Robustly Improving Bandit Algorithms with Confounded and Selection Biased Offline Data: A Causal Approach

Wen Huang, Xintao Wu

University of Arkansas

{wenhuang, xintaowu}@uark.edu

Abstract

This paper studies bandit problems where an agent has access to offline data that might be utilized to potentially improve the estimation of each arm's reward distribution. A major obstacle in this setting is the existence of compound biases from the observational data. Ignoring these biases and blindly fitting a model with the biased data could even negatively affect the online learning phase. In this work, we formulate this problem from a causal perspective. First, we categorize the biases into confounding bias and selection bias based on the causal structure they imply. Next, we 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 show that prior causal bounds could help consistently reduce the asymptotic regret.

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 (Li et al. 2010), online advertising (Tang et al. 2013; Avadhanula et al. 2021), and anomaly detection (Ding, Li, and Liu 2019). 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 (Chen et al. 2023). 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 (Pearl 2009) 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 (Zhang and Bareinboim 2017; Sharma et al. 2020; Tennenholtz et al. 2021) only focused on one specific bias and have not dealt with compound biases in the offline data. It was shown in (Bareinboim, Tian, and Pearl 2014) 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 (Correa and Bareinboim 2017) or the accessibility of external unbiasedly measured data (Bareinboim and Tian 2015). In this paper, we address this limitation by non-parametrically bounding the target conditional causal effect when confounding and selection biases can not be mitigated 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 one contextual bandit algorithm (LinUCB-PCB) and one non-contextual bandit algorithm (UCB-PCB) that are enhanced with prior causal bounds. We theoretically show under mild conditions both 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.

Background

Our work is based on Pearl's structural causal model (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 $v = f_V(Pa(V), \mathbf{u}_V)$ 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 variables 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 $X = x$. The causal model $\mathcal{M}$ is associated with a causal graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$. 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. 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 the outcome variables $P(\mathbf{y}|do(\mathbf{x}))$ can be computed by the truncated factorization formula (Pearl 2009),

$$P(\mathbf{y}|do(\mathbf{x})) = \prod_{Y \in \mathbf{Y}} P(y|Pa(Y))\delta_{\mathbf{X}=\mathbf{x}}, \tag{1}$$

where $\delta_{\mathbf{X}=\mathbf{x}}$ means assigning attributes in $\mathbf{X}$ involved in the term ahead with the corresponding values in $\mathbf{x}$. The post intervention distribution $P(\mathbf{y}|do(\mathbf{x}))$ is identifiable if it can be uniquely computed from observational distributions $P(\mathbf{V})$.

Confounding Bias occurs when there exist hidden variables that simultaneously determine exposure variables 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. (Shpitser, VanderWeele, and Robins 2012) further generalized the backdoor criterion to identify the causal effect $P(\mathbf{y}|do(\mathbf{x}))$ if all non-proper causal paths are blocked.

Definition 2 (Generalized Backdoor Criterion (Shpitser, VanderWeele, and Robins 2012))**.** *A set of variables $\mathbf{Z}$ satisfies the adjustment criterion relative to $(\mathbf{X}, \mathbf{Y})$ in $\mathcal{G}$ if: (i) no element in $\mathbf{Z}$ is a descendant in $\mathcal{G}_{\overline{\mathbf{X}}}$ of any $W \notin \mathbf{X}$ lying on a proper causal path from $\mathbf{X}$ to Y. (ii) all non-causal paths in $\mathcal{G}$ from $\mathbf{X}$ to Y are blocked by $\mathbf{Z}$.*

In Definition 2 $\mathcal{G}_{\overline{\mathbf{X}}}$ denotes the graph resulting from removing all incoming edges to $\mathbf{X}$ in $\mathcal{G}$, and a causal path from a node in $\mathbf{X}$ to $\mathbf{Y}$ is called proper if it does not intersect $\mathbf{X}$ except at the starting point. The causal effect can thus be computed by controlling for a set of covariates $\mathbf{Z}$.

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} P(\mathbf{y}|\mathbf{x}, \mathbf{z})P(\mathbf{z}) \tag{2}$$

Sample Selection Bias arises with a biased selection mechanism, e.g., choosing users based on a certain time or location. To accommodate for SCM framework, we introduce a node S in a causal graph representing a binary indicator of entry into the observed data, and denote the causal graph augmented with selection node S as $\mathcal{G}_s$. Generally speaking, the target distribution $P(\mathbf{y}|do(\mathbf{x}))$ is called s-recoverable if it can be computed from the available (biased) observational distributions $P(\mathbf{V}|S = 1)$ in the augmented graph $\mathcal{G}_s$. To recover causal effects in the presence of confounding and sample selection bias, (Correa, Tian, and Bareinboim 2018) studied the use of generalized adjustment criteria and introduced a sufficient and necessary condition for recovering causal effects from biased distributions.

Theorem 1 (Generalized Adjustment for Causal Effect (Correa, Tian, and Bareinboim 2018))**.** *Given a causal diagram $\mathcal{G}$ augmented with selection variable S, disjoint sets of variables $\mathbf{Y}, \mathbf{X}, \mathbf{Z}$, for every model compatible with $\mathcal{G}$, we have*

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathbf{Z}} P(\mathbf{y}|\mathbf{x}, \mathbf{z}, S = 1)P(\mathbf{z}|S = 1) \tag{3}$$

if and only if the adjustment variable set $\mathbf{Z}$ satisfies the four criterion shown in (Correa, Tian, and Bareinboim 2018).

Instead of identifying causal effect in presence of selection bias by adjustment, (Correa, Tian, and Bareinboim 2019) proposed a parallel procedure to justify whether a causal quantity is identifiable and recoverable from selection bias using axiomatical c-components factorization (Tian and Pearl 2002). However, both techniques require strong graphical condition to obtain the unbiased estimation of the true conditional causal effect when both confounding and selection biases exist.

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 bi-directed edge indicates there exists unobserved confounder(s) between the two connected nodes. 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 2 (C-component Factorization (Tian and Pearl 2002))**.** *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:*

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\boldsymbol{C}\backslash\mathbf{Y}} Q[\boldsymbol{C}] = \sum_{\boldsymbol{C}\backslash\mathbf{Y}} \prod_{i=1}^{l} Q[\boldsymbol{C}_i] \tag{4}$$

where $\mathbf{X}, \mathbf{Y} \subset \mathbf{V}$ could be arbitrary sets, $\boldsymbol{C} = An(\mathbf{Y})_{\mathcal{G}_{\mathbf{V}\backslash\mathbf{X}}}$ denotes the ancestor node set of $\mathbf{Y}$ in sub-graph $\mathcal{G}_{\mathbf{V}\backslash\mathbf{X}}$, and $\boldsymbol{C}_1, ..., \boldsymbol{C}_l$ are the c-components of $\mathcal{G}_{\boldsymbol{C}}$.

(Bareinboim, Tian, and Pearl 2014) showed that $P(\mathbf{y}|do(\mathbf{x}))$ is recoverable and could be computed by Equation 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.

Related Works

Causal Inference under Confounding and Selection Biases. (Bareinboim, Tian, and Pearl 2014) firstly studied the use of covariate adjustment for simultaneously dealing with both confounding and selection biases based on the SCM. (Correa and Bareinboim 2017) developed a set of complete conditions to recover causal effects in two cases: when none of the covariates are measured externally, and when all of them are measured without selection bias. (Correa, Tian, and Bareinboim 2018) further studied a general case when only a subset of the covariates require external measurement. They developed adjustment-based technique that 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 recovering causal effects under certain graph conditions, our work focuses on bounding causal effects under compound biases, which is needed in various application domains.

Combining Offline Evaluation and Online Learning in Bandit Setting. Recently there are research works that focus on confounding issue in bandit setting (Bareinboim, Forney, and Pearl 2015; Tennenholtz et al. 2021). It is shown in (Bareinboim, Forney, and Pearl 2015) that in MAB problems, neglecting unobserved confounders will lead to a suboptimal arm selection strategy. They also demonstrated that one can not 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. (Tennenholtz et al. 2021) studied a linear bandit problem where the agent is provided with partially observed offline

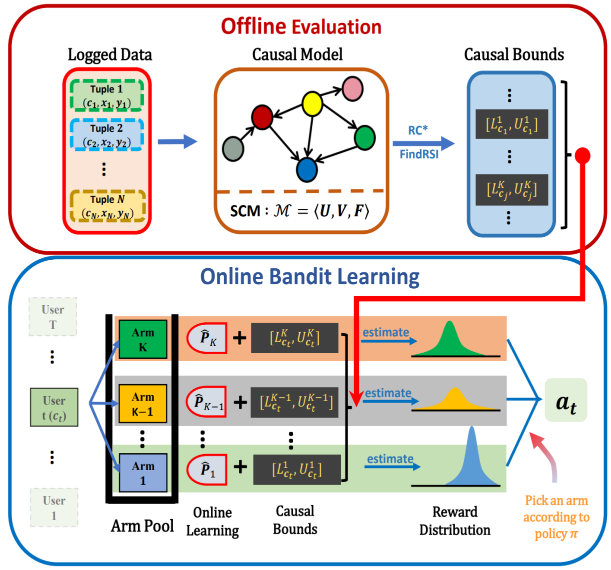

Figure 1: An illustration graph of our proposed framework.

data. (Zhang and Bareinboim 2017, 2021) derived causal bounds based on structural causal model and used them to guide arm selection in online bandit algorithms. (Sharma et al. 2020) 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. (Li et al. 2021; Tang and Xie 2021) 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 regret cost. However, none of those proposed algorithms take selection bias and confounding bias simultaneously into consideration during offline evaluation phase.

Algorithm Framework

An overview of our framework is illustrated in Figure 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 1) 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.

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 $\mathbf{x}_{a_t}$ based on previous observations. The agent then receives the reward Y_t and observes values for all non-intervened variables. We define the the expected mean reward of pulling an arm a with with feature value $\mathbf{x}_a$ given user context $\mathbf{c}$ as the conditional causal effect shown below:

$$u_{a,\mathbf{c}} = \mathbb{E}[Y|do(\mathbf{X} = \mathbf{x}_a), \mathbf{c}] \tag{5}$$

When the offline data are available, we can leverage them as prior estimators of the reward for each arm 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.

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 (Tian 2004) provides a complete procedure to compute conditional causal effects using observational distributions.

$$P(Y = y|do(\mathbf{x}), \mathbf{c}) = P_{\mathbf{x}}(y|\mathbf{c}) = \frac{P_{\mathbf{x}}(y, \mathbf{c})}{P_{\mathbf{x}}(\mathbf{c})} \tag{6}$$

where $P_{\mathbf{x}}(y|\mathbf{c})$ is the abbreviated form of the conditional causal effect $P(Y = y|do(\mathbf{x}), \mathbf{c})$. (Tian 2004) 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 6 is identifiable **if and only if** $P_{\mathbf{x}}(y, \mathbf{c})$ is identifiable. However, the cond-identify algorithm (Tian 2004) is not applicable for the scenario with the presence of selection bias.

Algorithm 1 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 1 decompose the target causal effect following c-component factorization and recursively call our RC* algorithm (shown in Algorithm 2) 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. We also include discussions regarding the assumptions on causal graph in (Huang and Wu 2023).

Algorithm 1: 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 $P_{\mathbf{x}}(y, \mathbf{c}) = \sum_{\mathbf{D}\setminus\{\mathbf{Y},\mathbf{C}\}} \prod_{i=1}^{l} Q[\mathbf{D}_i]$ following Equation 4.
7: **for** each $\mathbf{D}_i$ **do**
8: $\quad L_{Q(\mathbf{D}_i)}, U_{Q(\mathbf{D}_i)} = \text{RC*}(\mathbf{D}_i, P(\mathbf{v}|S = 1), \mathcal{G})$
9: **end for**
10: Update L_q, U_q according to Theorem 3.
11: *// Substitute Intervention*
12: $\mathcal{D} = \text{FindRSI}(\mathbf{x}, \mathbf{c}, y, \mathcal{G})$
13: **if** $\mathcal{D} \neq \phi$ **then**
14: $\quad$ Update L_w, U_w according to Theorem 4.
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** $L_{\mathbf{x},\mathbf{c}} = \max\{\hat{L}_q, \hat{L}_w\}, U_{\mathbf{x},\mathbf{c}} = \min\{\hat{U}_q, \hat{U}_w\}$

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 2 is designed based on the RC algorithm in (Correa, Tian, and Bareinboim 2019) to accommodate 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 2 marginalize out the non-ancestors of $\mathbf{E} \cup S$ since they do not affect the recoverability results. From Lemma 3 in (Bareinboim and Tian 2015), each c-component in line 7 is recoverable since none of them contains ancestors of S. Line 13 calls the Identify function proposed by (Tian and Pearl 2003) 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 $Q[\mathbf{E}]$ can not be computed from the biased observational data in theory. These situations are referred to as nonrecoverable situations. We address this nonrecoverable challenge by nonparametrically 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 4.

Note that in line 9 of RC* algorithm, we bound the target c-component by $[0, 1]$ since under semi-Markovian models

Algorithm 2: 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*$(\mathbf{E}, \frac{P}{\prod_i Q[\boldsymbol{C}_i]}, \mathcal{G}_{\mathbf{V}\backslash \boldsymbol{C}})$

it is challenging to find a tight bound for $Q[\mathbf{E}]$ when $\boldsymbol{C} = \emptyset$. One future direction is to further apply a non-parametric bounding technique similar to (Wu, Zhang, and Wu 2019). That is, choosing certain probability distributions in the truncated formula that are the source of unrecoverability, and set specific domain values for a carefully chosen variable set to allow these distributions achieve their maximum/minimum values. Finally we list the causal bounds derived from calling RC* algorithm in the follow Theorem:

Theorem 3 (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:*

$$\begin{aligned} L_q &= \sum_{\mathbf{D}\backslash\{\mathbf{Y},\mathbf{C}\}} \prod_{i=1}^{l} L_{Q[D_i]}/P_{\mathbf{x}}(\mathbf{c}) \\ U_q &= \sum_{\mathbf{D}\backslash\{\mathbf{Y},\mathbf{C}\}} \prod_{i=1}^{l} U_{Q[D_i]}/P_{\mathbf{x}}(\mathbf{c}) \end{aligned} \tag{7}$$

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

$$\begin{aligned} P_{\mathbf{x}}(y, \mathbf{c}) &\leq \max_{\mathbf{w}^*\in\mathbf{W}} P_{\mathbf{x}}(y, \mathbf{c}|\mathbf{w}^*) \\ P_{\mathbf{x}}(y, \mathbf{c}) &\geq \min_{\mathbf{w}^*\in\mathbf{W}} P_{\mathbf{x}}(y, \mathbf{c}|\mathbf{w}^*) \end{aligned} \tag{8}$$

We then investigate whether the action/observation exchange rule of do-calculus (Pearl 2009) and the corresponding graph conditions could be extended in the presence of selection bias and list the results in the following lemma.

Lemma 1 (Action/Observation Rule under Selection Bias)**.** *If the graphical condition $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}, S|\mathbf{X}, \mathbf{W})_{\mathcal{G}_{\overline{\mathbf{X}}\overline{\mathbf{Z}(\mathbf{w})}}}$ is satisfied in $\mathcal{G}$, the following equivalence between two post-interventional distributions holds:*

$$P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{w}), \mathbf{z}, S = 1) = P(y|do(\mathbf{x}), \mathbf{W}, \mathbf{z}, S = 1) \tag{9}$$

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. $\mathbf{Z}(\mathbf{W})$ 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 9, 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 4.

Theorem 4 (Causal Bound with Substitute Interventions)**.** *Given a set of variables corresponding to recoverable substitute interventions: $\mathcal{D} = \{\mathbf{W}|P_{\mathbf{x},\mathbf{w}}(y, \mathbf{c})$ is recoverable$\}$, the target conditional intervention $P_{\mathbf{x}}(y|\mathbf{c})$ is bounded by*

$$\begin{aligned} L_w &= \max_{\mathbf{W}\in\mathcal{D}} \min_{\mathbf{w}^*\in\mathbf{W}} P_{\mathbf{x},\mathbf{w}^*}(y, \mathbf{c})/P_{\mathbf{x}}(\mathbf{c}) \\ U_w &= \min_{\mathbf{W}\in\mathcal{D}} \max_{\mathbf{w}^*\in\mathbf{W}} P_{\mathbf{x},\mathbf{w}^*}(y, \mathbf{c})/P_{\mathbf{x}}(\mathbf{c}) \end{aligned} \tag{10}$$

We list our procedure of finding all the recoverable substitute interventions in Algorithm 3. Basically the main function FindRSI in Algorithm 3 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 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 dashed node C_1 denotes the confounder that affects both I_1 and Y simultaneously. To bound the conditional causal

Algorithm 3: 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$\}$.
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 1 **then**
    $\mathcal{D} = \mathcal{D} \cup \{\mathbf{W}\}$
  end if
end for

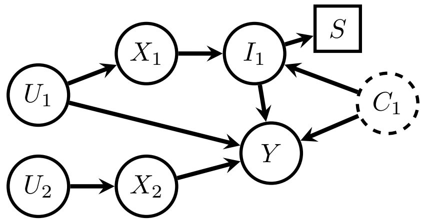

Figure 2: Causal graph for synthetic data.

effect $p_{x_1,x_2}(y|u_1,u_2)$ via c-component factorization, we first identify the set $\mathbf{D} = An(\mathbf{Y})_{\mathcal{G}_{\mathbf{V}\setminus\mathbf{X}}} = \{Y, U_1, U_2\}$. The target intervention could be expressed as

$$\begin{aligned} p_{x_1,x_2}(y|u_1,u_2) &= p_{x_1,x_2}(y,u_1,u_2)/p(u_1,u_2) \\ &= (Q[Y]\cdot Q[U_1]\cdot Q[U_2])/p(u_1,u_2) \end{aligned} \tag{11}$$

We then call RC* algorithm to bound each c-component and return the bound for each arm according to Theorem 3. For bounding causal effects via substitute interventions, we call FindRSI to find a valid variable set $\mathcal{D} = \{I_1\}$. According to Theorem 4, we can obtain the bound for each arm.

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. We define the concatenation of user and arm feature as $\mathbf{x}_{\mathbf{c}_t,a} = [\mathbf{c}_t, \mathbf{x}_a] \in \mathbb{R}^d$ when the agent picks arm a based on user profile $\mathbf{c}_t$ at time $t \in [T]$ and use its simplified notion $\mathbf{x}_{t,a}$ to be consistent with previous work like LinUCB. Under the linear assumption, the binary reward is generated by a Bernoulli distribution $Y_a \sim Ber(\langle \boldsymbol{\theta}, \mathbf{x}_{t,a}\rangle)$ parameterized by $\boldsymbol{\theta}$. Let $a_t = \arg\max_{a\in\mathcal{A}} \mathbb{E}[Y_a]$, the expected cumulative regret up to time T is defined as:

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

At each round the agent pulls an arm based on the user context, observes the reward, and aims to minimize the expected cumulative regret $\mathbb{E}[R(T)]$ after T rounds. We next conduct regret analysis and show our strategy could consistently reduce the long-term regret with the guide of a prior causal bound for each arm's reward distribution.

LinUCB Algorithm with Prior Causal Bounds

LinUCB (Chu et al. 2011) 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 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 4 shows the pseudo-code of our LinUCB-PCB algorithm. The truncated upper confidence bound shown in line 10 of Algorithm 4 contains strong prior information about the true reward distribution implied by the prior causal bound, thus leading to a lower asymptotic regret bound. We include the proof details of Theorems 5 and 6 in (Huang and Wu 2023).

Algorithm 4: LinUCB algorithm with Prior Causal Bounds (LinUCB-PCB)

1: **Input:** Time horizon T, arm set $\mathcal{A}$, prior causal bounds $\{[L_{a,\mathbf{c}}, U_{a,\mathbf{c}}]\}_{a,\mathbf{c}\in\{\mathcal{A},\mathcal{C}\}}$, $\alpha \in \mathbb{R}^+$.
2: **for** t = 1,2,3,...,T **do**
3:     Observe $\mathbf{x}_{t,a} \in \mathbb{R}^d$ for user profile c_t and every arm a
4:     **for** $a \in \mathcal{A}$ **do**
5:         **if** a is new **then**
6:             $A_a \leftarrow \mathbf{I_d}$, $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$
7:         **end if**
8:         $\hat{\boldsymbol{\theta}}_a \leftarrow A_a^{-1}\mathbf{b}_a$
9:         $UCB_a(t) \leftarrow \hat{\boldsymbol{\theta}}_a^{\mathrm{T}}\mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^{\mathrm{T}} A_a^{-1}\mathbf{x}_{t,a}}$
10:        $\overline{UCB}_a(t) \leftarrow \min\{UCB_a(t), U_{a,\mathbf{c}_t}\}$
11:        *//Truncated UCB*
12:    **end for**
13:    Pull arm $a_t \leftarrow argmax_{a\in\mathcal{A}}\overline{UCB}_a(t)$, and observe a reward r_{t,a_t}
14:    $A_{a_t} \leftarrow A_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^{\mathrm{T}}$, $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_{t,a_t}\mathbf{x}_{t,a_t}$
15: **end for**

Theorem 5. *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:

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

where M denotes the upper bound of $||\boldsymbol{\theta}_a||_2$ for all arms, λ denotes the penalty factor corresponding to the ridge regression estimator $\hat{\boldsymbol{\theta}}_a$.

We follow the standard procedure of deriving the expected regret bound for linear contextual bandit algorithms in (Abbasi-Yadkori, Pál, and Szepesvári 2011) and (Lattimore and Szepesvári 2020). We next discuss the potential improvement in regret that can be achieved by applying LinUCB-PCB algorithm in comparison to original LinUCB algorithm.

Theorem 6. *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.*

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.

Extension We have also investigated leveraging the developed causal bounds to further improve one state-of-the-art contextual bandit algorithm (Hao, Lattimore, and Szepesvari 2020) and one classical non-contextual bandit algorithm (Lattimore and Szepesvári 2020). We include our developed OAM-PCB and UCB-PCB algorithms as well as the corresponding regret analysis in (Huang and Wu 2023).

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 2. We generate 30000 data points following the conditional probability table in (Huang and Wu 2023) to simulate the confounded and selection biased setting. After conducting the preferential exclusion indicated by the selection mechanism, there are approximately 15000 data points used for offline evaluation.

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 2 which only takes into account confounding bias (Biased), and a naive conditional probability estimate derived without considering both confounding and selection biases (CP). We further report the causal bounds and the estimated reward among 16 different values of the context vector in Figure 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.

Online Bandit Learning We use 15000 samples from the generated data to simulate the online bandit learning process. In Figure 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

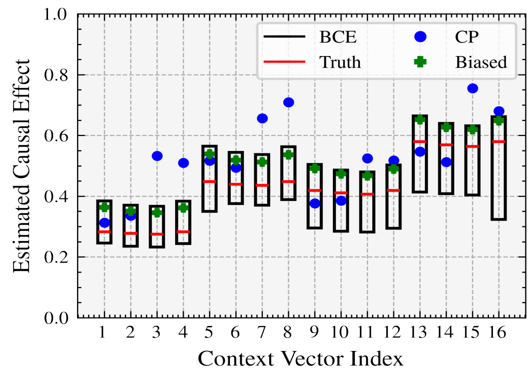

Figure 3: Comparison results for offline evaluation under confounding and selection biases.

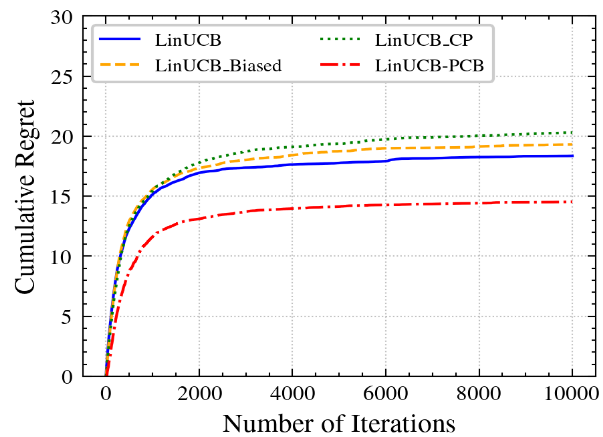

Figure 4: Comparison results for contextual linear bandit.

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.

Conclusion

This work studies 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 techniques to derive causal bounds 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 future work, we will study incorporating causal bounds into advanced bandit algorithms such as state-of-the-art linear contextual bandits (Hao, Lattimore, and Szepesvari 2020), contextual bandits under non-linearity assumption (Zhou, Li, and Gu 2020), and bandits with adversarial feedback (Luo et al. 2023), to comprehensively demonstrate the generalization ability of our approach.

Ethics Statement

Our research could benefit online recommendation system providers so that they can mitigate biases from multiple sources while conducting recommendations. Our research could also benefit users of online recommendation systems as we aim to eliminate the influence of biases and achieve accurate personalized recommendation. Since fairness could be regarded as a certain type of bias, our research could be further extended to prevent users from receiving biased recommendations, especially for those from disadvantaged groups. To the best knowledge, we do not see any negative ethical impact from our paper.

Acknowledgements

This work was supported in part by NSF under grants 1910284, 1940093, 1946391, 2137335, and 2147375.

References

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

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

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

Bareinboim, E.; and Tian, J. 2015. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

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

Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; and He, X. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3): 1–39.

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 208–214.

Correa, J.; and Bareinboim, E. 2017. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Correa, J.; Tian, J.; and Bareinboim, E. 2018. Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Correa, J. D.; Tian, J.; and Bareinboim, E. 2019. Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 2744–2751.

Ding, K.; Li, J.; and Liu, H. 2019. Interactive anomaly detection on attributed networks. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 357–365.

Hao, B.; Lattimore, T.; and Szepesvari, C. 2020. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics*, 3536–3545. PMLR.

Huang, W.; and Wu, X. 2023. Robustly Improving Bandit Algorithms with Confounded and Selection Biased Offline Data: A Causal Approach. *arXiv preprint arXiv:2312.12731*.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

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

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

Luo, H.; Tong, H.; Zhang, M.; and Zhang, Y. 2023. Improved High-Probability Regret for Adversarial Bandits with Time-Varying Feedback Graphs. In *International Conference on Algorithmic Learning Theory*, 1074–1100. PMLR.

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

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

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

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

Tang, Q.; and Xie, H. 2021. A Robust Algorithm to Unifying Offline Causal Inference and Online Multi-armed Bandit Learning. In *2021 IEEE International Conference on Data Mining (ICDM)*, 599–608. IEEE.

Tennenholtz, G.; Shalit, U.; Mannor, S.; and Efroni, Y. 2021. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, 430–439. PMLR.

Tian, J. 2004. Identifying Conditional Causal Effects. In Chickering, D. M.; and Halpern, J. Y., eds., *UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence, Banff, Canada, July 7-11, 2004*, 561–568. AUAI Press.

Tian, J.; and Pearl, J. 2002. A General Identification Condition for Causal Effects. In Dechter, R.; Kearns, M. J.; and Sutton, R. S., eds., *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*, 567–573. AAAI Press / The MIT Press.

Tian, J.; and Pearl, J. 2003. On the identification of causal effects.

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

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

Zhang, J.; and Bareinboim, E. 2021. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.

Zhou, D.; Li, L.; and Gu, Q. 2020. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, 11492–11502. PMLR.