pearl, *Finding minimal d-separators*. Citeseer, 1998.

[98] Y. Lu, A. Meisami, and A. Tewari, "Causal bandits with unknown graph structure," *arXiv preprint arXiv:2106.02988*, 2021.

[99] I. Shpitser and J. Pearl, "Complete identification methods for the causal hierarchy," *Journal of Machine Learning Research*, vol. 9, no. Sep, pp. 1941–1979, 2008.

[100] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang, "Stochastic bandits with linear constraints," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2827–2835.

[101] W. Huang, L. Zhang, and X. Wu, "Achieving counterfactual fairness for causal bandit," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6952–6959.

[102] J. K. Lunceford and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in medicine*, vol. 23, no. 19, pp. 2937–2960, 2004.

[103] C. Avin, I. Shpitser, and J. Pearl, "Identifiability of path-specific effects," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2, no. August, pp. 357–363, 2005.

[104] I. Shpitser and J. Pearl, "Effects of treatment on the treated: Identification and generalization," *arXiv preprint arXiv:1205.2615*, 2012.

[105] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson, "The tetrad project: Constraint based aids to causal model specification," *Multivariate Behavioral Research*, vol. 33, no. 1, pp. 65–117, 1998.

[106] L. Tang, R. Rosales, A. Singh, and D. Agarwal, "Automatic ad format selection via contextual bandits," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, pp. 1587–1594.

[107] V. Avadhanula, R. Colini Baldeschi, S. Leonardi, K. A. Sankararaman, and O. Schrijvers, "Stochastic bandits for multi-platform budget optimization in online advertising," in *Proceedings of the Web Conference 2021*, 2021, pp. 2805–2817.

[108] K. Ding, J. Li, and H. Liu, "Interactive anomaly detection on attributed networks," in *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019, pp. 357–365.

[109] E. Bareinboim and J. Tian, "Recovering causal effects from selection bias," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

[110] J. Tian, "Identifying conditional causal effects," in *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004*. AUAI Press, 2004, pp. 561–568.

[111] J. Tian and J. Pearl, "On the identification of causal effects," 2003.

[112] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[113] B. Hao, T. Lattimore, and C. Szepesvari, "Adaptive exploration in linear contextual bandit," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 3536–3545.

[114] D. B. Rubin, "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, Mar. 2005.

[115] Y. Wang and D. M. Blei, "The blessings of multiple causes," *arXiv preprint arXiv:1805.06826*, 2018.

[116] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[117] W. Huang, Y. Wu, and X. Wu, "Multi-cause discrimination analysis using potential outcomes," in *Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings 13*. Springer, 2020, pp. 224–234.

[118] J. Zhang and E. Bareinboim, "Equality of Opportunity in Classification: A Causal Approach," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 3675–3685.

[119] L. Zhang, Y. Wu, and X. Wu, "Situation Testing-Based Discrimination Discovery: A Causal Inference Approach," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, vol. 2016-Janua. IJCAI/AAAI Press, 2016, pp. 2718–2724.

[120] S. Verma and J. Rubin, "Fairness Definitions Explained," in *Proceedings of the International Workshop on Software Fairness*, ser. FairWare '18. New York, NY, USA: ACM, 2018, pp. 1–7.

[121] H. Heidari, V. Nanda, and K. Gummadi, "On the Long-term Impact of Algorithmic Decision Policies: Effort Unfairness and Feature Segregation through Social Learning," in *International Conference on Machine Learning*, 2019, pp. 2692–2701. [Online]. Available: http://proceedings.mlr.press/v97/heidari19a.html

[122] G. W. Imbens, "The role of the propensity score in estimating dose-response functions," *Biometrika*, vol. 87, no. 3, pp. 706–710, 2000.

[123] M. Lichman, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2013.

[124] D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette, "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in medicine*, vol. 32, no. 19, pp. 3388–3414, 2013.

[125] L. Zhang, Y. Wu, and X. Wu, "Achieving Non-Discrimination in Data Release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. New York, New York, USA: ACM Press, Nov. 2017, pp. 1335–1344.

[126] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

A Appendix

A.1 Nomenclature and Assumptions for Chapter 5

In our regret bound analysis of D-UCB and F-UCB algorithms, we follow several standard assumptions [100] to guarantee the correctness and the simplicity of the proofs.

Assumption 1. For all $t \in [T]$, both the error term of reward and the error term of counterfactual fairness discrepancy follow 1-sub-Gaussian distribution.

Assumption 2. For all $t \in [T]$, both the mean of reward and the mean of counterfactual fairness discrepancy are within $[0, 1]$.

Assumption 3. There exists a safe policy π_0, i.e., $\pi_0 \in \Pi_t$ such that $\Delta_{\pi_0} \leq \tau$ for each $t \in [T]$.

The last assumption introduces the existence of a safe policy at each round, which plays an important role in the regret bound analysis of F-UCB. The nomenclature used for the proof part is shown in Table A.1.

A.2 Proof of Theorem 5

Proof. Following the definition we can further define the expected reward mean of a certain policy as

$$\mu_\pi = \mathbb{E}_{a\sim\pi}[\mu_a|do(a)] = \mathbb{E}_{a\sim\pi}\left[\sum_{i=1}^{|\mathbf{Z}|}\mathbb{E}[R|\mathbf{W}=\mathbf{w}_i]P(\mathbf{Z}=\mathbf{z}_i|a)\right]$$

and the policy applied at each time t as $\pi_t = \text{argmax}_{\pi\in\Pi_t}\,\mathbb{E}_{a\sim\pi}[\mu_a]$.

Table A.1: Nomenclature.

$\mathcal{A}_t$	Arm set at time t
$\mathbf{A}, \mathbf{X}, R$	Arm features, user features, and reward
$\mathbf{W}$	d-separation set that separates R from $(\mathbf{A} \cup \mathbf{X}) \backslash \mathbf{W}$
$\mathbf{Z}$	The difference between the d-separation set $\mathbf{W}$ and $\mathbf{A} \cup \mathbf{X}$.
δ	With probability at least $1-\delta$ that the true reward is less than the estimated upper confidence bound for an arm a at time t
δ_E	With probability at least $1-\delta_E$ that the regret of causal fair bandit is bounded
δ'	With probability at least $1-\delta'$ that the true counterfactual discrepancy is less than its estimated upper confidence bound for an arm a at time t
$UCB_a(t)$	Upper confidence bound of the reward for action a based on the observed values up to time t
μ_a	Expected mean reward for arm a
μ_{a,s^*}	Estimated mean reward for arm a if $gender = s^*$ given the user's profile
μ_π	Expected mean reward for taking policy π
$\hat{\mu}_\pi(t)$	Estimated mean reward of a policy π based on the observed values up to time t
$\mathcal{R}_T$	Cumulative regret up to time T
α_r	Parameter that controls the scale of the confidence interval of reward
α_c	Parameter that controls the scale of the confidence interval of counterfactual discrepancy
γ_t	Parameter that could be tuned to ensure the fairness of a certain policy
E	Event under which all the true rewards are less than the estimated upper confidence bound
E_{cf}	Event under which all the counterfactual discrepancies are less than the estimated upper confidence bound

Let $N_{\mathbf{w}}(t) = \sum_{s=1}^{t} \mathbf{I}_{\mathbf{W}_s=\mathbf{w}}$ denote the count for a certain domain value of $\mathbf{W}$ up to time t. Further we define the mean of the reward related to a d-separation set domain value as $\mu_{\mathbf{w}} = \mathbb{E}[R|\mathbf{W} = \mathbf{w}]$ and its estimated value as $\hat{\mu}_{\mathbf{w}}(t) = \frac{1}{N_{\mathbf{w}}(t)} \sum_{s=1}^{t} R_{a_s} \mathbf{I}_{\mathbf{W}_s=\mathbf{w}}$.

We also define the upper confidence bound of the reward for each arm and the upper confidence bound for each policy:

$$UCB_{\mathbf{w}}(t) = \hat{\mu}_{\mathbf{w}}(t) + \sqrt{\frac{2\log(1/\delta)}{1 \vee N_{\mathbf{w}}(t)}}$$

$$UCB_a(t) = \sum_{\mathbf{Z}} UCB_{\mathbf{w}}(t) P(\mathbf{z}|\mathbf{x}_{t,a})$$

$$\mathbb{E}_{a\sim\pi}[UCB_a(t)] = \mathbb{E}_{a\sim\pi}\left[\sum_{\mathbf{Z}} UCB_{\mathbf{w}}(t) P(\mathbf{z}|\mathbf{x}_{t,a})\right]$$

Let E be the event that for all time $t \in [T]$ and value index $i \in [|\mathbf{W}|]$, we have

$$|\hat{\mu}_{\mathbf{w}_i}(t) - \mu_{\mathbf{w}_i}| \leq \sqrt{\frac{2\log(1/\delta)}{1 \vee N_{\mathbf{w}_i}(t)}}$$

Since $\hat{\mu}_{\mathbf{w}_i}(t)$ is the sample mean estimator of $\mu_{\mathbf{w}_i}$, and the error term follows sub-Gaussian distribution, we can show

$$P\left(|\hat{\mu}_{\mathbf{w}_i}(t) - \mu_{\mathbf{w}_i}| \geq \sqrt{\frac{2\log(1/\delta)}{1 \vee N_{\mathbf{w}_i}(t)}}\right) =$$
$$\mathbb{E}\left[P\left(|\hat{\mu}_{\mathbf{w}_i}(t) - \mu_{\mathbf{w}_i}| \geq \sqrt{\frac{2\log(1/\delta)}{1 \vee N_{\mathbf{w}_i}(t)}}\middle| \mathbf{w}_{(1)}, ..., \mathbf{w}_{(t)}\right)\right] \leq \mathbb{E}[2\delta] = 2\delta$$

where $\mathbf{w}_{(t)}$ denotes the observed values at time t. Thus by summing up the probabilities through all domain values of $t \in [T]$ and $i \in [|\mathbf{W}|]$, using union bound criteria we have

$P(E^c) = 1 - P(E) \leq 2\delta T|\mathbf{W}|$. The above result implies a lower probability bound for event E. The cumulative regret could be decomposed as

$$\begin{aligned}\mathcal{R}_T &= \sum_{t=1}^{T}(\mu_{a^*} - \mu_{a_t}) \\ &= \sum_{t=1}^{T}\left(\mu_{a^*} - UCB_{a_t}(t) + UCB_{a_t}(t) - \mu_{a_t}\right)\end{aligned}$$

Following the rule of optimism in the face of uncertainty, under event E we have

$$\begin{aligned}\mu_{a^*} &= \sum_{i=1}^{|\mathbf{Z}|}\mathbb{E}[R|\mathbf{W} = \mathbf{w}_i]P(\mathbf{Z} = \mathbf{z}_i|a^*) \\ &\leq \sum_{i=1}^{|\mathbf{Z}|}UCB_{\mathbf{w}_i}(t)P(\mathbf{Z} = \mathbf{z}_i|a^*) = UCB_{a^*}(t)\end{aligned}$$

As $UCB_{a^*}(t) \leq UCB_{a_t}(t)$ always holds due to OFU arm picking strategy, we have $\mu_{a^*} - UCB_{a_t}(t) \leq 0$.

With probability at least $1 - 2\delta T|\mathbf{W}|$, the cumulative regret can thus be further bounded by

$$\begin{aligned}\mathcal{R}_T &\leq \sum_{t=1}^{T}(UCB_{a_t}(t) - \mu_{a_t}) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{|\mathbf{Z}|}(UCB_{\mathbf{w}_i}(t) - \mu_{\mathbf{w}_i})P(\mathbf{Z} = \mathbf{z}_i|a_t) \\ &\leq \sum_{t=1}^{T}\sum_{i=1}^{|\mathbf{W}|}\sqrt{\frac{8\log(1/\delta)}{1 \vee N_{\mathbf{w}_i}(t)}}P(\mathbf{Z} = \mathbf{z}_i|a_t) \\ &= \sum_{t=1}^{T}\sum_{i=1}^{|\mathbf{W}|}\sqrt{\frac{8\log(1/\delta)}{1 \vee N_{\mathbf{w}_i}(t)}}\left(P(\mathbf{Z} = \mathbf{z}_i|a_t) - \mathbf{I}_{Z_{(t)}=Z_i}\right) + \sum_{t=1}^{T}\sum_{i=1}^{|\mathbf{W}|}\sqrt{\frac{8\log(1/\delta)}{1 \vee N_{\mathbf{w}_i}(t)}}\left(\mathbf{I}_{Z_{(t)}=Z_i}\right)\end{aligned} \tag{A.1}$$

The second part of Equation A.1 is bounded by

$$\begin{aligned}\sum_{t=1}^{T}\sum_{i=1}^{|\mathbf{W}|}\sqrt{\frac{8\log(1/\delta)}{1\vee N_{\mathbf{w}_i}(t)}}\mathbf{I}_{Z_{(t)}=Z_i} &\leq \sum_{i=1}^{|\mathbf{W}|}\int_0^{N_{\mathbf{w}_i}(T)}\sqrt{\frac{8\log(1/\delta)}{s}}ds \\ &\leq \sum_{i=1}^{|\mathbf{W}|}\sqrt{32N_{\mathbf{w}_i}(T)\log(1/\delta)} \\ &\leq \sqrt{32|\mathbf{W}|T\log(1/\delta)}\end{aligned}$$

We will use the following proposition called Azuma's inequality to derive the bound of the first term of Equation A.1.

Proposition 1. Suppose $\{M_k : k = 0, 1, 2...\}$ is a martingale and $|M_k - M_{k-1}| < c_k$ almost surely, then for all $t \in [T]$ and positive value ϵ we have:

$$P(|M_t - M_0| > \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2\sum_{k=1}^{t} c_k^2}\right)$$

For the first part, we further define

$$M_t = \sum_{s=1}^{t}\sum_{i=1}^{|\mathbf{W}|}\sqrt{\frac{8\log(1/\delta)}{1\vee N_{\mathbf{w}_i}(t)}}\left(P(\mathbf{Z}=\mathbf{z}_i|a_t) - \mathbf{I}_{Z_{(t)}=Z_i}\right)$$

with $M_0 = 0$, since $\{M_t\}_{t=0}^{T}$ is a martingale sequence, we have

$$|M_t - M_{t-1}|^2 = \left|\sum_{i=1}^{|\mathbf{W}|}\sqrt{\frac{8\log(1/\delta)}{1\vee N_{\mathbf{w}_i}(t)}}\Big(P(\mathbf{Z}=\mathbf{z}_i|a_t) - \mathbf{I}_{Z_{(t)}=Z_i}\Big)\right|^2 \leq 32\log(1/\delta)$$

which shows $|M_t - M_{t-1}|$ is bounded for any $t \in [T]$. Applying Azuma's inequality, we have

$$\begin{aligned} & P\left(|M_T - M_0| > \sqrt{|\mathbf{W}| T log(T)} \log(T)\right) \\ =& P\left(|M_T| > \sqrt{|\mathbf{W}| T \log(T)} \log(T)\right) \leq \exp(-\frac{|\mathbf{W}| \log^3(T)}{32 \log(1/\delta)}) \end{aligned}$$

The formula above gives a high probability bound of the first part. Now we can combine the bounds of two parts in Equation A.1 to derive the high probability bound of $\mathcal{R}_T$. Since $P(E^c) \leq 2\delta T|\mathbf{W}|$, applying union bound rule, with probability at least $1 - 2\delta T|\mathbf{W}| - \exp(-\frac{|\mathbf{W}| \log^3(T)}{32 \log(1/\delta)})$, the regret is bounded by:

$$\mathcal{R}_T \leq \sqrt{|\mathbf{W}| T \log(T)} \log(T) + \sqrt{32 |\mathbf{W}| T \log(1/\delta)} \tag{A.2}$$

□

Corollary 3. By setting $\delta = 1/T^2$, the causal bandit algorithm achieves $\tilde{O}(\sqrt{|\mathbf{W}| \cdot T})$ regret bound.

Proof. Plugging in the value $\delta = 1/T^2$, with probability at least $1 - 2|\mathbf{W}|/T - \exp(-\frac{|\mathbf{W}| \log^2(T)}{64})$, the regret is bounded by

$$R_T \leq 16 \sqrt{|\mathbf{W}| T \log(T)} \log(T)$$

The above formula thus leads to $\tilde{O}(\sqrt{|\mathbf{W}| \cdot T})$ long-term expected regret.

□

A.3 Proof of Theorem 6

Proof. If a set of attributes $\mathbf{B} \subseteq \mathbf{X} \backslash \{S\}$ are descendants of S, $\mathbb{E}[R(a, s^*)|\mathbf{x}_t]$ is not identifiable. According to Proposition 2 in [9], we have that

$$\mathbb{E}[R(a, s^*)|\mathbf{x}_t] \leq \sum_{\mathbf{I}} \frac{P(\mathbf{x}_{t,a}, \mathbf{i})}{P(\mathbf{x}_{t,a})} \max_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{x}_{t,a} \backslash s_t, \mathbf{i}]\}$$

It follows that

$$\begin{aligned}\mathbb{E}[R(a, s^*)|\mathbf{x}_t] &\leq \sum_{\mathbf{I}} P(\mathbf{i}|\mathbf{x}_{t,a}) \max_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{x}_{t,a} \backslash s_t, \mathbf{i}]\} \\ &= \sum_{\mathbf{Z}, \mathbf{I} \backslash \mathbf{Z}} P(\mathbf{z}|\mathbf{x}_{t,a}) P(\mathbf{i} \backslash \mathbf{z}|, \mathbf{z}, \mathbf{x}_{t,a}) \max_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{w} \backslash s_t]\},\end{aligned}$$

where $\mathbf{Z}$ and $\mathbf{W}$ are defined following Section 3.2 in the main paper. We claim that $\mathbf{W}$ has no intersection with $\mathbf{I} \backslash \mathbf{Z}$. Otherwise, there exists an attribute $I \in \mathbf{I}$ which belongs to $\mathbf{W}$ but not $\mathbf{Z}$. This contradicts to the definition of $\mathbf{Z}$, which is given by $\mathbf{W}$ subtracting $\mathbf{A} \cup \mathbf{X}$. Thus, it follows that

$$\begin{aligned}\mathbb{E}[R(a, s^*)|\mathbf{x}_t] &\leq \sum_{\mathbf{Z}} \max_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{w} \backslash s_t]\} P(\mathbf{z}|\mathbf{x}_{t,a}) \cdot \sum_{\mathbf{I} \backslash \mathbf{Z}} P(\mathbf{i} \backslash \mathbf{z}|, \mathbf{z}, \mathbf{x}_{t,a}) \\ &= \sum_{\mathbf{Z}} \max_{\mathbf{b}} \{\mathbb{E}[R|s^*, \mathbf{w} \backslash s_t]\} P(\mathbf{z}|\mathbf{x}_{t,a})\end{aligned}$$

□

A.4 Proof of Theorem 7

Proof. Similar to the regret analysis of causal bandit, we decompose the cumulative regret $\mathcal{R}_T$ into two parts.

$$\begin{aligned}\mathcal{R}_T &= \sum_{t=1}^{T} \left(\mathbb{E}_{a\sim\pi^*}[\mu_a] - \mathbb{E}_{a\sim\pi_t}[\mu_a]\right) \\ &= \left(\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi^*}[\mu_a] - \mathbb{E}_{a\sim\pi_t}[UCB_a(t)]\right) + \left(\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[UCB_a(t)] - \mathbb{E}_{a\sim\pi_t}[\mu_a]\right) \end{aligned} \tag{A.3}$$

We will further bound $\mathcal{R}_T$ by proving the first term of Equation A.3 is less than 0 and the second term could be bounded by adopting the upper confidence bound analysis approach. In the fair bandit setting we introduce another event E_{cf} that implies the estimation error of the counterfactual discrepancy is bounded. That is, for all time $t \in [T]$ and a policy π, with probability at least $1 - \delta'$,

$$|\hat{\Delta}_\pi(t) - \Delta_\pi| \leq \sqrt{\frac{8\log(1/\delta')}{1 \vee N_a(t)}}$$

First we define the inflated upper confidence bound with scale parameters for the mean reward and fairness discrepancy as

$$UCB_a(t) = \hat{\mu}_a + \alpha_r \beta_a(t)$$

$$UCB_{\Delta_\pi}(t) = \hat{\Delta}_\pi + \alpha_c \beta_a(t), \textit{ where } \beta_a(t) = \sqrt{2\log(1/\delta')/N_a(t)}$$

Notice that the event E_{cf} will always happen if the event E happens. Under event

$E \cap E_{cf} = E$ and with the assumption that $\alpha_r, \alpha_c \geq 1$, we have

$$
\begin{aligned}
(\alpha_r - 1)\beta_a(a) \leq \epsilon_a^r(t) \leq (\alpha_r + 1)\beta_a(a),\ \forall a \in \mathcal{A}_t \\
(\alpha_c - 1)\beta_a(a) \leq \epsilon_a^c(t) \leq (\alpha_c + 1)\beta_a(a),\ \forall a \in \mathcal{A}_t
\end{aligned} \tag{A.4}
$$

where $\epsilon_a^r(t)$ and $\epsilon_a^c(t)$ are the error term of the reward and counterfactual discrepancy. If the optimal policy belongs to the fair policy subspace, which means $\pi^* \in \bar{\Phi}_t$, we can easily get:

$$
\mathbb{E}_{a\sim\pi^*}[\mu_a] \leq \mathbb{E}_{a\sim\pi^*}[UCB_a(t)] \leq \mathbb{E}_{a\sim\pi_t}[UCB_a(t)]
$$

Now we assume $\pi^* \notin \bar{\Phi}_t$, that is, $\mathbb{E}_{a_t\sim\pi^*}[UCB_{\Delta_{\pi^*}}(t)] > \tau$. Let $\pi^* = \rho^*\bar{\pi}^* + (1-\rho^*)\pi_0$, where $\bar{\pi}^*$ denotes the optimal policy in the policy subspace $\bar{\Phi}_t \setminus \pi_0$.

Consider the mixed policy of π^* and π_0, denoted as $\tilde{\pi} = \gamma_t\pi^* + (1-\gamma_t)\pi_0 = \gamma_t\rho^*\bar{\pi}^* + (1-\gamma_t\rho^*)\pi_0$, where $\gamma_t \in [0,1]$ is the maximum value to ensure $\tilde{\pi} \in \Phi_t$. One feasible solution for γ_t is

$$
\begin{aligned}
\gamma_t &= \frac{\tau - \Delta_{\pi_0}}{\rho^*\mathbb{E}_{a_t\sim\bar{\pi}_t}[UCB_{\Delta_{\bar{\pi}_t}}(t)] - \rho^*\Delta_{\pi_0}} \\
&= \frac{\tau - \Delta_{\pi_0}}{\mathbb{E}_{a_t\sim\bar{\pi}_t}[\rho^*(\bar{c}_a + \epsilon_a^c(t))] - \rho^*\Delta_{\pi_0}} \\
&\geq \frac{\tau - \Delta_{\pi_0}}{\tau - \Delta_{\pi_0} + \rho^*(1+\alpha_c)\mathbb{E}_{a_t\sim\bar{\pi}^*}[\beta_a(t)]}
\end{aligned}
$$

Denote $\frac{\tau-\Delta_{\pi_0}}{\tau-\Delta_{\pi_0}+\rho^*(1+\alpha_c)\mathbb{E}_{a\sim\bar{\pi}^*}[\beta_a(t)]}$ as Γ. From the design of the integrated policy $\tilde{\pi}_t$ we further

have:

$$\mathbb{E}_{a\sim\pi_t}[UCB_a(t)] \geq \gamma_t \mathbb{E}_{a_t\sim\pi^*}[UCB_a(t)] + (1-\gamma_t)UCB_a(t) \tag{A.5}$$

$$\geq \Gamma \times \mathbb{E}_{a_t\sim\pi^*}[UCB_a(t)] \tag{A.6}$$

$$= \Gamma \times (\mathbb{E}_{a_t\sim\pi^*}[\mu_a] + \mathbb{E}_{a_t\sim\pi^*}[\epsilon_{a_t}])$$

$$\geq \Gamma \times (\mathbb{E}_{a_t\sim\pi^*}[\mu_a] + (\alpha_r - 1)\mathbb{E}_{a_t\sim\pi^*}[\beta_t]) \tag{A.7}$$

$$\geq \frac{\tau - \Delta_{\pi_0}}{\tau - \Delta_{\pi_0} + (1+\alpha_c)\mathbb{E}_{a\sim\pi^*}[\beta_a(t)]} \times (\mathbb{E}_{a_t\sim\pi^*}[\mu_a] + (\alpha_r - 1)\mathbb{E}_{a_t\sim\pi^*}[\beta_t]) \tag{A.8}$$

For the above derivation, Equation A.6 holds because $UCB_a(t) > 0$, Equation A.7 is the consequence of Equation A.4, Equation A.8 is derived based on the fact that $\mathbb{E}_{a_t\sim\pi^*}[\beta_t] = \rho^*\mathbb{E}_{a_t\sim\bar{\pi}^*}[\beta_t] + (1-\rho^*\beta_0(t)) \geq \rho^*\mathbb{E}_{a_t\sim\bar{\pi}^*}[\beta_t]$. Denote the term in Equation A.8 as C_0. Let $C_1 = \mathbb{E}_{a\sim\pi^*}[\beta_a(t)]$, it holds that

$$C_0 = \frac{\tau - \Delta_{\pi_0}}{\tau - \Delta_{\pi_0} + (1+\alpha_c)C_1} \times (\mathbb{E}_{a_t\sim\pi^*}[\mu_a] + (\alpha_r - 1)C_1)$$

$C_0 > \mathbb{E}_{a\sim\pi^*}[\mu_a]$ is satisfied if and only if:

$$(\tau - \Delta_{\pi_0})(\alpha_r - 1)C_1 \geq (1+\alpha_c)C_1\mathbb{E}_{a\sim\pi^*}[\mu_a]$$

Since $\mathbb{E}_{a\sim\pi^*}[\mu_a] \leq 1$, the above inequation holds if $(\tau - \Delta_{\pi_0})(\alpha_r - 1)C_1 \geq 1 + \alpha_c$. Thus, by setting $\delta_E = 4|\mathbf{W}|T\delta$ and $\alpha_r, \alpha_c \geq 1, \alpha_c \leq \tau(\alpha_r - 1)$, following simple union bound rule implies that with probability at least $1 - \frac{\delta_E}{2}$, we have

$$\sum_{t=1}^{T} (\mathbb{E}_{a\sim\pi^*}[\mu_a] - \mathbb{E}_{a_t\sim\pi_t}[UCB_a(t)]) \leq 0$$

Next we will derive the bound of the second term in Equation A.3. The result is given by the following proposition.

Proposition 2. If $\delta_E = 4|\mathbf{W}|T\delta$ for a $\delta \in (0,1)$, then with probability at least $1 - \frac{\delta_E}{2}$, we have

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[UCB_a(t)] - \mathbb{E}_{a\sim\pi_t}[\mu_a] \leq (\alpha_r + 1)\left(2\sqrt{2T|\mathbf{W}|\log(1/\delta)} + 4\sqrt{T\log(2/\delta_E)\log(1/\delta)}\right) \tag{A.9}$$

Under the conditions in the proposition, we have $P(E) \geq 1 - \frac{\delta_E}{2}$. Under the event E, we have $R_{a_t} \in [\hat{\mu}_{a_t} - \beta_a(t), \hat{\mu}_{a_t} + \beta_a(t)]$ for all $t \in [T]$ and $a \in \mathcal{A}$. Thus for all t we could further derive

$$\mathbb{E}_{a\sim\pi_t}[UCB_a(t)] - \mathbb{E}_{a\sim\pi_t}[\mu_a] \leq (\alpha_r + 1)\mathbb{E}_{a\sim\pi_t}[\beta_a(t)]$$

Let $\mathcal{F}_{t-1}$ be the σ-algebra defined up to the choice of π_t and a'_t be another choice picked from policy $\pi_t|\mathcal{F}_{t-1}$. a'_t is conditionally independent of a_t, which means $a'_t \perp\!\!\!\perp a_t|\mathcal{F}_{t-1}$. By definition the following equality holds:

$$\mathbb{E}_{a\sim\pi_t}[\beta_a(t)] = \mathbb{E}_{a'_t\sim\pi_t}[\beta_a(t)|\mathcal{F}_{t-1}]$$

Setting $A_t = \mathbb{E}_{a'_t\sim\pi_t}[\beta_a(t)|\mathcal{F}_{t-1}] - \beta_{a_t}(t)$, $M_t = \sum_{s=1}^{t} A_t$ is thus a martingale sequence with $|M_t - M_{t-1}| = |A_t| \leq 2\sqrt{2\log(1/\delta)}$. Thus applying Azuma-Hoeffding inequality implies:

$$P\left(\sum_{t=1}^{T} \mathbb{E}_{a\sim\pi_t}[\beta_a(t)] \geq \sum_{t=1}^{T} \beta_{a_t}(t) + 4\sqrt{T\log(2/\delta_E)\log(1/\delta)}\right) \leq \delta_E/2$$

We denote the event that describes the results of the above inequality as E_A. In the equation above, the sum of the adaptive scaling parameter could be decomposed as follows:

$$\sum_{t=1}^{T} \beta_{a_t}(t) = \sum_{a \in \mathcal{A}} \sum_{t=1}^{T} \mathbf{I}_{\{a_t = a\}} \beta_a(t) = \sum_{i \in |\mathbf{W}|} \sum_{t=1}^{T} \mathbf{I}_{\{\mathbf{W}_t = \mathbf{w}_i\}} \beta_{\mathbf{w}_i}(t)$$

Under event E, for each domain value of the d-separation set $\mathbf{W}$ we have:

$$\sum_{t=1}^{T} \mathbf{I}_{\{\mathbf{W}_t = \mathbf{w}_i\}} \beta_{\mathbf{w}_i}(t) = \sqrt{2\log(1/\delta)} \sum_{t=1}^{N_{\mathbf{w}_i}(T)} \frac{1}{\sqrt{t}} \leq 2\sqrt{2N_{\mathbf{w}_i}(T)\log(1/\delta)}$$

Since $\sum_{i \in |\mathbf{W}|} N_{\mathbf{w}_i}(T) = T$, using the fact that arithmetic mean is less than quadratic mean we have:

$$\sum_{i \in |\mathbf{W}|} 2\sqrt{2N_{\mathbf{w}_i}(T)\log(1/\delta)} \leq 2\sqrt{2T|\mathbf{W}|\log(1/\delta)}$$

Conditioning on the event $E \cap E_A$ whose probability satisfies $P(E \cap E_A) \geq 1 - \delta_E$, we have

$$P\left(\mathbb{E}_{a\sim\pi_t}[UCB_a(t)] - \mathbb{E}_{a\sim\pi_t}[\mu_a] \geq (\alpha_r + 1)\left(2\sqrt{2T|\mathbf{W}|\log(1/\delta)} + 4\sqrt{T\log(2/\delta_E)\log(1/\delta)}\right)\right) \leq \delta_E/2 \tag{A.10}$$

which is exactly the result of Proposition 2.

Finally, combining the theoretical derivation of the two parts above leads to the cumulative regret bound shown in Theorem 7. □

A.5 Proof of Theorem 8

To prove Theorem 8, we firstly introduce the subscript **nd** to denote all the variables in such set that are not descendants of any variable in $\mathbf{I}$, that is, for an arbitrary set $\mathbf{A}$, let $\mathbf{A_{nd}} = \{A \in \mathbf{A} | A \notin De(\mathbf{I})\}$, where $De(\mathbf{I})$ denotes the decedents of $\mathbf{I}$. Similarly we define

the subscript $\mathbf{d}$ as all the variables in such set that are descendants of any variable in $\mathbf{I}$. We then define several subsets of $\mathbf{Z}$ and conditional independence related to them.

1. $\mathbf{Z} = \mathbf{Z^M} \cup \mathbf{Z}^\intercal, \mathbf{Z^S} = \{Z \in \mathbf{Z^M} | (Z \perp\!\!\!\perp S | \mathbf{Z}^\intercal)\}$

2. $\mathbf{C}$: Replace the bi-directional edge with a node and two arrows into two connecting nodes, starting from the augmented node. $\mathbf{C}$ denotes the set of all new variables introduced by the augmentation process.

3. $\mathbf{L}_1$: Variables in $(\mathbf{V} \cup \mathbf{C}) \backslash (\mathbf{Z} \cup \mathbf{I} \cup \mathbf{U} \cup Y)$ such that:

   - are d-connected to Y given $\mathbf{Z}^\intercal, \mathbf{Z^S}$ in $\mathcal{G} \backslash \mathbf{I}$;
   - are not descendants of $\mathbf{I}$;
   - are ancestors of some $Z \in \mathbf{Z}^\intercal \cup \mathbf{Z^S}$.

4. $\mathbf{L}_2$: Variables in $(\mathbf{V} \cup \mathbf{C}) \backslash (\mathbf{Z} \cup \mathbf{I} \cup \mathbf{U} \cup Y)$ such that:

   - are d-connected to Y given $\mathbf{Z}$ in $\mathcal{G} \backslash \mathbf{I}$;
   - are independent of $\mathbf{I}$ given $\mathbf{Z}^\intercal$, $\mathbf{Z^S}$ and S on $\mathcal{G}_{\overline{\mathbf{I}(\mathbf{Z^S}, \mathbf{Z}^\intercal, S)}}$;
   - are ancestors of some $Z \in \mathbf{Z}$.

5. $\mathbf{Z^X} = \left\{ Z \in \mathbf{Z} \backslash (\mathbf{Z}^\intercal \cup \mathbf{Z^S}) | (Z \perp\!\!\!\perp \mathbf{I} | \mathbf{Z^S}, \mathbf{Z}^\intercal, S)_{\mathcal{G}_{\overline{\mathbf{I}(\mathbf{Z^S}, \mathbf{Z}^\intercal, S)}}} \right\}$.

6. $\mathbf{Z^Y} = \mathbf{Z} \backslash (\mathbf{Z^S} \cup \mathbf{Z}^\intercal \cup \mathbf{Z^X})$.

Suppose in the causal graph $\mathcal{G}$ there are sets of variables $\mathbf{Z}, \mathbf{X}$ and Y, such that $(\mathbf{Z}, \mathbf{Z}^\intercal)$ is a valid adjustment set that meets the condition in Theorem 8. Then, the following conditional independence formula hold in the augmented graph of $\mathcal{G}$:

1. $(Y \perp\!\!\!\perp \mathbf{Z}^\intercal_{\mathbf{d}}, \mathbf{Z^S_d} | \mathbf{L}_1, \mathbf{Z}^\intercal_{\mathbf{nd}}, \mathbf{Z^S_{nd}}, \mathbf{U}, \mathbf{I})_{\mathcal{G}_{\overline{\mathbf{I}}}}$

2. $(Y \perp\!\!\!\perp S|\mathbf{Z}^{\intercal}, \mathbf{U}, \mathbf{I})_{\mathcal{G}_{\overline{\mathbf{I}}}}$

3. $(Y \perp\!\!\!\perp S|\mathbf{Z}^{\intercal}, \mathbf{Z}^{\mathbf{S}}, \mathbf{U}, \mathbf{I})_{\mathcal{G}_{\overline{\mathbf{I}}}}$

4. $(\mathbf{L}_1 \perp\!\!\!\perp \mathbf{I}|\mathbf{U}, \mathbf{Z}^{\mathbf{S}}, \mathbf{Z}^{\intercal})_{\mathcal{G}_{\overline{\mathbf{I}(\mathbf{Z}^{\mathbf{S}}, \mathbf{Z}^{\intercal})}}}$

5. $(Y \perp\!\!\!\perp \mathbf{Z}^{\mathbf{Y}}|\mathbf{Z}^{\intercal}, \mathbf{Z}^{\mathbf{S}}, \mathbf{U}, \mathbf{L}_2, \mathbf{Z}^{\mathbf{X}}, \mathbf{I}, S)_{\mathcal{G}_{\overline{\mathbf{I}}}}$

6. $(\mathbf{L}_2 \perp\!\!\!\perp \mathbf{I}|\mathbf{I}, \mathbf{U}, S)_{\mathcal{G}_{\overline{\mathbf{I}(\mathbf{Z}, S)}}}$

Based on the above theoretical results, we are ready to give the derivation of Equation 6.2 in Theorem 8. Each step and intermediate equations are derived based on rules of conditional independence and basic laws of conditional probability and marginalization in probability theory and Bayesian inference. Firstly, we marginalize the target distribution on $\mathbf{L}_1, \mathbf{Z}^{\intercal}_{\mathbf{nd}}$ and $\mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}$ to derive Equation A.11.

$$\begin{aligned} & P(Y = y|do(\mathbf{I} = \mathbf{i}), \mathbf{U} = \mathbf{u}) \\ &= \sum_{\mathbf{L}_1, \mathbf{Z}^{\intercal}_{\mathbf{nd}}, \mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}} P(y|do(\mathbf{i}), \mathbf{u}, \mathbf{l}_1, \mathbf{z}^{\intercal}_{\mathbf{nd}}, \mathbf{z}^{\mathbf{S}}) P(\mathbf{l}_1, \mathbf{z}^{\intercal}_{\mathbf{nd}}, \mathbf{z}^{\mathbf{S}}|do(\mathbf{i}), \mathbf{u}) \\ &= \sum_{\mathbf{L}_1, \mathbf{Z}^{\intercal}_{\mathbf{nd}}, \mathbf{Z}^{\mathbf{S}}_{\mathbf{nd}}} P(y|do(\mathbf{i}), \mathbf{u}, \mathbf{l}_1, \mathbf{z}^{\intercal}_{\mathbf{nd}}, \mathbf{z}^{\mathbf{S}}_{\mathbf{nd}}) P(\mathbf{l}_1, \mathbf{z}^{\intercal}_{\mathbf{nd}}, \mathbf{z}^{\mathbf{S}}_{\mathbf{nd}}|\mathbf{u}) \end{aligned} \tag{A.11}$$

Based on independence condition 1 we then leverage variables $\mathbf{Z}^{\intercal}_{\mathbf{d}}, \mathbf{Z}^{\mathbf{S}}_{\mathbf{d}}$ to form $\mathbf{Z}^{\intercal}, \mathbf{Z}^{\mathbf{S}}$ in the summation. Based on independence condition 2 and 3 we can introduce the selection variable S and *do* operator in Equation A.12.

$$
\begin{aligned}
&= \sum_{\mathbf{L}_1,\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{l}_1,\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}})P(\mathbf{l}_1,\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}}|\mathbf{u}) \\
&= \sum_{\mathbf{L}_1,\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{l}_1,\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}})P(\mathbf{l}_1|\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}}) \times P(\mathbf{z}^{\mathbf{S}}|\mathbf{z}^\intercal,\mathbf{u})P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{L}_1,\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{l}_1,\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}})P(\mathbf{l}_1|\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}}) \times P(\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{L}_1,\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{l}_1,\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}})P(\mathbf{l}_1|do(\mathbf{i}),\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}}) \times P(\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u})
\end{aligned}
\tag{A.12}
$$

By independence condition 4 we can sum out the term $\mathbf{L}_1$ and introduce S in the first term of Equation A.12. We then condition on $\mathbf{L}_2$, $\mathbf{Z}^{\mathbf{Y}}$ and $\mathbf{Z}^{\mathbf{X}}$. By the subset segmentation $\mathbf{Z}^{\mathbf{Y}} = \mathbf{Z}\backslash(\mathbf{Z}^{\mathbf{S}} \cup \mathbf{Z}^\intercal \cup \mathbf{Z}^{\mathbf{X}})$ and independence condition 5 we are able to merge the terms related to $\mathbf{Z}$ in the summation and get Equation A.13.

$$
\begin{aligned}
&= \sum_{\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}})P(\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}},S=1)P(\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{Z}^\intercal,\mathbf{Z}^{\mathbf{S}},\mathbf{L}_2,\mathbf{Z}^{\mathbf{X}}} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}},\mathbf{l}_2,\mathbf{z}^{\mathbf{X}},S=1) \times P(\mathbf{l}_2,\mathbf{z}^{\mathbf{X}}|\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}},S=1)P(\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{Z},\mathbf{L}_2} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{z},\mathbf{l}_2,S=1)P(\mathbf{l}_2,\mathbf{z}^{\mathbf{Y}},\mathbf{z}^{\mathbf{X}},\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}},S=1) \times P(\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{Z},\mathbf{L}_2} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{z},\mathbf{l}_2,S=1)P(\mathbf{l}_2|\mathbf{z}^{\mathbf{Y}},\mathbf{z}^{\mathbf{X}},\mathbf{u},\mathbf{z}^\intercal,\mathbf{z}^{\mathbf{S}},S=1) \times P(\mathbf{z}^{\mathbf{Y}},\mathbf{z}^{\mathbf{X}},\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u}) \\
&= \sum_{\mathbf{Z},\mathbf{L}_2} P(y|do(\mathbf{i}),\mathbf{u},\mathbf{z},\mathbf{l}_2,S=1)P(\mathbf{l}_2|do(\mathbf{i}),\mathbf{z},\mathbf{u},S=1) \times P(\mathbf{z}^{\mathbf{Y}},\mathbf{z}^{\mathbf{X}},\mathbf{z}^{\mathbf{S}}|\mathbf{u},\mathbf{z}^\intercal,S=1)P(\mathbf{z}^\intercal|\mathbf{u})
\end{aligned}
\tag{A.13}
$$

Finally, using independence condition 6 we sum out $\mathbf{L}_2$ in the summation. Since the adjustment pair $(\mathbf{Z},\mathbf{Z}^\intercal)$ satisfies the generalized adjustment criterion for conditional intervention, we are able to introduce S in the first two terms and derive Equation A.14,

which is the same as Equation 6.2 in Theorem 8.

$$
\begin{aligned}
&= \sum_{\mathbf{Z}} P(y|do(\mathbf{i}), \mathbf{u}, \mathbf{z}, \mathbf{l}_2, S=1) P(\mathbf{z}^{\mathbf{Y}}, \mathbf{z}^{\mathbf{X}}, \mathbf{z}^{\mathbf{S}}|\mathbf{u}, \mathbf{z}^{\intercal}, S=1) P(\mathbf{z}^{\intercal}|\mathbf{u}) \\
&= \sum_{\mathbf{Z}} P(y|do(\mathbf{i}), \mathbf{u}, \mathbf{z}, \mathbf{l}_2, S=1) P(\mathbf{z}\backslash\mathbf{z}^{\intercal}|\mathbf{u}, \mathbf{z}^{\intercal}, S=1) P(\mathbf{z}^{\intercal}|\mathbf{u}) \\
&= \sum_{\mathbf{Z}} P(y|\mathbf{i}, \mathbf{z}, \mathbf{u}, S=1) P(\mathbf{z}\backslash\mathbf{z}^{\intercal}|\mathbf{z}^{\intercal}, \mathbf{u}, S=1) P(\mathbf{z}^{\intercal}|\mathbf{u})
\end{aligned}
\tag{A.14}
$$

A.6 Proof of Theorem 10

The path-specific causal effect can be computed from the observational data if and only if the recanting witness criterion is not satisfied [103]. Note that to calculate the second term $P(y|do(\mathbf{i}_1))$ in the presence of confounding bias and selection bias, we can directly follow the adjustment formula shown in Equation 3 and obtain

$$
P(y|do(\mathbf{i}_1)) = \sum_{\mathbf{Z}} P(y|\mathbf{i}_1, \mathbf{z}, S=1) P(\mathbf{z}\backslash\mathbf{z}^{\intercal}|\mathbf{z}^{\intercal}, S=1) P(\mathbf{z}^{\intercal}) \tag{A.15}
$$

We then aim to compute the second term $P(Y = y|do(\mathbf{i}_2|_{\pi}, \mathbf{i}_1|_{\bar{\pi}}))$ in the presence of confounding and selection biases if some unbiased observations can be further collected. To derive Equation 6.8, We first follow the axiomatic truncated factorization formula of an intervention:

$$
P(y|do(\mathbf{i}_1)) = \sum_{\mathbf{v}'} \prod_{V \in \mathbf{V}\backslash\mathbf{I}} P(v|\mathbf{pa}_V) \delta_{\mathbf{I}=\mathbf{i}_1} \tag{A.16}
$$

where $\mathbf{V}' = \mathbf{V}\backslash\{\mathbf{I}, Y\}$, $\mathbf{pa}_V$ denotes the realization of the parents of node V, and $\delta_{\mathbf{I}=\mathbf{i}_1}$ denotes assigning variables in $\mathbf{I}$ involved in the term ahead with the corresponding values in $\mathbf{i}$. We then divide the children of $\mathbf{I}$ into two sets: C_{π} and $C_{\bar{\pi}}$. Specifically, C_{π} contains each of $\mathbf{I}$'s children C, where the causal path $I \to C$ is a segment of a path in π. $C_{\bar{\pi}}$ contains each

of $\mathbf{I}$'s children C, where either C is not included in any path from I to Y, or the causal path $I \to C$ is a segment of a path not in π. Finally, we replace values $\mathbf{i}_1$ with $\mathbf{i}_2$ for the terms corresponding to nodes in C_π, and keep values $\mathbf{i}_1$ unchanged for the terms corresponding to nodes in $C_{\bar{\pi}}$. We split and classify nodes that belong to different subsets and factorize $P(Y = y|do(\mathbf{i}_2|_\pi, \mathbf{i}_1|_{\bar{\pi}}))$ into the following probability truncation formula:

$$\sum_{\mathbf{v}'} \Big(\prod_{G \in C_\pi} P(g|\mathbf{i}_2, \mathbf{pa}_G \backslash \{\mathbf{I}\}) \prod_{H \in C_{\bar{\pi}}} P(h|\mathbf{i}_1, \mathbf{pa}_H \backslash \{\mathbf{I}\}) \prod_{O \in \mathbf{V}_c \backslash\ CH_\mathbf{I}} P(o|\mathbf{pa}_O) \prod_{Z \in \mathbf{Z}} P(z|\mathbf{pa}_Z) \prod_{Q \in \mathbf{V}' \backslash \{\mathbf{V}_c \cup \mathbf{Z}\}} P(q|\mathbf{pa}_Q) \Big) \tag{A.17}$$

where $\mathbf{V}_c$ denotes the nodes that lie on the causal paths except $\mathbf{I}$. Note that the above computation requires that C_π and $C_{\bar{\pi}}$ are two disjoint subsets. Thus, the recanting witness criterion is not satisfied, and the path-specific treatment effect transmitted solely through causal paths is identifiable and can be computed according to Equation A.17.

Let $\mathbf{PA}$ denote Y's parent nodes along all causal paths, $\mathbf{PA}_\pi$ denote Y's parent nodes that lie in π, and $\mathbf{PA}_{\bar{\pi}}$ denote the remaining parents along the causal paths. We can compute

$P(Y = y|do(\mathbf{i}_2|_\pi, \mathbf{i}_1|_{\bar{\pi}}))$ by adjusting on a valid covariate set $\mathbf{Z}$:

$$
\begin{aligned}
& P(Y = y|do(\mathbf{i}_2|_\pi, \mathbf{i}_1|_{\bar{\pi}})) && \text{(A.18)} \\
&= \sum_{\mathbf{Z}\cup\mathbf{PA}} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{Z})P(\mathbf{pa}_{\bar{\pi}}|\mathbf{i_1}, \mathbf{z})P(y|\mathbf{pa}, \mathbf{z})P(\mathbf{z}) && \text{(a)} \\
&= \sum_{\mathbf{Z}\cup\mathbf{PA}} \frac{P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z})}{P(\mathbf{pa}_\pi|\mathbf{i_1}, \mathbf{z})} P(\mathbf{pa}_\pi|\mathbf{i_1}, \mathbf{z})P(\mathbf{pa}_{\bar{\pi}}|\mathbf{i_1}, \mathbf{z})P(y|\mathbf{pa}, \mathbf{z})P(\mathbf{z}) && \text{(b)} \\
&= \sum_{\mathbf{Z}\cup\mathbf{PA}} \frac{P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z})}{P(\mathbf{pa}_\pi|\mathbf{i_1}, \mathbf{z})} P(\mathbf{pa}|\mathbf{i}_1, \mathbf{z})P(y|\mathbf{pa}, \mathbf{z}, \mathbf{i}_1)P(\mathbf{z}) && \text{(c)} \\
&= \sum_{\mathbf{Z}\cup\mathbf{PA}} \frac{P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z})}{P(\mathbf{pa}_\pi|\mathbf{i_1}, \mathbf{z})} P(y, \mathbf{pa}|\mathbf{i}_1, \mathbf{z})P(\mathbf{z}) && \text{(d)} \\
&= \sum_{\mathbf{Z}\cup\mathbf{PA}} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z})P(y, \mathbf{pa}_{\bar{\pi}}|\mathbf{pa}_\pi, \mathbf{i}_1, \mathbf{z})P(\mathbf{z}) && \text{(e)} \\
&= \sum_{\mathbf{Z}} P(\mathbf{z}) \sum_{\mathbf{PA}_\pi} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z}) \sum_{\mathbf{PA}_{\bar{\pi}}} P(y, \mathbf{pa}_{\bar{\pi}}|\mathbf{pa}_\pi, \mathbf{i}_1, \mathbf{z}) && \text{(f)} \\
&= \sum_{\mathbf{Z}} \sum_{\mathbf{PA}_\pi} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z})P(y|\mathbf{pa}_\pi, \mathbf{i}_1, \mathbf{z})P(\mathbf{z}) && \text{(g)} \\
&= \sum_{\mathbf{Z}} \left(\sum_{\mathbf{PA}_\pi} P(\mathbf{pa}_\pi|\mathbf{i_2}, \mathbf{z}, S = 1)P(y|\mathbf{pa}_\pi, \mathbf{i}_1, \mathbf{z}, S = 1) \right) \times P(\mathbf{z}\backslash\mathbf{z}^\intercal|\mathbf{z}^\intercal, S = 1)P(\mathbf{z}^\intercal) && \text{(h)}
\end{aligned}
$$

For Equation a, since $\mathbf{PA} \cup \mathbf{Z}$ is a valid set that d-separates Y and $\mathbf{I}$, based on d-separation criteria [78], and the general product rule in probability, the path-specific causal effect in Equation A.17 can be rewritten by conditioning and marginalizing on $\mathbf{PA} \cup \mathbf{Z}$. Equation b is derived by dividing then multiplying the term $P(\mathbf{pa}_\pi|\mathbf{i_1}, \mathbf{z})$. Equation c is derived by rules of conditional independence and the fact $\mathbf{PA} = \mathbf{PA}_\pi \cup \mathbf{PA}_{\bar{\pi}}$. Equations d, e and f are derived by basic laws of conditional probability. Equation g is derived by marginalizing out the term $\mathbf{pa}_{\bar{\pi}}$ in summation. Equation h is derived by $(\mathbf{Z}, \mathbf{Z}^\intercal)$ satisfies the generalized adjustment criterion in Theorem 2 such that $(\mathbf{Y} \perp\!\!\!\perp S|\mathbf{Z}^\intercal)_{\mathcal{G}^{pbd}_{\mathbf{YI}}}$.

A.7 Proof of Theorem 14

Proof. To derive the regret bound of LinUCB-PCB algorithm, we follow existing research works (e.g., [50, 84]) to make four common assumptions defined as follows:

1. The error term ϵ_t follows 1-sub-Gaussian distribution for each time point.
2. $\{\alpha_t\}_{i=1}^n$ is a non-decreasing sequence with $\alpha_1 \geq 1$.
3. $||\boldsymbol{\theta}^*||_2 < M$ for all time points and arms.
4. There exists a $\delta \in (0,1)$ such that with probability $1-\delta$, for all $t \in [T], \boldsymbol{\theta}^* \in \mathcal{C}_t$ where $\mathcal{C}_t$ satisfies Equation 4.5.

According to the arm selection strategy and OFU principle, the regret at each time t is bounded by:

$$\begin{aligned} reg_t &= \mathbf{x}_{t,a}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_t - \mathbf{x}_{t,a}^{\mathrm{T}}\boldsymbol{\theta}^* \\ &\leq \mathbf{x}_{t,a}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_t + \alpha_t||\mathbf{x}_{t,a}||_{A_t^{-1}} - \mathbf{x}_{t,a}^{\mathrm{T}}\boldsymbol{\theta}^* \\ &\leq \mathbf{x}_{t,a}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_t + \alpha_t||\mathbf{x}_{t,a}||_{A_t^{-1}} - (\mathbf{x}_{t,a}^{\mathrm{T}}\hat{\boldsymbol{\theta}}_{\boldsymbol{t}} - \alpha_t||\mathbf{x}_{t,a}||_{A_t^{-1}}) \\ &\leq 2\alpha_t||\mathbf{x}_{t,a}||_{A_t^{-1}} \end{aligned}$$

Summing up the regret at each bound, with probability at least $1-\delta$ the cumulative regret up to time T is bounded by:

$$R_T = \sum_{t=1}^{T} reg_t \leq \sqrt{T\sum_{t=1}^{T} reg_t^2} \leq 2\alpha_T\sqrt{T\sum_{t=1}^{T}||\mathbf{x}_{t,a}||_{A_t^{-1}}^2} \tag{A.19}$$

Since $\{\alpha_t\}_{i=1}^n$ is a non-decreasing sequence, we can enlarge each element α_t to α_T to obtain the inequalities in Equation A.19. By applying the inequalities from Lemma 2 and 3

we could further relax the regret bound up to time T to:

$$\begin{aligned} R_T &\leq 2\alpha_T \sqrt{2Tlog\frac{|A_t|}{\lambda^d}} \\ &\leq 2\alpha_T \sqrt{2Td(log(\lambda + TL^2/d) - log\lambda)} \\ &= 2\alpha_T \sqrt{2Tdlog(1 + TL^2/(d\lambda))} \end{aligned} \tag{A.20}$$

Following the result of Lemma 1, by loosing the determinant of A_t according to Lemma 3, Lemma 4 provides a suitable choice for α_T up to time T. By plugging in the RHS from Equation 4.5 we derive the cumulative regret bound:

$$R_T \leq \sqrt{2Tdlog(1 + TL^2/(d\lambda))} \times 2(\sqrt{\lambda}M + \sqrt{2log(1/\delta) + dlog(1 + TL^2/(d\lambda))})$$

□

Finally, by plugging in $\delta = 1/T$ we obtain the regret bound of LinUCB-PCB algorithm shown in Theorem 14:

$$R_T \leq Cd\sqrt{T}log(TL)$$

A.8 Proof of Theorem 16

Proof. We first define $\Delta_{max} = \max_{a,c} \Delta_a^c$ and $\mathcal{A}_c^- = \{a : U_{a,c} \geq \mu_c^*\}$. According to Lemma 3.2 in [113] G_t is guaranteed to be invertible since the arm set $\mathcal{A}$ is assumed to span $\mathbb{R}^d$. The regret during the initialization is at most $d\Delta_{max} \approx o(log(T))$. We can thus ignore the regret during the initialization phase in the remaining proof.

To prove the regret during the exploration-exploitation phase, we first define the event

$\mathcal{B}_t$ as follows:

$$\mathcal{B}_t = \left\{ \exists t \geq l, \exists a \in \mathcal{A}, s.t.\ |a^\top \hat{\boldsymbol{\theta}}_t - a^\top \boldsymbol{\theta}| \geq ||a||_{G_t^{-1}} f_n^{1/2} \right\} \tag{A.21}$$

By choosing $\delta = 1/T$, from Lemma A.2 in [113] we have $P(\mathcal{B}_t) \leq 1/T$. We thus decompose the cumulative regret by applying optimal allocation matching (oam) policy with respect to event $\mathcal{B}_t$ as follows:

$$R_{\pi_{\text{oam}}}(T) = \mathbb{E}[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \Delta_a^{c_t} \mathbb{1}(a_t = a)] = \mathbb{E}[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \Delta_a^{c_t} \mathbb{1}(a_t = a, \mathcal{B}_t)] + \mathbb{E}[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \Delta_a^{c_t} \mathbb{1}(a_t = a, \mathcal{B}_t^c)] \tag{A.22}$$

The first term of Equation A.22 could be asymptotically bounded by $o(log(T))$:

$$\begin{aligned} &\limsup_{T \to \infty} \frac{\mathbb{E}[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \Delta_a^{c_t} \mathbb{1}(a_t = a, \mathcal{B}_t)]}{log(T)} \\ =& \limsup_{T \to \infty} \frac{\mathbb{E}[\sum_{t=1}^{T} \Delta_{a_t}^{c_t} \mathbb{1}(\mathcal{B}_t)]}{log(T)} \\ \leq& \limsup_{t \to \infty} \frac{\Delta_{max} \sum_{t=1}^{T} P(\mathcal{B}_t)}{log(T)} \leq \limsup_{T \to \infty} \frac{\Delta_{max} \sum_{t=1}^{T} 1/T}{log(T)} \\ =& \limsup_{T \to \infty} \frac{\Delta_{max}}{log(T)} = 0 \end{aligned} \tag{A.23}$$

To bound the second term in Equation A.22, we further define the event $\mathcal{D}_{t,c_t}$ by

$$\mathcal{D}_{t,c_t} = \left\{ \forall a \in \mathcal{A}, ||a||_{G_t^{-1}}^2 \leq \max\{ \frac{\widehat{\Delta}_{min}^2(t-1)}{f_n}, \frac{(\widehat{\Delta}_a^{c_t}(t-1))^2}{f_n} \} \right\} \tag{A.24}$$

At time t the algorithm exploits under event $\mathcal{D}_{t,c_t}$. Under event $\mathcal{D}_{t,c_t}^c$ the algorithm explores at round t. We then further decompose the second term in Equation A.22 into the

sum of exploitation regret and exploration regret:

$$\begin{aligned}\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t=a,\mathcal{B}_t^c)] = &\underbrace{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t})]}_{\text{exploitation regret}}\\ &+\underbrace{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t}^c)]}_{\text{exploration regret}}\end{aligned} \tag{A.25}$$

We then bound those two terms by Lemma 8 and Lemma 9 accordingly.

Lemma 8. The exploitation regret satisfies

$$\limsup_{T\to\infty}\frac{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t})]}{log(T)}=0 \tag{A.26}$$

Lemma 9. The exploration regret satisfies

$$\limsup_{T\to\infty}\frac{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t}^c)]}{log(T)}\leq\mathcal{V}(\boldsymbol{\theta},\mathcal{A}) \tag{A.27}$$

where $\mathcal{V}(\boldsymbol{\theta},\mathcal{A})$ is defined in Theorem 16.

Combining the bounds of exploitation and exploration regrets leads to the results below:

$$\limsup_{T\to\infty}\frac{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t=a,\mathcal{B}_t^c)]}{log(T)}\leq\mathcal{V}(\boldsymbol{\theta},\mathcal{A}) \tag{A.28}$$

□

Finally, combining the results in Equation A.23 leads to the asymptotic regret bound

of Algorithm 11:

$$R_{\pi_{\text{oam}}}(T) \leq log(T) \cdot \mathcal{V}(\boldsymbol{\theta}, \mathcal{A}) \tag{A.29}$$

A.8.1 Proof of Lemma 8

Proof. Under the event β_t^c defined in Equation A.21, we have

$$\max_{a \in \mathcal{A}} |\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}, a \rangle| \leq ||a||_{G_t^{-1}} f_n^{1/2} \tag{A.30}$$

We further bound $||a||_{G_t^{-1}}$ under event $\mathcal{D}_{t,c_t}$ by:

$$||a||_{G_t^{-1}} \leq \max\left(\frac{\widehat{\Delta}_{min}^2}{f_n}, \frac{\widehat{\Delta}_a^c(t)^2}{f_n}\right) = \frac{(\widehat{\Delta}_a^c(t))^2}{f_n} \tag{A.31}$$

We further define τ_a for each $a \in \mathcal{A}$ as

$$\tau_a = \min\left\{N : \forall t \geq d, \mathcal{D}_t, c_t \text{ occurs}, N_a(t) \geq N, \text{implies} |\langle \hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}, a \rangle| \leq \frac{\Delta_{min}}{2}\right\} \tag{A.32}$$

and decompose the exploitation regret with respect to the event $\{N_{\hat{a}_c^*(t)}(t) \geq \tau_{\hat{a}_c^*(t)}\}$ defined in [113] as follows:

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T} \sum_{a \in \mathcal{A}_{c_t}^-} \Delta_a^{c_t} \mathbb{1}(a_t = a, \mathcal{B}_t^c, \mathcal{D}_{t,c_t})\Big] \\
\leq &\mathbb{E}\Big[\sum_{c=1}^{|\mathcal{C}|} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_c^-} \Delta_a^c \mathbb{1}(a_t = a, \mathcal{B}_t^c, \mathcal{D}_{t,c}, N_{\hat{a}_c^*(t)}(t) \geq \tau_{\hat{a}_c^*(t)})\Big] \\
+ &\mathbb{E}\Big[\sum_{c=1}^{|\mathcal{C}|} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_c^-} \Delta_a^c \mathbb{1}(a_t = a, \mathcal{B}_t^c, \mathcal{D}_{t,c}, N_{\hat{a}_c^*(t)}(t) < \tau_{\hat{a}_c^*(t)})\Big]
\end{aligned} \tag{A.33}$$

When $a_c^* = \hat{a}_c^*(t)$ the first term in Equation A.33 equals 0, we next bound the second term:

$$\begin{aligned}
&\mathbb{E}\Bigg[\sum_{c=1}^{|\mathcal{C}|}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_c^-}\Delta_a^c \mathbb{1}(a_t = a, \mathcal{B}_t^c, \mathcal{D}_{t,c}, N_{\hat{a}_c^*(t)}(t) < \tau_{\hat{a}_c^*(t)})\Bigg] \\
\leq &\mathbb{E}\Bigg[\sum_{c=1}^{|\mathcal{C}|}\sum_{t=1}^{T}\mathbb{1}(a_t = a, \mathcal{B}_t^c, \mathcal{D}_{t,c}, N_{\hat{a}_c^*(t)}(t) < \tau_{\hat{a}_c^*(t)})\Bigg]\Delta_{max} \leq \sum_{c=1}^{|\mathcal{C}|}\sum_{a\in\mathcal{A}}\mathbb{E}[\tau_a]\Delta_{max} \leq \sum_{a\in\mathcal{A}}\mathbb{E}(\tau_a)\Delta_{max}
\end{aligned} \tag{A.34}$$

Combining the results together leads to the desired results:

$$\limsup_{T\to\infty}\frac{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a^{c_t}\mathbb{1}(a_t = a, \mathcal{B}_t^c, \mathcal{D}_{t,c_t})]}{log(T)} \leq \limsup_{T\to\infty}\frac{|\mathcal{A}|\Delta_{max}(8(1+1/log(n)) + 4cdlog(dlog(n)))}{\Delta_{min}^2 log(T)} = 0 \tag{A.35}$$

□

A.8.2 Proof of Lemma 9

Proof. Let $\mathcal{M}_t$ denote the set that records the index of action sets that has not been fully explored until round t:

$$\mathcal{M}_t = \left\{m : \exists a \in \mathcal{A}^m, N_a(t) \leq \min\{f_n/\hat{\Delta}_{min}^2(t), T_a(\hat{\Delta}(t))\}\right\} \tag{A.36}$$

Under the event $\mathcal{D}_{t,c_t}^c$ we have $\mathcal{M}_t \neq \emptyset$. We decompose the exploration regret into two terms: regret under unwasted exploration and wasted exploration, according to whether c_t

belongs to $\mathcal{M}_t$.

$$\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t}^c)]=$$
$$\underbrace{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t}^c,c_t\in\mathcal{M}_t)]}_{\text{unwasted exploration}}+\underbrace{\mathbb{E}[\sum_{t=1}^{T}\sum_{a\in\mathcal{A}}\Delta_a\mathbb{1}(a_t=a,\mathcal{B}_t^c,\mathcal{D}_{t,c_t}^c,c_t\notin\mathcal{M}_t)]}_{\text{wasted exploration}} \tag{A.37}$$

Following the proof procedure of Lemma B.1 and B.2 in [113] by substituting with the reduced arm set $\mathcal{A}_c^- = \{a : U_{a,c} \geq \mu_c^*\}$, we show that the regret regarding the wasted explorations is bounded by $o(log(T))$, and the regret regarding to the unwasted explorations is bounded by $log(T)\cdot\mathcal{V}(\boldsymbol{\theta},\mathcal{A})$. Combing the bounds of these two terms leads to our conclusion. □

A.9 Non-contextual Bandit with Prior Causal Bounds

Our prior causal bounds can also be incorporated into non-contextual bandits. In non-contextual bandit setting, the goal is to calculate $\mathbb{E}[Y|do(\mathbf{X}=\mathbf{x})]$ for each $\mathbf{x}$ and identify the best arm. Recall Equation 7.1 in the main text we have $P_{\mathbf{x}}(y|\mathbf{c}) = \frac{P_{\mathbf{x}}(y,\mathbf{c})}{P_{\mathbf{x}}(\mathbf{c})}$. Simply replacing the outcome variable with y and removing the term $P_{\mathbf{x}}(\mathbf{c})$ from Equation 7.2 and 7.5 in the main text leads to the causal bound for the interventional distribution $p_{\mathbf{x}}(y)$.

A.9.1 Proof of Theorem 17

Proof. We first decompose the cumulative regret up to time T:

$$R(T)=\sum_{a=1}^{k}\Delta_a\mathbb{E}[N_a(T)] \tag{A.38}$$

Let E_a be the event defined by:

$$E_a = \left\{ \mu^* < \min_{t \in [T]} UCB^*(t, \delta) \right\} \cup \left\{ \hat{\mu}_{au_a} + \sqrt{\frac{2}{u_a} log(\frac{1}{\delta})} < \mu^* \right\}$$

where $u_a \in [T]$ is a constant. Since $N_a(T) \leq T$, we have

$$\mathbb{E}[N_a(T)] = \mathbb{E}[\mathbb{1}\{E_a\} N_a(T)] + \mathbb{E}[\mathbb{1}\{E_a^c\} N_a(T)] \leq u_a + P(E_i^c) T \tag{A.39}$$

We will show that if E_a occurs, the number of times arm a is played up to time T is upper bounded (Lemma 10), and the complement event E_a^c occurs with low probability (Lemma 11).

Lemma 10. If E_a occurs, the times that arm a is played is bounded by:

$$\mathbb{E}[N_a(T)] \leq \begin{cases} 0, & if \;\; U_a < \mu^* \\ 3 + \dfrac{16 log(T)}{\Delta_a^2}, & otherwise \end{cases} \tag{A.40}$$

Lemma 11.

$$P(E_a^c) \leq T\delta + \exp\left(- \frac{u_a c^2 \Delta_a^2}{2} \right)$$

Combining the results of the two lemmas we have

$$P(E_i^c) \leq T\delta + \exp\left(- \frac{u_a c^2 \Delta_a^2}{2} \right) \tag{A.41}$$

Then by substituting the result from the two lemmas into Equation A.39, we have

$$\mathbb{E}[N_a(T)] \leq u_a + T\bigg(T\delta + \exp\bigg(-\frac{u_a c^2 \Delta_a^2}{2}\bigg)\bigg) \tag{A.42}$$

We next aim to choose a suitable value for $u_a \in [T]$. Directly solving Equation A.50 and taking the minimum value in the solution space leads to a legit value of u_a:

$$u_a = \left\lceil \frac{2log(1/\delta)}{(1-c)^2 \Delta_a^2} \right\rceil \tag{A.43}$$

Then we take $\delta = 1/n^2$ and u_a with the value in Equation A.43 to get the following equation:

$$\mathbb{E}[N_a(T)] \leq u_a + 1 + T^{1-2c^2/(1-c)^2} = \left\lceil \frac{2log(1/\delta)}{(1-c)^2 \Delta_a^2} \right\rceil + 1 + T^{1-2c^2/(1-c)^2} \tag{A.44}$$

By substituting $c = 1/2$ in the above equation we obtain

$$\mathbb{E}[N_a(T)] \leq 3 + \frac{16log(T)}{\Delta_a^2} \tag{A.45}$$

Finally, substituting $\mathbb{E}[N_a(T)]$ with the bound above for each arm $a \in \mathcal{A}$ in Equation A.38 leads to the desired regret bound:

$$R(T) \leq 3\sum_{a=1}^{k} \Delta_a + \sum_{a:\mathbf{U}_a \geq \mu^*} \frac{16log(T)}{\Delta_a}$$

□

A.9.2 Proof of Lemma 10

Proof. We derive the proof by contradiction. Suppose $N_a(T) > u_a$, there would exist a round $t \in [T]$ where $N_a(t-1) = u_a$ and $a_t = a$. By the definition of E_a we have

$$UCB_a(t-1,\delta) = \hat{u}_a(t-1) + \sqrt{\frac{2log(1/\delta)}{N_a(t-1)}} = \hat{u}_{au_a} + \sqrt{\frac{2log(1/\delta)}{u_a}} < u^* < UCB^*(t-1,\delta) \quad \text{(A.46)}$$

We thus have $a_t = \text{argmax}_i\, UCB_i(t-1,\delta) \neq a$, which leads to a contradiction. As a result, if E_a occurs we have $N_a(T) \leq u_a$.

□

A.9.3 Proof of Lemma 11

Proof. According to the definition E_a^c is defined as:

$$E_a^c = \underbrace{\left\{\mu^* \geq \min_{t\in[T]} UCB^*(t,\delta)\right\}}_{\text{term 1}} \cup \underbrace{\left\{\hat{\mu}_{au_a} + \sqrt{\frac{2log(1/\delta)}{u_a}} \geq \mu^*\right\}}_{\text{term 2}} \quad \text{(A.47)}$$

We decompose term 1 according to the definition of $UCB^*(t,\delta)$:

$$\left\{\mu^* \geq \min_{t\in[T]} UCB^*(t,\delta)\right\} \subset \left\{\mu^* \geq \min_{s\in[T]} \hat{u}_{1s} + \sqrt{\frac{2log(1/\delta)}{s}}\right\} = \bigcup_{s\in[T]} \left\{\mu^* \geq \hat{u}_{1s} + \sqrt{\frac{2log(1/\delta)}{s}}\right\} \quad \text{(A.48)}$$

We next apply corollary 5.5 in [112] and leverage union bound rule of independent random variables to further upper bound term 1 by $n\delta$ as follows:

$$P\left(u^* \geq \min_{t\in T} UCB^*(t,\delta)\right) \leq P\left(\bigcup_{s\in[T]} \left\{\mu^* \geq \hat{u}_{1s} + \sqrt{\frac{2log(1/\delta)}{s}}\right\}\right) \leq \sum_{s=1}^{T} P\left(\mu^* \geq \hat{u}_{1s} + \sqrt{\frac{2log(1/\delta)}{s}}\right) \leq n\delta \quad \text{(A.49)}$$

Next we aim to bound term 2 in Equation A.47. We proceed by assuming u_a is large

enough such that

$$\Delta_a - \sqrt{\frac{2log(1/\delta)}{u_a}} \geq c\Delta_a \tag{A.50}$$

for some constant $c \in (0, 1)$ to be chosen later. Since $u^* = u_a + \Delta_a$, according to corollary 5.5 in [112] we have

$$\begin{aligned} P\left(\hat{\mu}_{au_a} + \sqrt{\frac{2log(1/\delta)}{u_a}} \geq \mu^*\right) &= P\left(\hat{\mu}_{au_a} - u_a \geq \Delta_a - \sqrt{\frac{2log(1/\delta)}{u_a}}\right) \\ &\leq P\left(\hat{\mu}_{au_a} - u_a \geq c\Delta_a\right) \leq \exp(-\frac{u_a c^2 \Delta_a^2}{2}) \end{aligned} \tag{A.51}$$

□