Achieving Counterfactual Fairness for Causal Bandit

Wen Huang, Lu Zhang, Xintao Wu

University of Arkansas
{wenhuang, lz006, xintaowu}@uark.edu

Abstract

In online recommendation, 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. 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.

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 (Celis et al. 2018; Liu et al. 2017; Zhu, Hu, and Caverlee 2018), 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 (Celis et al. 2018; Liu et al. 2017; Zhu, Hu, and Caverlee 2018; Joseph et al. 2016, 2018; Jabbari et al. 2017; Burke 2017; Burke, Sonboli, and Ordonez-Gauger 2018; Ekstrand et al. 2018). Among them, Joseph et al. (2016) 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 (Joseph et al. 2018) for infinite and contextual bandits, (Jabbari et al. 2017) for reinforcement learning, (Liu et al. 2017) 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 (Huang et al. 2020) 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 (Lattimore, Lattimore, and Reid 2016; Sen et al. 2017; Lee and Bareinboim 2018, 2019; Lu et al. 2020), 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, Lu et al. (2020) 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 (Correa and Bareinboim 2020) 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 (Lu et al. 2020) 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 (Huang et al. 2020). The results demonstrate the advantage of our causal based fair bandit algorithm on achieving individual level fairness in online recommendation.

Background

Our work is based on Pearl's structural causal models (Pearl 2009) which describes the causal mechanisms of a system as a set of structural equations.

Definition 1 (Structural Causal Model (SCM) (Pearl 2009))**.** *A causal model $\mathcal{M}$ is a triple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F} \rangle$ where 1) $\mathbf{U}$ is a set of hidden contextual variables that are determined by factors outside the model; 2) $\mathbf{V}$ is a set of observed variables that are determined by variables in $\mathbf{U} \cup \mathbf{V}$; 3) $\mathbf{F}$ is a set of equations mapping from $\mathbf{U} \times \mathbf{V}$ to $\mathbf{V}$. Specifically, for each $V \in \mathbf{V}$, there is an equation $f_V \in \mathbf{F}$ mapping from $\mathbf{U} \times (\mathbf{V} \backslash V)$ to V, i.e., $v = f_V(Pa(V), \mathbf{u}_V)$, where $Pa(V)$ is a realization of a set of observed variables called the parents of V, and $\mathbf{u}_V$ is a realization of a set of hidden variables.*

Quantitatively measuring causal effects is facilitated with the do-operator (Pearl 2009), which simulates the physical interventions that force some variable to take certain values. Formally, the intervention that sets the value of X to x is denoted by $do(x)$. In a SCM, intervention $do(x)$ is defined as the substitution of equation $x = f_X(Pa(X), \mathbf{u}_X)$ with constant $X = x$. For an observed variable Y other than X, its variant under intervention $do(x)$ is denoted by $Y(x)$. The distribution of $Y(x)$, also referred to as the post-intervention distribution of Y, is denoted by $P(Y(x))$. The soft intervention (also known as the conditional action, policy intervention) extends the hard intervention such that it forces variable X to take a new functional relationship in responding to some other variables (Correa and Bareinboim 2020). Denoting the soft intervention by π, the post-interventional distribution of X given its parents is denoted by $P_\pi(X|Pa(X))$. More generally, the new function could receive as inputs the variables other than the original parents $Pa(X)$, as long as they are not the descendants of X. The distribution of Y after performing the soft intervention is denoted by $P(Y(\pi))$.

With intervention, the counterfactual effect measures the causal effect while the intervention is performed conditioning on only certain individuals or groups specified by a subset of observed variables $\mathbf{O} = \mathbf{o}$. Given a context $\mathbf{O} = \mathbf{o}$, the counterfactual effect of the value change of X from x_1 to x_2 on Y is given by $\mathbb{E}[Y(x_2)|\mathbf{o}] - \mathbb{E}[Y(x_1)|\mathbf{o}]$.

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 (Spirtes, Glymour, and Scheines 2000) connects the causal graph with conditional independence.

Definition 2 (d-Separation (Spirtes, Glymour, and Scheines 2000))**.** *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 $\mathbf{W}$.*

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 $\mathbf{x}_{t,a}$ 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.

Modeling Arm Selection via Soft Intervention

In bandit algorithms, we often choose an arm that maximizes the expectation of the conditional reward, $a_t = \arg\max_a \mathbb{E}[R|\mathbf{x}_{t,a}]$. 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.

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 1, where $\mathbf{A}$ represents the arm features, $\mathbf{X}$ represents the user features,

[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.

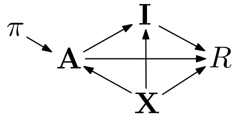

Figure 1: Graph structure for contextual bandit. Node π denotes the soft intervention conducted on arm selection.

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 (Correa and Bareinboim 2020), it can be further decomposed as follows:

$$\begin{aligned} \mu_\pi = \mathbb{E}[R(\pi)|\mathbf{x}_t] &= \sum_{\mathbf{a}} P_\pi(\mathbf{a}|\mathbf{x}_t) \cdot \mathbb{E}[R(\mathbf{a})|\mathbf{x}_t] \\ &= \mathbb{E}_{\mathbf{a}\sim\pi}\left[\mathbb{E}[R(\mathbf{a})|\mathbf{x}_t]\right] \end{aligned} \tag{1}$$

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.

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 (Pearl 2009), we can decompose μ_a as follows.

$$\begin{aligned} \mu_a = \mathbb{E}[R|do(\mathbf{a}), \mathbf{x}_t] &= \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{z}, do(\mathbf{a}), \mathbf{x}_t] P(\mathbf{z}|do(\mathbf{a}), \mathbf{x}_t) \\ &= \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{z}, \mathbf{a}, \mathbf{x}_t] P(\mathbf{z}|\mathbf{a}, \mathbf{x}_t) = \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{z}, \mathbf{a}, \mathbf{x}_t] P(\mathbf{z}|\mathbf{x}_{t,a}) \\ &= \sum_{\mathbf{Z}} \mathbb{E}[R|\mathbf{w}] P(\mathbf{z}|\mathbf{x}_{t,a}) \end{aligned} \tag{2}$$

where the last step is due to the d-separation. Similar to (Lu et al. 2020), 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

$$\hat{\mu}_\pi(t) = \mathbb{E}_{\mathbf{a}\sim\pi}\left[\sum_{\mathbf{Z}} \hat{\mu}_{\mathbf{w}}(t) \cdot P(\mathbf{z}|\mathbf{x}_{t,a})\right] \tag{3}$$

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

$$\pi_t = \arg\max_{\pi \in \Pi_t} \mathbb{E}_{\mathbf{a}\sim\pi}[UCB_a(t)] \tag{4}$$

$$UCB_a(t) = \sum_{\mathbf{Z}} UCB_{\mathbf{w}}(t) P(\mathbf{z}|\mathbf{x}_{t,a}) \tag{5}$$

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

$$UCB_{\mathbf{w}}(t) = \hat{\mu}_{\mathbf{w}}(t) + \sqrt{\frac{2\log(1/\delta)}{1 \vee N_{\mathbf{w}}(t)}} \tag{6}$$

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

$$\hat{\mu}_{\mathbf{w}}(t) = \frac{1}{T_{\mathbf{w}}(t)} \sum_{k=1}^{t} r_t \mathbb{1}_{\mathbf{w}_k = \mathbf{w}} \quad \text{where} \quad T_{\mathbf{w}}(t) = \sum_{k=1}^{t} \mathbb{1}_{\mathbf{w}_k = \mathbf{w}} \tag{7}$$

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

Algorithm 1: D-UCB: Causal Bandit based on d-separation

```
1: Input: Policy space $\Pi$, confidence level parameter $\delta$, 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 $\pi_t$ following Eq. (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. (7).
7: end for
```

Theorem 1 (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. [2] The proof of Theorem 1 follows the general regret analysis framework of the UCB algorithm (Auer, Cesa-Bianchi, and Fischer 2002). 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. □

Algorithm 1 shows the pseudo code of the D-UCB. In Line 2, according to Theorem 1, 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 $\pi_t = \arg\max_{\pi \in \Pi_t} \mathbb{E}_{\mathbf{a}\sim\pi}[UCB_a(t)]$. 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 (Geiger, Verma, and Pearl 1990). We leverage the algorithm of finding a minimum cost separator (Tian, Paz, and Pearl 1998) 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 $P(\mathbf{z}|\mathbf{x_{t,a}})$ 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 (Lu, Meisami, and Tewari 2021) tries 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.

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 3. *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 unidentifiable from observational data without the knowledge of structure equations (Shpitser and Pearl 2008). In (Wu, Zhang, and Wu 2019), 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 (Wu, Zhang, and Wu 2019), our counterfactual fairness is identifiable if $\mathbf{X}\backslash\{S\}$ are not descendants of S. In this case, similar to Eq. (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. (2), we denote $\mu_{a,s^*} = \mathbb{E}[R(a, s^*)|\mathbf{x}_t]$, which can be

[2] Due to space limits, we only include proof sketches. Refer to the appendix in (Huang, Zhang, and Wu 2021) for proof details of all theorems.

decomposed using the do-calculus as

$$\begin{aligned}\mu_{a,s^*} &= \mathbb{E}[R(a,s^*)|\mathbf{x}_t] \\ &= \sum_{\mathbf{Z}} \mathbb{E}[R|s^*, \mathbf{w}\backslash s_t] \cdot P(\mathbf{z}|s^*, \mathbf{x}_{t,a}\backslash s_t) \qquad (8)\end{aligned}$$

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

$$\hat{\Delta}_\pi(t) = \left|\mathbb{E}_{\mathbf{a}\sim\pi}[\hat{\mu}_{a,s^+}(t)] - \mathbb{E}_{\mathbf{a}\sim\pi}[\hat{\mu}_{a,s^-}(t)]\right| \qquad (9)$$

In the case where μ_{a,s^*} is not identifiable, based on Proposition 2 in (Wu, Zhang, and Wu 2019) we derive the lower and upper bounds of μ_{a,s^*} as presented in the following theorem.

Theorem 2. *Given a causal graph as shown in Figure 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*

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

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

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

$$UCB_{\Delta_\pi}(t) = \hat{\Delta}_\pi(t) + \sum_{\mathbf{Z}} \sqrt{\frac{8\log(1/\delta)}{1 \vee N_{\mathbf{w}}(t)}} P(\mathbf{z}|\mathbf{x}_{t,a}) \qquad (10)$$

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:

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

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. The Assumption 3 in Appendix gives the definition of a safe policy π_0, which refers to 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 2 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 $\pi_t = \arg\max_{\pi\in\bar{\Phi}_t} \mathbb{E}_{\mathbf{a}\sim\pi}[UCB_a(t)]$.

Algorithm 2: F-UCB: Fair Causal Bandit

```
1:  Input: Policy space $\Pi$, fairness threshold $\tau$, confidence level
      parameter $\delta$, 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. (3) and the
          estimated fairness discrepancy using Eq. (9).
6:     end for
7:     Determine the conservative fair policy subspace $\bar{\Phi}_t$.
8:     Find the optimal policy following Eq. (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 3 (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:*

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

Proof Sketch. Our derivation of the regret upper bound of F-UCB follows the proof idea of bandits with linear constraints (Pacchiano et al. 2021), 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. □

Remark. In Theorem 3, 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. In the appendix of (Huang, Zhang, and Wu 2021) we show the numerical relation α_c and α_r should satisfy in order to synchronously bound the uncertainty caused by the error terms. The values taken in Theorem 3 is one feasible solution with α_c taking the minimum value under the constraint domain space.

The general framework we proposed (Eq. (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 paper 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.

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.

Email Campaign Dataset

We adopt the Email Campaign data as used in previous works (Lu et al. 2020). 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. Figure 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 our experiment, we set $\delta = 1/t^2$ for each $t \in [T]$. In the appendix of (Huang, Zhang, and Wu 2021) we show the domain values of all 11 attributes and their conditional probability tables.

Figure 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 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 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 subspace 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.

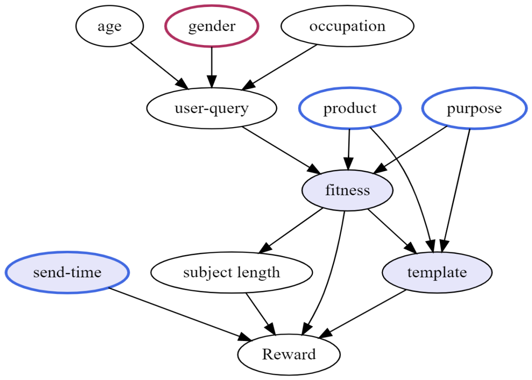

Figure 2: Graph structure under Email Campaign Data. Nodes with blue frame denote the variables that can be intervened. The node with red frame is the sensitive attribute. Light shaded nodes denote the minimal d-separation set.

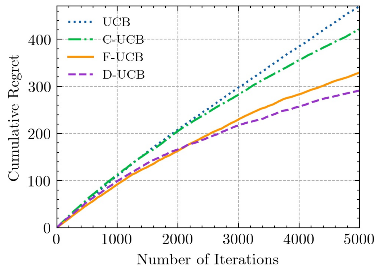

Figure 3: Comparison of bandit algorithms ($\tau = 0.3$ for F-UCB).

Table 1 shows how the cumulative regret of F-UCB ($T = 5000$ rounds) varies with the fairness threshold τ. The values in Table 1 (and Table 2) are obtained by averaging the results over 5 trials. The larger the τ, the smaller the cumulative regret. In the right block of Table 1, 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.

Adult-Video Dataset

We further compare the performance of F-UCB algorithm with Fair-LinUCB (Huang et al. 2020) on Adult-Video dataset. We follow the settings of (Huang et al. 2020) by combining two publicly available datasets: Adult dataset

Table 1: 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

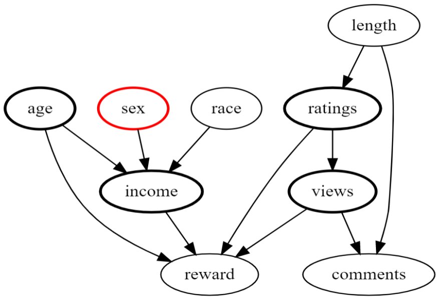

Figure 4: Graph structure for Adult-Video Data

Table 2: 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

and Youtube video dataset. We include in the appendix of (Huang, Zhang, and Wu 2021) detailed information about datasets and experiment. 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 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 set $\delta = 1/t^2$ for each $t \in [T]$. The cumulative regret is added up through 5000 rounds.

We observe from Table 2 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 (Huang et al. 2020) in our experiment and Lin-UCB based algorithm will better catch the reward distribution under this setting.

Related Work

Causal Bandits. There have been a few research works of studying how to learn optimal interventions sequentially by representing the relationship between interventions and outcomes as a causal graph along with associated conditional distributions. Lattimore, Lattimore, and Reid (2016) introduced the causal bandit problems in which interventions are treated as arms in a bandit problem but their influence on the reward, along with any other observations, is assumed to conform to a known causal graph. Specifically they focus on the setting that observations are only revealed after selecting an intervention (and hence the observed features cannot be used as context) and the distribution of the parents of the reward is known under those interventions. Lee and Bareinboim (2018) developed a way to choose an intervention subset based on the causal graph structure as a brute-force way to apply standard bandit algorithms on all interventions can suffer huge regret. Lee and Bareinboim (2019) studied a relaxed version of the structural causal bandit problem when not all variables are manipulable. Sen et al. (2017) considered best intervention identification via importance sampling. Instead of forcing a node to take a specific value, they adopted soft intervention that changes the conditional distribution of a node given its parent nodes. Lu et al. (2020) proposed two algorithms, causal upper confidence bound (C-UCB) and causal Thompson Sampling (C-TS), and showed that they have improved cumulative regret bounds compared with algorithms that do not use causal information. They focus on causal relations among interventions and use causal graphs to capture the dependence among reward distribution of these interventions.

Fair Machine Learning. Fairness in machine learning has been a research subject with rapid growth and attention recently. Related but different from our work include long term fairness (e.g., (Liu et al. 2018)), 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 (Bower et al. 2017; Emelianov et al. 2019; Dwork and Ilvento 2019; Dwork, Ilvento, and Jagadeesan 2020), which primarily consider the combination of multiple non-adaptive sequential decisions and evaluate fairness at the end of the pipeline, and fair sequential learning (Joseph et al. 2016), which sequentially considers each individual and makes decision for them. Liu et al. (2018) 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. Hu and Rangwala (2020) 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.

Conclusions

In our paper, 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.

Acknowledgments

This work was supported in part by NSF 1910284, 1920920, 1940093, and 1946391.

Ethics Statement

Our research could benefit online recommendation system providers so that they can make fair recommendations while still achieving low regret. Our research could also benefit users of online recommendation systems as we achieve counterfactual fairness. Counterfactual fairness is a causality-based user-side fairness notion. Our research could prevent users from receiving biased recommendations especially for those from disadvantage groups.

References

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3): 235–256.

Bower, A.; Kitchen, S. N.; Niss, L.; Strauss, M. J.; Vargas, A.; and Venkatasubramanian, S. 2017. Fair Pipelines. *CoRR*, abs/1707.00391.

Burke, R. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093*.

Burke, R.; Sonboli, N.; and Ordonez-Gauger, A. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *FaccT'18*.

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

Correa, J. D.; and Bareinboim, E. 2020. A Calculus for Stochastic Interventions: Causal Effect Identification and Surrogate Experiments. In *AAAI'20*, 10093–10100.

Dwork, C.; and Ilvento, C. 2019. Fairness Under Composition. In *ITCS'19*, volume 124 of *LIPIcs*, 33:1–33:20.

Dwork, C.; Ilvento, C.; and Jagadeesan, M. 2020. Individual Fairness in Pipelines. In *FORC'20*, 7:1–7:22.

Ekstrand, M. D.; Tian, M.; Kazi, M. R. I.; Mehrpouyan, H.; and Kluver, D. 2018. Exploring author gender in book rating and recommendation. In *RecSys'18*, 242–250.

Emelianov, V.; Arvanitakis, G.; Gast, N.; Gummadi, K. P.; and Loiseau, P. 2019. The Price of Local Fairness in Multistage Selection. In *IJCAI'19*, 5836–5842.

Geiger, D.; Verma, T.; and Pearl, J. 1990. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*.

Hu, Q.; and Rangwala, H. 2020. Metric-Free Individual Fairness with Cooperative Contextual Bandits. *CoRR*, abs/2011.06738.

Huang, W.; Labille, K.; Wu, X.; Lee, D.; and Heffernan, N. 2020. Achieving User-Side Fairness in Contextual Bandits. *arXiv preprint arXiv:2010.12102*.

Huang, W.; Zhang, L.; and Wu, X. 2021. Achieving Counterfactual Fairness for Causal Bandit. *CoRR*, abs/2109.10458.

Jabbari, S.; Joseph, M.; Kearns, M. J.; Morgenstern, J.; and Roth, A. 2017. Fairness in Reinforcement Learning. In *ICML'17*.

Joseph, M.; Kearns, M. J.; Morgenstern, J.; Neel, S.; and Roth, A. 2018. Meritocratic Fairness for Infinite and Contextual Bandits. In *AIES'18*, 158–163. ACM.

Joseph, M.; Kearns, M. J.; Morgenstern, J. H.; and Roth, A. 2016. Fairness in Learning: Classic and Contextual Bandits. In *NeurIPS*.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal Bandits: Learning Good Interventions via Causal Inference. In *NeurIPS'16*.

Lee, S.; and Bareinboim, E. 2018. Structural Causal Bandits: Where to Intervene? In *NeurIPS'18*, 2573–2583.

Lee, S.; and Bareinboim, E. 2019. Structural Causal Bandits with Non-Manipulable Variables. In *AAAI'19*.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed Impact of Fair Machine Learning. In *ICML'18*.

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

Lu, Y.; Meisami, A.; and Tewari, A. 2021. Causal Bandits with Unknown Graph Structure. *arXiv preprint arXiv:2106.02988*.

Lu, Y.; Meisami, A.; Tewari, A.; and Yan, W. 2020. Regret Analysis of Bandit Problems with Causal Background Knowledge. In *UAI'20*, 141–150.

Pacchiano, A.; Ghavamzadeh, M.; Bartlett, P.; and Jiang, H. 2021. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, 2827–2835. PMLR.

Pearl, J. 2009. *Causality*. Cambridge university press.

Sen, R.; Shanmugam, K.; Dimakis, A. G.; and Shakkottai, S. 2017. Identifying Best Interventions through Online Importance Sampling. In *ICML'17*, 3057–3066.

Shpitser, I.; and Pearl, J. 2008. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9: 1941–1979.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*, volume 81. MIT press.

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

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *IJCAI'19*, 1438–1444.

Zhu, Z.; Hu, X.; and Caverlee, J. 2018. Fairness-aware tensor-based recommendation. In *CIKM'18*, 1153–1162.