TECHNICAL REPORT
R-61
First version: Jun, 2020
Last version: Jan, 2021

Bounding Causal Effects on Continuous Outcome

Junzhe Zhang, Elias Bareinboim
Causal Artificial Intelligence Laboratory
Columbia University
{junzhez,eb}@cs.columbia.edu

Abstract

We investigate the problem of bounding causal effects from experimental studies in which treatment assignment is randomized but the subject compliance is imperfect. It is well known that under such conditions, the actual causal effects are not point-identifiable due to uncontrollable unobserved confounding. In their seminal work, Balke and Pearl (1994) derived the tightest bounds over the causal effects in this settings by employing an algebra program to derive analytic expressions. However, Pearl's approach assumes the primary outcome to be discrete and finite. Solving such a program could be intractable when high-dimensional context variables are present. In this paper, we present novel non-parametric methods to bound causal effects on the continuous outcome from studies with imperfect compliance. These bounds could be generalized to settings with the high-dimensional context.

Introduction

One of the most common methods for policy learning used throughout the empirical sciences is the use of randomization of the treatment assignment. This method is considered the gold standard within many disciplines and can be traced back, at least, to Fisher (Fisher 1935) and Neyman (Neyman 1923). Whenever human subjects are at the center of the experiment, unfortunately, issues of *non-compliance* arise, namely, subjects do not necessarily follow the experimental protocol and end up doing what they want. It is well-understood that under such conditions, *confounding bias* will emerge. For instance, subjects who did not comply with the treatment assignment may be precisely those who would have responded adversely to the treatment. Therefore, the actual *causal effects* of the treatment, when it is applied uniformly to the population, might be substantially less effective than of the data reveals.

To cope with this bias, analysts may resort to exploit theoretical assumptions underlying the interactions between compliance and response (Wright 1928; Angrist, Imbens, and Rubin 1996). The problem of identifying causal effects from observed data provided with causal assumptions about the data-generating mechanisms, represented in the form of a *directed acyclic causal diagram* (Pearl 2000, Ch. 1.2), has been extensively studied in the causal inference literature. Several criteria and algorithms have been developed (Pearl 2000; Spirtes, Glymour, and Scheines 2000; Bareinboim and Pearl 2016). For example, a criterion called *back-door* (Pearl 2000, Ch. 3.2.2) permits one to determine whether causal effects can be obtained by covariate adjustment and subsequent inverse probability weighting. This condition is also known as *conditional ignorability* and *unconfoundeness* (Rosenbaum and Rubin 1983). Efficient estimators were developed based on the propensity score (Rosenbaum and Rubin 1983; Bang and Robins 2005) and off-policy learning (Dudík, Langford, and Li 2011; Li, Munos, and Szepesvari 2015; Munos et al. 2016; Thomas and Brunskill 2016).

By and large, the combination of causal assumptions and observational data does not always allow one to point-identify the causal effect, called the *non-identifiable*. That is, there exists more than one parametrization of the target effect that are compatible with the same observational data and qualitative assumptions (Pearl 2000, Def. 3.2.2). A causal effect is partially identifiable if it is not identifiable, but the set of its possible values is smaller than the original parameter space. Inferring about the treatment effect in the partially identifiable settings has been a target of growing interest in the domains of causal inference (Balke and Pearl 1995; Chickering and Pearl 1996; Richardson et al. 2014; Cinelli et al. 2019), and more recently, in machine learning (Kallus and Zhou 2018; Kallus, Puli, and Shalit 2018). Among these works, two approaches are often employed: (1) bounds are derived for the target effect under minimal assumptions; or (2) additional untestable assumptions are invoked under which the causal effect is identifiable, and then sensitivity analysis is conducted to assess how the target causal effect varies as the untestable assumptions are changed. This paper focuses on the bounding approach.

(Robins 1989; Manski 1990) derived the first informative bounds over the causal effects from studies with imperfect compliance, under a set of non-parametric assumptions called *instrumental variables*. In their seminal work (Balke and Pearl 1994a, 1997), Balke and Pearl improved earlier results by employing a computer algebraic program to derive analytic expressions of the causal bounds, which are provably optimal. Despite the optimality guarantees provided in their treatment, there are still significant challenges in performing the partial identification of the causal effects with the presence of instrumental variables. First, Pearl's approaches assume the outcome is discrete and finite, which is often not the case in many practical applications. Second, in settings with the high-dimensional context, solving the formulated program is often intractable due to computational and sample complexity issues.

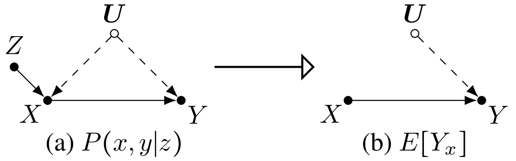
The goal of this paper is to overcome these challenges. We investigate the partial identification of the causal effect on the continuous outcome, with the presence of instrumental variables and the high-dimensional context. More specifically, our contributions are as follows. (1) We identify a set of novel non-parametric assumptions that explicate the inherent independence relationships among the latent counterfactual variables (also called the potential outcomes) when instrumental variables are present. (2) Using the proposed model, we formulate the linear programs that bound the target causal effect on the continuous outcome from studies with imperfect compliance, which is provable optimal. (3) We provide efficient estimation procedures for the derived bounds from finite observational sample under high-dimensional context. Finally, we apply the derived causal bounds to various bandit learning algorithms (Gittins 1979), showing that they could consistently improve the convergence for identifying the optimal treatment. Our results are validated on the International Stroke Trial data (Carolei et al. 1997). Given the space constraints, all proofs are provided in the full technical report (Zhang and Bareinboim 2020).

Preliminaries

In this section, we introduce the basic notations and definitions used throughout the paper. We use capital letters to denote variables (X) and small letters for their values (x). Let $\boldsymbol{\mathcal{X}}$ stand for the domain of X and $|\boldsymbol{\mathcal{X}}|$ for its dimension. We use $P(x)$ to represent probabilities $P(X = x)$.

The basic semantical framework of our analysis rests on *structural causal models* (SCM) (Pearl 2000, Ch. 7). A SCM M is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u})\rangle$, where $\boldsymbol{U}$ is a set of exogenous (unobserved) variables and $\boldsymbol{V}$ is a set of endogenous (observed) variables. $\mathcal{F}$ is a set of structural functions where $f_{V_i} \in \mathcal{F}$ decides the values of $V_i \in \boldsymbol{V}$ taking as argument a combination of other endogenous and exogenous variables (i.e., $V_i \leftarrow f_{V_i}(Pa_{V_i}, U_{V_i}), Pa_{V_i} \subseteq \boldsymbol{V}, U_{V_i} \subseteq \boldsymbol{U}$). The values of $\boldsymbol{U}$ are drawn from the distribution $P(\boldsymbol{u})$, inducing an observational distribution $P(\boldsymbol{v})$ over $\boldsymbol{V}$. Each SCM is associated with a causal diagram in the form of a directed acyclic graph G, where nodes represent variables and arrows stand for functional relationships (e.g., see Fig. 6). By convention, whenever clear from the context, the exogenous $\boldsymbol{U}$ are left implicit. The bi-directed arrows between V_i and V_j indicate the existence of an unobserved confounder (UC) U_k affecting both V_i and V_j, i.e., $V_i \leftarrow U_k \rightarrow V_j$

An intervention on a set of endogenous variables $\boldsymbol{X}$, denoted by $do(\boldsymbol{x})$, is an operation where values of $\boldsymbol{X}$ are set to constants $\boldsymbol{x}$, regardless of how they were ordinarily determined (through the functions $\{f_X : \forall X \in \boldsymbol{X}\}$). For a SCM M, let $M_{\boldsymbol{x}}$ be a modified sub-model of M under intervention $do(\boldsymbol{x})$. The potential outcome of Y to intervention $do(\boldsymbol{x})$, denoted by $Y_{\boldsymbol{x}}(\boldsymbol{u})$, is the solution for Y with $\boldsymbol{U} = \boldsymbol{u}$ in the sub-model $M_{\boldsymbol{x}}$; it can be read as the counterfactual sentence "the value that Y would have obtained in situation $\boldsymbol{U} = \boldsymbol{u}$, had $\boldsymbol{X}$ been $\boldsymbol{x}$." Statistically, averaging $\boldsymbol{u}$ over the distribution $P(\boldsymbol{u})$ leads to the interventional distribution $P(y_{\boldsymbol{x}})$. For a detailed survey on the structural causal models, we refer readers to (Pearl 2000, Ch. 7).



(a) $P(x, y|z)$ (b) $E[Y_x]$

Figure 1: Causal diagram of the instrumental variable (IV) model: Z represents the (randomized) treatment assigned, X the treatment actually received, and Y the outcome.

One fundamental problem in causal inference is to estimate $P(y_{\boldsymbol{x}})$ from the combination of the observational distribution $P(\boldsymbol{v})$ and causal diagram G. An interventional distribution $P(y_{\boldsymbol{x}})$ is identifiable from G if for any pair of SCMs M_1 and M_2 compatible with G, $P_{M_1}(y_{\boldsymbol{x}}) = P_{M_2}(y_{\boldsymbol{x}})$ whenever $P_{M_1}(\boldsymbol{v}) = P_{M_2}(\boldsymbol{v})$ (Pearl 2000, pp. 77). In other words, $P(y_{\boldsymbol{x}})$ are non-identifiable if there exists a pair of SCMs that give arise to the same $P(\boldsymbol{v})$ and G but induce different distributions $P(y_{\boldsymbol{x}})$.

New Bounds on Causal Effects

We will focus on the a special type of SCM called the *instrumental variable* (IV) models which represent experimental studies with imperfect compliance (Pearl 2000, Ch. 8.2). Fig. 6a shows the causal diagram of the IV model where Z represents the (randomized) treatment assigned, X the treatment actually received, and Y the observed outcome; exogenous variables $\boldsymbol{U}$ summarize the unknown factors about an individual subject that affect both X and Y. The values of Y are continuous, decided by a function $y \leftarrow f_Y(x, \boldsymbol{u})$ bounded in $[0, 1]$. We assume Z, X are both finite. For each $Z = z$, values of X are decided by an unknown mechanism $x \leftarrow f_X(z, \boldsymbol{u})$. The data collected from the studies are summarized as the observational distribution $P(x, y|z)$.

Given $P(x, y|z)$, we are interested in inferring the expected outcome on Y by of performing a treatment $do(x)$, i.e., the causal effect $E[Y_x]$. Fig. 6 graphically describes this learning settings. Unfortunately, the non-identifiability of the treatment effect $E[Y_x]$ from the surrogate Z and UCs between X and Y was shown in (Bareinboim and Pearl 2012). To overcome this challenge, we will consider the problem of partial identification in IV models (Manski 2003). Instead of pin-pointing the target quantity, the goal of partial identification is to derive bounds on the parameter space of the causal effect $E[Y_x]$ from the observational data $P(x, y|z)$, called the *causal bound*.

Restricted Instrumental Variable Models

Let $\mathcal{M}_{\text{IV}}[P(x, y|z)]$ denote a set of IV models described in Fig. 6a which are compatible with distribution $P(x, y|z)$; therefore, for any $M \in \mathcal{M}_{\text{IV}}[P(x, y|z)]$, $P_M(x, y|z) = P(x, y|z)$. We could derive causal bounds $E[Y_x] \in [l_x, h_x]$

by solving the optimization problems as follows:

$$\left.\begin{array}{l} l_x = \min\limits_{M \in \mathcal{M}_{\text{OBS}}} E_M[Y_x] \\ h_x = \max\limits_{M \in \mathcal{M}_{\text{OBS}}} E_M[Y_x] \end{array}\right| \mathcal{M}_{\text{OBS}} = \mathcal{M}_{\text{IV}}[P(x,y|z)] \quad (1)$$

The challenge of solving Eq. (1) is that the parametric forms of the exogenous variables $\boldsymbol{U}$ and structural functions $\mathcal{F}$ are not explicitly specified. $\mathcal{M}_{\text{IV}}[P(x,y|z)]$ could be infinitely large, making it hard to derive the bounds $[l_x, h_x]$.

We will provide efficient methods to overcome this challenge. In particular, we propose a new non-parametric representation for the pair $\boldsymbol{U}, \mathcal{F}$. Let $Y_{\boldsymbol{\mathcal{X}}}$ denote a vector of potential responses $(Y_{x_0}, \dots, Y_{x_{|\boldsymbol{\mathcal{X}}|-1}})$ where each $x_i \in \boldsymbol{\mathcal{X}}$ and let $y_{\boldsymbol{\mathcal{X}}}$ be its realizations; $X_{\boldsymbol{\mathcal{Z}}} = x_{\boldsymbol{\mathcal{Z}}}$ is similarly defined. We first describe a family of IV models where $\boldsymbol{U}, \mathcal{F}$ are well-specified from counterfactuals $X_{\boldsymbol{\mathcal{Z}}}, Y_{\boldsymbol{\mathcal{X}}}$.

Definition 1 (Restricted IV (RIV) Model)**.** A restricted instrumental variable model is a SCM $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, P(\boldsymbol{u}) \rangle$ in Fig. Fig. 6a where $\boldsymbol{V} = \{X, Y, Z\}$, $\boldsymbol{U} = \{X_{\boldsymbol{\mathcal{Z}}}, Y_{\boldsymbol{\mathcal{X}}}\}$. Given a vector $x_{\boldsymbol{\mathcal{Z}}}$, x_z is an element in $x_{\boldsymbol{\mathcal{Z}}}$ at the index $z_i = z$; similarly, y_x is an element in $y_{\boldsymbol{\mathcal{X}}}$ at $x_i = x$. Values of X, Y are decided by functions $f_X, f_Y \in \mathcal{F}$ defined as:

$$x \leftarrow f_X(z, x_{\boldsymbol{\mathcal{Z}}}) = x_z, \qquad y \leftarrow f_Y(x, y_{\boldsymbol{\mathcal{X}}}) = y_x, \qquad (2)$$

Y_{x_i} are mutually independent given $X_{\boldsymbol{\mathcal{Z}}}$, i.e., for any x_i,

$$Y_{x_i} \perp\!\!\!\perp \{Y_{x_j} : \forall x_j \neq x_i\} | X_{\boldsymbol{\mathcal{Z}}} \qquad (3)$$

At first glance, the conditional independence among $Y_{\boldsymbol{\mathcal{X}}}$ in Def. 1 may be surprising since it seems to impose additional constraints about the exogenous $\boldsymbol{U}$ in the original IV model. We will show in sequel that this restriction indeed captures the natural properties of the optimization problem in Eq. (1). Fig. 2a shows the graphical representation of a RIV model. The square labeled with $\boldsymbol{\mathcal{X}}$ indicates that there are $|\boldsymbol{\mathcal{X}}|$ nodes Y_{x_i} of this kind. The counterfactuals $X_{\boldsymbol{\mathcal{Z}}}, Y_{\boldsymbol{\mathcal{X}}}$ are the exogenous variables $\boldsymbol{U}$ affecting the treatment X and outcome Y, respectively. $X_{\boldsymbol{\mathcal{Z}}}, Y_{\boldsymbol{\mathcal{X}}}$ are correlated; each potential reward node Y_{x_i} is d-separated from other nodes Y_{x_j} where $x_j \neq x_i$ given $X_{\boldsymbol{\mathcal{Z}}}$ (Pearl 2000, Def. 1.2.3). The counterfactual distribution $P(x_{\boldsymbol{\mathcal{Z}}}, y_{\boldsymbol{\mathcal{X}}})$ can be written as

$$P(x_{\boldsymbol{\mathcal{Z}}}, y_{\boldsymbol{\mathcal{X}}}) = P(x_{\boldsymbol{\mathcal{Z}}}) \prod_{x_i \in \boldsymbol{\mathcal{X}}} P(y_{x_i} | x_{\boldsymbol{\mathcal{Z}}}).$$

Let $\mathcal{M}_{\text{RIV}}[P(x,y|z)]$ denote a set of RIV models compatible with $P(x,y|z)$. We will show that solving Eq. (1) is equivalent to optimizing $E[Y_x]$ over the feasible region $\mathcal{M}_{\text{OBS}} = \mathcal{M}_{\text{RIV}}[P(x,y|z)]$.

Theorem 1. *Given $P(x,y|z)$, for any IV model $M_1 \in \mathcal{M}_{\text{IV}}[P(x,y|z)]$, there exists an IV model $M_2 \in \mathcal{M}_{\text{RIV}}[P(x,y|z)]$ such that (s.t.) $E_{M_1}[Y_x] = E_{M_2}[Y_x]$, and vice versa.*

Thm. 1 says that for any IV model M of Fig. 6a inducing the observational data $P(x,y|z)$, we could always reduce it into a RIV model in $\mathcal{M}_{\text{RIV}}[P(x,y|z)]$ while preserving its treatment effects $E[Y_x]$. Optimizing Eq. (1) within the feasible region $\mathcal{M}_{\text{OBS}} = \mathcal{M}_{\text{RIV}}[P(x,y|z)]$ thus induces causal bounds $E[Y_x] \in [l_x, h_x]$. The sharpness of $[l_x, h_x]$ follows immediately from Def. 1, i.e., there exist SCMs $M_1, M_2 \in \mathcal{M}_{\text{IV}}[P(x,y|z)]$ such that $E_{M_1}[Y_x] = l_x, E_{M_2}[Y_x] = h_x$.

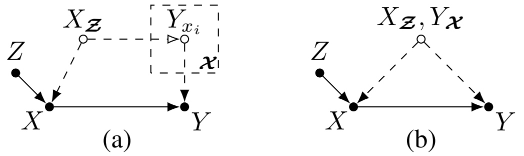

$X_{\boldsymbol{\mathcal{Z}}}$
Y_{x_i}
$\boldsymbol{\mathcal{X}}$
Z
X
Y
(a)
$X_{\boldsymbol{\mathcal{Z}}}, Y_{\boldsymbol{\mathcal{X}}}$
Z
X
Y
(b)

Figure 2: Causal diagrams of (a) a RIV model where Y_{x_i} are mutually independent given $X_{\boldsymbol{\mathcal{Z}}}$; (b) an unrestricted RIV model. The square labelled with $\boldsymbol{\mathcal{X}}$ represents $|\boldsymbol{\mathcal{X}}|$ nodes of which only a single example Y_{x_i} is shown explicitly.

A Linear Program Formulation

We now turn our attention to solving the optimization problem in Eq. (1) within the feasible region $\mathcal{M}_{\text{OBS}} = \mathcal{M}_{\text{RIV}}[P(x,y|z)]$. We will use probabilities $P(x_{\boldsymbol{\mathcal{Z}}})$ and $E[Y_{x_i}|x_{\boldsymbol{\mathcal{Z}}}]P(x_{\boldsymbol{\mathcal{Z}}})$ as unknown parameters. Basic probabilistic properties and $y \leftarrow f_Y(x, \boldsymbol{u}) \in [0,1]$ imply:

$$\sum_{x_{\boldsymbol{\mathcal{Z}}}} P(x_{\boldsymbol{\mathcal{Z}}}) = 1, \; 0 \leq E[Y_{x_i}|x_{\boldsymbol{\mathcal{Z}}}]P(x_{\boldsymbol{\mathcal{Z}}}) \leq P(x_{\boldsymbol{\mathcal{Z}}}) \leq 1 \quad (4)$$

By Def. 1, $P(x,y|z)$ could be written as linear combinations of $P(x_{\boldsymbol{\mathcal{Z}}})$ and $E[Y_{x_i}|x_{\boldsymbol{\mathcal{Z}}}]P(x_{\boldsymbol{\mathcal{Z}}})$ as

$$P(x|z) = \sum_{x_{\boldsymbol{\mathcal{Z}}}} I_{x_z = x} P(x_{\boldsymbol{\mathcal{Z}}}), \qquad (5)$$

$$E[Y|x,z]P(x|z) = \sum_{x_{\boldsymbol{\mathcal{Z}}}} I_{x_z = x} E[Y_x|x_{\boldsymbol{\mathcal{Z}}}]P(x_{\boldsymbol{\mathcal{Z}}}). \qquad (6)$$

where $I_{\{.\}}$ is an indicator function. Similarly, the causal effects $E[Y_x]$ can be written as a linear function:

$$E[Y_x] = \sum_{x_{\boldsymbol{\mathcal{Z}}}} E[Y_x|x_{\boldsymbol{\mathcal{Z}}}]P(x_{\boldsymbol{\mathcal{Z}}}).$$

Eq. (1) is reducible to linear programs (LP) optimizing the objective function $E[Y_x]$ subject to probabilistic constraints Eq. (4) and observational constraints Eq. (6), i.e.,

$$\left.\begin{array}{l} l_x = \min E[Y_x] \\ h_x = \max E[Y_x] \end{array}\right| \text{subject to Eqs. (4) to (6)} \qquad (7)$$

Solving such a linear program leads to a valid causal bound over the expected reward $E[Y_x]$.

Comparision with Existing Methods The idea of modeling the exogenous variables $\boldsymbol{U}$ and functional relationships in $\mathcal{F}$ using its projection to the latent counterfactuals $X_{\boldsymbol{\mathcal{Z}}}, Y_{\boldsymbol{\mathcal{X}}}$ has been explored in the literature, including the canonical partition (Balke and Pearl 1994b) and principal stratification (Frangakis and Rubin 2002). These methods can be seen as a RIV model without the independence restriction among $Y_{\boldsymbol{\mathcal{X}}}$ (see Fig. 2b). For discrete X, Y, Z, representing $P(x_{\boldsymbol{\mathcal{Z}}}, y_{\boldsymbol{\mathcal{X}}})$ of this unrestricted model requires a probability table of size $\mathcal{O}(|\boldsymbol{\mathcal{X}}|^{|\boldsymbol{\mathcal{Z}}|}|\boldsymbol{\mathcal{Y}}|^{|\boldsymbol{\mathcal{X}}|})$; while a RIV model (Def. 1) requires a table of only $\mathcal{O}(|\boldsymbol{\mathcal{X}}|^{|\boldsymbol{\mathcal{Z}}|+1}|\boldsymbol{\mathcal{Y}}|)$, removing the exponential dependence on $|\boldsymbol{\mathcal{X}}|$. In addition, our representation does not require a particular parametric form of outcome (e.g., Y is discrete). This allows one to derive causal bounds on the continuous outcome by employing standard LP methods, which are provably optimal.

Contextual Settings

We now study the contextual IV (CIV) model shown in Fig. 3a where an additional context $\boldsymbol{C}$ is now observed. We denote by $do(\pi(x|\boldsymbol{c}))$ a stochastic intervention where values of treatment are decided following a conditional distribution $\pi(x|\boldsymbol{c})$ mapping from $\boldsymbol{C}$ to X. The expected reward $E[Y_{\pi(x|\boldsymbol{c})}]$ induced by $do(\pi(x|\boldsymbol{c}))$ is given by

$$E[Y_{\pi(x|\boldsymbol{c})}] = \sum_{x,\boldsymbol{c}} E[Y_x|\boldsymbol{c}]\pi(x|\boldsymbol{c})P(\boldsymbol{c}). \tag{8}$$

We are interested in inferring the causal effect $E[Y_{\pi(x|\boldsymbol{c})}]$ from the observational distribution $P(x,y,\boldsymbol{c}|z)$. The non-identifiability of this learning settings, shown in Fig. 3, has been acknowledged in the causal inference literature (Tian 2008; Correa and Bareinboim 2019).

We thus consider the partial identification problem that bounds $E[Y_{\pi(x|\boldsymbol{c})}]$ from $P(x,y,\boldsymbol{c}|z)$ in CIV models. Among quantities in Eq. (8), $\pi(x|\boldsymbol{c}), P(\boldsymbol{c})$ are provided. It is thus sufficient to bound the conditional causal effect $E[Y_x|\boldsymbol{c}]$. Let $\mathcal{M}_{\text{CIV}}[P(x,y,\boldsymbol{c}|z)]$ denote contextual IV models that are compatible with the observational distribution $P(x,y,\boldsymbol{c}|z)$. We show that the formulation of Def. 1 are also applicable in the contextual settings.

Theorem 2. *Given $P(x,y,\boldsymbol{c}|z)$, fix a context $\boldsymbol{C} = \boldsymbol{c}$, for any CIV model $M_1 \in \mathcal{M}_{\text{CIV}}[P(x,y,\boldsymbol{c}|z)]$, there exists an IV model $M_2 \in \mathcal{M}_{\text{RIV}}[P(x,y|\boldsymbol{c},z)]$ such that $E_{M_1}[Y_x|\boldsymbol{c}] = E_{M_2}[Y_x]$, and vice versa.*

Thm. 2 implies that bounding $E[Y_x|\boldsymbol{c}]$ from $P(x,y,\boldsymbol{c}|z)$ in CIV models is equivalent to bounding $E[Y_x]$ from the observational data $P(x,y|\boldsymbol{c},z)$ in RIV models. For any $M \in \mathcal{M}_{\text{CIV}}[P(x,y,\boldsymbol{c}|z)]$, one could always translate it into a solution of Eq. (1) within the feasible region $\mathcal{M}_{\text{RIV}}[P(x,y|\boldsymbol{c},z)]$. Let $E[Y_x|\boldsymbol{c}] \in [l_x(\boldsymbol{c}), h_x(\boldsymbol{c})]$ denote the solutions of Eq. (1) with $\mathcal{M}_{\text{OBS}} = \mathcal{M}_{\text{RIV}}[P(x,y|\boldsymbol{c},z)]$. Causal bounds $E[Y_{\pi(x|\boldsymbol{c})}] \in [l_\pi, h_\pi]$ are computable from $[l_x(\boldsymbol{c}), h_x(\boldsymbol{c})]$ following Eq. (8).

Theorem 3. *Given $P(x,y,\boldsymbol{c}|z)$, there exist CIV models M_1, $M_2 \in \mathcal{M}_{\text{CIV}}[P(x,y,\boldsymbol{c}|z)]$ such that $E_{M_1}[Y_{\pi(x|\boldsymbol{c})}] = l_\pi$ and $E_{M_2}[Y_{\pi(x|\boldsymbol{c})}] = h_\pi$.*

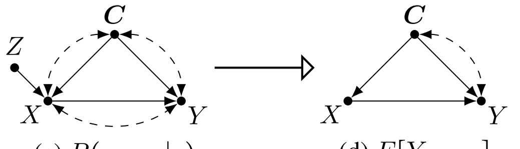
Thm. 3 guarantees that $[l_\pi, h_\pi]$ are optimal in CIV models. Suppose there exists a bound $E[Y_{\pi(x|\boldsymbol{c})}] \in [l'_\pi, h'_\pi]$ strictly contained in $[l_\pi, h_\pi]$. We can always find CIV models M_1, M_2 compatible with $P(x,y,\boldsymbol{c}|z)$ while their $E[Y_{\pi(x|\boldsymbol{c})}]$ lie outside $[l'_\pi, h'_\pi]$, which is a contradiction.

The conditional bound $E[Y_x|\boldsymbol{c}] \in [l_x(\boldsymbol{c}), h_x(\boldsymbol{c})]$ is obtainable by solving Eq. (7) where observational constraints $P(x|z), E[Y|x,z]P(x|z)$ in Eq. (6) are replaced with conditional quantities $P(x|\boldsymbol{c},z), E[Y|x,z,\boldsymbol{c}]P(x|z,\boldsymbol{c})$. Approximating solution to Eq. (7) generalizes the natural bounds in (Manski 1990) to the contextual settings.

Theorem 4. *For an CIV model M, given $P(x,y,\boldsymbol{c}|z)$, $\max_z l_\pi(z) \leq E[Y_{\pi(x|\boldsymbol{c})}] \leq \min_z h_\pi(z)$ where*

$$l_\pi(z) = \sum_{x,\boldsymbol{c}} E[Y|x,\boldsymbol{c},z]\pi(x|\boldsymbol{c})P(x,\boldsymbol{c}|z),$$

$$h_\pi(z) = l_\pi(z) + \sum_{x,\boldsymbol{c}} \pi(X \neq x|\boldsymbol{c})P(x,\boldsymbol{c}|z).$$



(c) $P(x,y,\boldsymbol{c}|z)$ (d) $E[Y_{\pi(x|\boldsymbol{c})}]$

Figure 3: Causal diagram of the contextual instrumental variable (CIV) model: Z represents the (randomized) treatment assigned, X the treatment actually received, and Y the observed outcome; $\boldsymbol{C}$ is an additional observed context.

Estimation in High-Dimensional Context

Causal bounds developed so far are functions of the observational distribution, which are identifiable from the sampling process, and so generally can be consistently estimated. Howevers, computational and sample complexity challenges could arise when context $\boldsymbol{C}$ is high-dimensional.

In this section, we will introduce robust estimation procedures to circumvent this issue. We first assume that the models $\hat{P}(x|\boldsymbol{c},z)$ and $\hat{E}[YI_{\{X=x\}}|\boldsymbol{c},z]$ of observational quantities $P(x|\boldsymbol{c},z)$ and $E[YI_{\{X=x\}}|\boldsymbol{c},z]$ are provided. Let $\{X_i, Y_i, \boldsymbol{C}_i, Z_i\}_{i=1}^n$ denote finite samples drawn from an observational distribution $P(x,y,\boldsymbol{c},z)$. For a sampled instance with context $\boldsymbol{C}_i$, we could compute functions $l_{X_i}(\boldsymbol{C}_i), h_{X_i}(\boldsymbol{C}_i)$ by solving LPs in Eq. (7) with observational constraints $P(x|z), E[Y|x,z]P(x|z)$ in Eq. (6) replaced with $\hat{P}(x|\boldsymbol{C}_i,z)$ and $\hat{E}[YI_{\{X=x\}}|\boldsymbol{C}_i,z]$. Since we consider only a fixed $\boldsymbol{C} = \boldsymbol{C}_i$, Eq. (7) could be solved efficiently despite of dimensionalities of context $\boldsymbol{C}$. Thm. 2 ensures that the conditional causal effect $E[Y_x|\boldsymbol{C}_i] \in [l_x(\boldsymbol{C}_i), h_x(\boldsymbol{C}_i)]$. We could obtain causal bounds $E[Y_{\pi(x|\boldsymbol{c})}] \in [l_\pi, h_\pi]$ by summing over finite samples. Formally, the empirical estimates $\hat{l}_\pi, \hat{h}_\pi$ of causal bounds l_π, h_π are given by:

$$\hat{l}_\pi = \frac{1}{n}\sum_{i=1}^n \sum_x \pi(x|\boldsymbol{C}_i) l_x(\boldsymbol{C}_i),\ \hat{h}_\pi = \frac{1}{n}\sum_{i=1}^n \sum_x \pi(x|\boldsymbol{C}_i) h_x(\boldsymbol{C}_i).$$

Lemma 1. *If $\hat{P}(x|\boldsymbol{c},z)$ and $\hat{E}[YI_{\{X=x\}}|\boldsymbol{c},z]$ are models of $P(x|\boldsymbol{c},z)$ and $E[YI_{\{X=x\}}|\boldsymbol{c},z]$, $\hat{l}_\pi, \hat{h}_\pi$ are consistent estimators of l_π, h_π.*

When the models of $P(x|\boldsymbol{c},z)$ and $E[YI_{\{X=x\}}|\boldsymbol{c},z]$ are not available, one could obtain an efficient approximation of the causal bounds using the empirical estimates of the natural bounds in Thm. 4. For any $Z = z$, let $n(z) = \sum_i^n I_{Z_i=z}$. The estimators $\hat{l}_\pi(z), \hat{h}_\pi(z)$ are defined as follows:

$$\hat{l}_\pi(z) = \frac{1}{n(z)} \sum_{i=1}^n Y_i I_{Z_i=z} \pi(X_i|\boldsymbol{C}_i),$$

$$\hat{h}_\pi(z) = \hat{l}_\pi(z) + \frac{1}{n(z)} \sum_{i=1}^n I_{Z_i=z} \pi(X \neq X_i|\boldsymbol{C}_i).$$

Lemma 2. *$\hat{l}_\pi(z), \hat{h}_\pi(z)$ are consistent estimators of functions $l_\pi(z), h_\pi(z)$ defined in Thm. 4.*

The causal effect $E[Y_{\pi(x|\boldsymbol{c})}]$ could then be bounded from the finite samples $\{X_i, Y_i, \boldsymbol{C}_i, Z_i\}_{i=1}^n$ by inequalities $\max_z \hat{l}_\pi(z) \leq E[Y_{\pi(x|\boldsymbol{c})}] \leq \min_z \hat{h}_\pi(z)$.

Bandit Algorithms with Causal Bounds

The causal bounds derived so far may seem to be uninformative since they do not immediately identify the optimal treatment in IV models. We will that this is not the case. More specifically, we will introduce a systematic procedure to incorporate causal bounds in online bandit algorithms (Auer, Cesa-Bianchi, and Fischer 2002; Sen, Shanmugam, and Shakkottai 2018; Audibert and Bubeck 2010) for identifying the optimal treatment. Our analysis reveals that causal bounds consistently improve the performance of bandit algorithms in various learning settings.

For the IV model of Fig. 6(a), we denote by μ_x the expected reward $E[Y_x]$ of performing a treatment $do(x)$. Let $\mu^* = \max_x \mu_x$ and let x^* denote the optimal treatment (so, $\mu_{x^*} = \mu^*$). An bandit algorithm learns the optimal treatment through repeated episodes of experiments $t = 1, \dots, T$. At each episode t, the algorithm performs an intervention $do(X_t)$ and observes an outcome Y_t. The cumulative regret $E[R_T]$ after T episodes is defined by $E[R_T] = T\mu^* - \sum_{t=1}^{T} E[Y_t]$, i.e., the loss due to the fact that the algorithm does not always play the optimal arm. A desirable asymptotic property of an algorithm is to have $\lim_{T \to \infty} E[R_T]/T = 0$, meaning that the procedure converges and finds the optimal treatment x^*.

We also consider the contextual IV model of Fig. 3(a). Let $\boldsymbol{\Pi}$ be a finite set of candidate policies $\{\pi_1, \dots, \pi_N\}$. Let $\mu_\pi = E[Y_{\pi(x|\boldsymbol{c})}]$ and let π^* denote the optimal policy $\pi^* = \arg\max_{\pi_i \in \boldsymbol{\Pi}} \mu_{\pi_i}$. At each episode $t = 1, \dots, T$, a bandit algorithm has access to a context $\boldsymbol{C}_t$, picks a policy π_t, assigns a treatment X_t following $\pi_t(x|\boldsymbol{C}_t)$ and observes a reward Y_t. Observational data $P(x, y, \boldsymbol{c}|z)$ are provided prior to the experiments. Similarly, the cumulative regret $E[R_T^{\boldsymbol{\Pi}}]$ after T episodes is defined as $E[R_T^{\boldsymbol{\Pi}}] = T\mu_{\pi^*} - \sum_{t=1}^{T} E[Y_t]$. We will assess and compare the performance of bandit algorithms in terms of the cumulative regrets.

Causal UCB

Our methods follow the well celebrated principle of *optimism in the face of uncertainty* (OFU). This principle leads to efficient bandit strategies, in the form of the upper confidence bound (UCB) algorithms (Auer, Cesa-Bianchi, and Fischer 2002). We will next describe UCB strategy in bandit models with a set of candidate policies $\boldsymbol{\Pi}$ (an arm choice x can be seen as a policy in $\boldsymbol{\mathcal{X}}$). At each round t, a UCB agent constructs a set of "plausible" models $\mathcal{M}_t$ that are consistent with the data. The agent then identifies the most "favorable" model from $\mathcal{M}_t$ and prescribes the optimal policy for the identified model. In practice, this decision is determined by the concentration bounds over possible models. An upper (lower) confidence bound $U_\pi(t)$ ($L_\pi(t)$) for a policy π at time t is the maximum (minimum) expected reward of π of models in $\mathcal{M}_t$. The agent then prescribes a policy based on $U_x(t), L_x(t)$ and observes a reward.

We now incorporate causal bounds using the OFU principle, called the Causal-UCB (for short, UCB^c). Let μ_{M_π} denote the expected reward of a policy $\pi \in \boldsymbol{\Pi}$ in model M. At each round t, we obtain a subset $\mathcal{M}_t^c$ from $\mathcal{M}_t$ by removing models inconsistent with the causal bounds, i.e., $\mathcal{M}_t^c = \{M \in \mathcal{M}_t : \mu_{M_\pi} \in [l_\pi, h_\pi]\}$. The agent then prescribes a decision that is optimal to the most favorable model in the subset $\mathcal{M}_t^c$. When the causal bounds are beneficial and significantly reduce the complexities of the search space $\mathcal{M}_t$, it is expected that a UCB^c agent outperforms the standard (non-causal) method. Alg. 1 describes an implementation of UCB^c in bandit settings. At trial t, we clip confidence bounds $U_\pi(t), L_\pi(t)$ using the causal bound $[l_\pi, h_\pi]$. The agent then proceeds ordinarily with the clipped bounds $\overline{U}_\pi(t), \overline{L}_\pi(t)$. Likewise, UCB^c in IV models of Fig. 6a follows Alg. 1 with π replaced with arm $x \in \boldsymbol{\mathcal{X}}$.

Algorithm 1: Causal-UCB (UCB^c)

```
Input: Causal bounds $\{[l_\pi, h_\pi]\}_{\pi \in \boldsymbol{\Pi}}$.
for all $t$ do
    For each $\pi \in \boldsymbol{\Pi}$, compute $\overline{U}_\pi(t), \overline{L}_\pi(t)$ as:
        $\overline{U}_\pi(t) = \max\{\min\{U_\pi(t), h_\pi\}, l_\pi\},$
        $\overline{L}_\pi(t) = \max\{\min\{L_\pi(t), h_\pi\}, l_\pi\},$        (9)
    where $U_\pi(t), L_\pi(t)$ are, respectively, the upper and lower confidence bounds for policy $\pi$.
    Play an arm $do(X_t \sim \pi_t)$ and observe $Y_t$.
end for
```

For the remainder of this section, we will apply the UCB^c strategy (Alg. 1) to the state-of-the-art bandit algorithms, showing its consistent improvements for various tasks. We refer the interested readers to the technical report (Zhang and Bareinboim 2020) for detailed implementations.

Multi-Armed Bandits We start with IV models of Fig. 6a. (Cappé et al. 2013) proposed the kl-UCB procedure that is applicable for bandit settings with bounded reward. It computes confidence bound $U_x(t)$ for each arm x using Cramer's theorem. The agent then plays an arm X_t with the largest $U_x(t)$. Let $\Delta_x = \mu^* - \mu_x$ and let $\boldsymbol{\mathcal{X}}^-$ be the subset $\{x \in \boldsymbol{\mathcal{X}} : \mu_x < \mu^*\}$. kl-UCB guarantees a bound on the cumulative regret of:

$$E[R_T] \leq \sum_{x \in \boldsymbol{\mathcal{X}}^-} \left(\frac{\Delta_x}{kl(\mu_x, \mu^*)} \right) \log(T) + o(\log(T)), \quad (10)$$

where $kl(\mu_x, \mu^*) = \mu_x \log(\mu_x/\mu^*) + (1 - \mu_x)\log((1 - \mu_x)/(1 - \mu^*))$, i.e., the Kullback-Leibler divergence between Bernoulli distributions with mean μ_x and μ^*.

We applied Causal UCB strategy (Alg. 1) to kl-UCB by replacing policy π with arm $x \in \boldsymbol{\mathcal{X}}$. We denote the resultant algorithm kl-UCB^c. At trial t, a kl-UCB^c agent pulls an arm X_t with the largest clipped confidence bound $\overline{U}_x(t)$ where $\overline{U}_x(t)$ is obtained following Eq. (9). Let $\boldsymbol{\mathcal{X}}^-_{h_x \geq c}$ denote a set $\{x \in \boldsymbol{\mathcal{X}}^- : h_x \geq c\}$. kl-UCB^c guarantees the asymptotic regret bound as follows:

Theorem 5. *Given $E[Y_x] \in [l_x, h_x]$, the regret $E[R_T]$ of kl-UCB^c after $T \geq 3$ is bounded by*

$$E[R_T] \leq \sum_{x \in \boldsymbol{\mathcal{X}}^-_{h_x \geq \mu^*}} \left(\frac{\Delta_x}{kl(\mu_x, \mu^*)} \right) \log(T) + o(\log(T)).$$

SCMs	IV models (Fig. 6)		Contextual IV (Fig. 3)
Tasks	Cumulative Regret $E[R_T]$	Best-Arm	Cumulative Regret $E[R_T^{\mathbf{\Pi}}]$
Standard	$\mathcal{O}\big(\sum_{x \in \boldsymbol{\mathcal{X}}^-} \big(\frac{\Delta_x}{kl(\mu_x, \mu^*)}\big) \log(T)\big)$	$\mathcal{O}\big(\sum_{x : \boldsymbol{\mathcal{X}}^-} \Delta_x^{-2} \log(\delta^{-1} K \log(\Delta_x^{-2}))\big)$	$\mathcal{O}\big(C\lambda(\mathbf{\Pi}^-) M^2 \log(T)\big)$
Causal	$\mathcal{O}\big(\sum_{x \in \boldsymbol{\mathcal{X}}^-_{h_x \geq \mu^*}} \big(\frac{\Delta_x}{kl(\mu_x, \mu^*)}\big) \log(T)\big)$	$\mathcal{O}\big(\sum_{x : \boldsymbol{\mathcal{X}}^-_{h_x \geq \mu_{\overline{1,2}}}} \Delta_x^{-2} \log(\delta^{-1} K \log(\Delta_x^{-2}))\big)$	$\mathcal{O}\big(C\lambda(\mathbf{\Pi}^-_{h_\pi \geq \mu_{\pi^*}}) M^2 \log(T)\big)$

Table 1: Summary of bandit results presented in this paper. “SCMs” represents the causal diagrams of the corresponding learning settings. “Obj.” stands for the objective that the target agent aims to optimize. “Standard” stands for the asymptotics of the standard UCB-style algorithms, “Causal” for the proposed strategy leveraging the observational distribution.

The regret bound in Thm. 5 is guaranteed to be smaller than Eq. (10) if $\boldsymbol{\mathcal{X}}^-_{h_x \geq \mu^*}$ is strictly contained in $\boldsymbol{\mathcal{X}}^-$. This result coincides with the optimal regret of B-kl-UCB (Zhang and Bareinboim 2017) when bounds over the expected reward are provided. Since $\Delta_x / kl(\mu_x, \mu^*) \leq 1/(2\Delta_x)$, the improvement of kl-UCB^c is significant when the gap Δ_x of $x \in \boldsymbol{\mathcal{X}}^-_{h_x < \mu^*}$ is small and close to zero.

Best Arm Identification We also consider the settings of pure exploration in IV models (Mannor and Tsitsiklis 2004). Rather than looking at the cumulative regret, we are concerned with PAC-style (“probably approximately correct”) bound on the sample complexity to identify the optimal treatment. In (Jamieson and Nowak 2014), a sampling strategy, called lil’LUCB, were provided, which finds the optimal arm with probability at least $1 - \frac{2+\epsilon}{\epsilon/2}(\log(1+\epsilon))^{-(1+\epsilon)}\delta$ in at most $\mathcal{O}(\sum_{x : \boldsymbol{\mathcal{X}}^-} \Delta_x^{-2} \log(\delta^{-1} |\boldsymbol{\mathcal{X}}| \log(\Delta_x^{-2})))$ trials for any $\epsilon \in (0, 1)$ and $\delta \in (0, \log(1+\epsilon)/e)$.

We apply Alg. 1 to lil’LUCB and denote the resulting algorithm lil’LUCB^c. Assume (without loss of generality) that arms are ordered such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_N$ (so, $\mu^* = \mu_1$). Let $\mu_{\overline{1,2}} = (\mu_1 + \mu_2)/2$. The following theorem provides the sample complexity analysis of lil’LUCB^c.

Theorem 6. *Given $E[Y_x] \in [l_x, h_x]$, with probability (w.p.) at least $1 - \frac{2+\epsilon}{\epsilon/2}(\log(1+\epsilon))^{-(1+\epsilon)}\delta$, lil’LUCB^c returns the optimal treatment x^* with $\mathcal{O}\Big(\sum_{x \in \boldsymbol{\mathcal{X}}^-_{h_x \geq \mu_{\overline{1,2}}}} \Delta_x^{-2} \log(\delta^{-1} |\boldsymbol{\mathcal{X}}| \log(\Delta_x^{-2}))\Big)$ samples.*

Compared with lil’LUCB, the sample complexity bound in Thm. 6 is tighter if $\boldsymbol{\mathcal{X}}^-_{\mu_{\overline{1,2}}} \subset \boldsymbol{\mathcal{X}}^-$.

Contextual Bandits Finally, we consider the regret minimization in the contextual IV models of Fig. 3a. (Sen, Shanmugam, and Shakkottai 2018) proposed a UCB-style algorithm for contextual bandits, called D-UCB, when a set of stochastic policies $\pi \in \mathbf{\Pi}$ are provided, i.e., $\pi(x|z) > 0$. At each round t, D-UCB estimates $U_\pi(t)$ for each policy π using the concentration bounds of the clipped importance sampling estimator (Sen, Shanmugam, and Shakkottai 2018), and apply the policy with the highest $U_\pi(t)$ estimation. Let $M(\pi_i, \pi_j)$ denote the log divergence between two arbitrary policies π_i, π_j (Sen, Shanmugam, and Shakkottai 2018, Def. 2), and let $M = \max_{\pi_i, \pi_j \in \mathbf{\Pi}} M(\pi_i, \pi_j)$. For any $\mathbf{\Pi}' \subseteq \mathbf{\Pi}$, let policies in $\mathbf{\Pi}'$ be ordered such that $\mu_{\pi_1} \geq \mu_{\pi_2} \geq \cdots \geq \mu_{\pi_n}$ where $n = |\mathbf{\Pi}'|$. For any $\pi \in \mathbf{\Pi}$, let $\Delta_\pi = \mu_{\pi^*} - \mu_\pi$. We define function $\lambda(\mathbf{\Pi}') = \Delta_{\pi_n}\gamma(\Delta_{\pi_n}) + \sum_{i=1}^{n-1} \Delta_{\pi_i}(\gamma(\Delta_{\pi_i}) - \gamma(\Delta_{\pi_{i+1}}))$, where $\gamma(\Delta) = \log^2(6/\Delta)/\Delta^2$. Let $\mathbf{\Pi}^-$ denote the set of sub-optimal policies $\{\pi \in \mathbf{\Pi} : \mu_\pi < \mu_{\pi^*}\}$. D-UCB obtains an asymptotic regret bound as follow:

$$E[R_T^{\mathbf{\Pi}}] \leq C\lambda(\mathbf{\Pi}^-) M^2 \log(T) + o(\log(T)), \qquad (11)$$

where C is a constant. We apply the causal strategy (Alg. 1) to D-UCB and denote the new procedure D-UCB^c. Let $\mathbf{\Pi}^-_{h_x \geq c}$ be a set of sub-optimal policies $\{\pi \in \mathbf{\Pi}^- : h_\pi \geq c\}$. We now provides the regret bound of D-UCB^c.

Theorem 7. *Given $E[Y_{\pi(x|\boldsymbol{c})}] \in [l_\pi, h_\pi]$, the regret $E[R_T^{\mathbf{\Pi}}]$ of D-UCB^c after $T \geq 2$ is bounded by*

$$E[R_T^{\mathbf{\Pi}}] \leq C\lambda(\mathbf{\Pi}^-_{h_x \geq \mu_{\pi^*}}) M^2 \log(T) + o(\log(T)). \qquad (12)$$

Compared with Eq. (15), the bound in Thm. 7 only differs in the hardness measure $\lambda(\mathbf{\Pi}^-_{\mu_{\pi^*}})$. The following lemma guarantees that $\lambda(\mathbf{\Pi}^-_{\mu_{\pi^*}})$ is never larger than $\lambda(\mathbf{\Pi}^-)$.

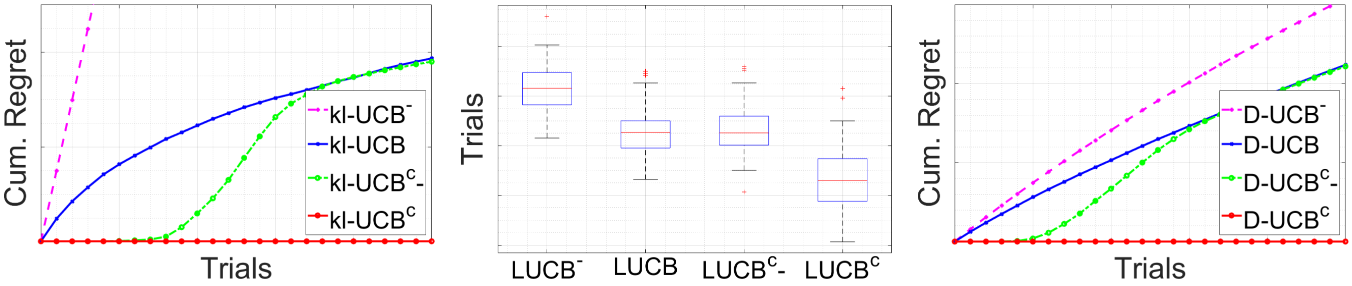
Lemma 3. *For any $\mathbf{\Pi}_1 \subset \mathbf{\Pi}_2 \subseteq \mathbf{\Pi}$, $\lambda(\mathbf{\Pi}_1) \leq \lambda(\mathbf{\Pi}_2)$. If $\Delta_{\pi_1} < \cdots < \Delta_{\pi_{|\mathbf{\Pi}|}}$, $\lambda(\mathbf{\Pi}_1) < \lambda(\mathbf{\Pi}_2)$.*

Thm. 7, together with Lem. 3, says that D-UCB^c improves over D-UCB if there exists some sub-optimal π with $h_\pi < \mu_{\pi^*}$ (given that μ_π of each $\pi \in \mathbf{\Pi}^-$ are not all equal).

We summarize in Table. 1 the results discussed in this section. “Standard” stands for the asymptotics of the standard algorithms and “Causal” is our strategy leveraging the causal bounds obtained from the observation. The interesting aspect is that the causal approach is guaranteed to rival the standard algorithms, even when the observation is biased towards the wrong decision. If the causal bounds are beneficial (e.g., $h_x < \mu^*$), our approach could eliminate a sub-optimal arm x (or policy π) early during the trials, thus outperforming the standard methods. Such improvements could be significant in more difficult instances, when the gap Δ_x between x and x^* is small, e.g., close to zero.

Experiments: International Stroke Trials

We now use a real-world dataset to investigate the performance of proposed bandit strategies. Specifically, we study the International Stroke Trial (IST) (Carolei et al. 1997), focusing on the effect of the aspirin allocation X on a composite score Y, a continuous value in $[0, 1]$ predicting the likelihood of patients’ recovery. We also measure the pre-treatment attributes $\boldsymbol{U}$, including age, gender, and conscious state. To emulate unobserved confounding, as discussed in the paper, we filter the IST data following a inclusion rule $f(x|z, \boldsymbol{u})$ and hide some columns of $\boldsymbol{U}$. We repeat this procedure and generate observational samples using 4 different



(a) Multi-Arm Bandits

(b) Best Arm ID

(c) Contextual Bandits

Figure 4: Simulations comparing solvers that are causally enhanced (kl-UCB^c, lil'LUCB^c, D-UCB^c) and standard (kl-UCB, lil'LUCB and D-UCB) on the International Stroke Trials data. Graphs are rendered in high resolution and can be zoomed in.

inclusion rules $\{f(x|z_i, \boldsymbol{u})\}_{i=1...4}$. To evaluate the performance of the different bandit strategies, we make bootstrap estimates of the patient's true response from the IST data. For details on the experimental setups, we refer readers to the full technical report (Zhang and Bareinboim 2020).

Multi-Armed Bandits The expected reward of not giving ($X = 0$) and giving aspirin ($X = 1$) are respectively $\mu_0 = 0.6201$ and $\mu_1 = 0.6948$, suggesting an increased chance of recovery from aspirin. The causal bounds $\mu_0 \in [0.5905, 0.6506], \mu_1 \in [0.4839, 0.7527]$ estimated using all $n = 9650$ samples do not permit the identification of the optimal treatment, since one is contained in the other. We deploy a kl-UCB^c agent provided with these causal bounds. For comparison, we also include a standard kl-UCB agent, a kl-UCB^c agent with the causal bounds estimated using 300 samples (kl-UCB^c−), and kl-UCB warm-started with the empirical estimates of $E[Y|x, z]$ (kl-UCB$^-$). The results (Fig. 4a) reveal a significant difference in the cumulative regret (CR) between kl-UCB (CR = 38.63) and kl-UCB^c (CR = 0.07); kl-UCB^c− agent (CR = 37.94) coincides with kl-UCB since the sample size $n = 300$ are not sufficient for obtaining any informative estimate of the causal bounds (Thm. 5); kl-UCB$^-$ (CR = 373.42) performs worst among all strategies due to unobserved confounding.

Best Arm ID We also run the lil'LUCB^c agent with empirical bounds estimated from $n = 9711$ observations (lil'LUCB^c-all) to efficiently identify the optimal treatment ($X = 1$). For comparison, we include the standard lil'LUCB, lil'LUCB^c with empirical bounds obtained from $n = 100$ observations (lil'LUCB^c-100), and lil'LUCB warm-started with estimates of $E[Y|x]$ from observations (lil'LUCB^o). The stopping times T of each algorithm are compared in Fig. 4(b). We can immediately note a dramatic difference in the sample complexities experienced by lil'LUCB^c ($T = 4.4 \times 10^3$) compared to lil'LUCB ($T = 6.5 \times 10^3$) and lil'LUCB^c-100 ($T = 6.5 \times 10^3$). lil'LUCB^o ($T = 8.2 \times 10^3$) performs worst among all algorithms.

Contextual Bandits Suppose we now have access to a context $\boldsymbol{C} = \{\text{sex}\}$. Our goal is to find the optimal treatment among two policies $\pi_0(x|\boldsymbol{c})$ and $\pi_1(x|\boldsymbol{c})$. We also include experiments for more involved candidate policies in the technical report (Zhang and Bareinboim 2020). We estimate causal bounds over μ_π using $n = 9650$ observational samples and provide them to a D-UCB^c agent. For comparision, we include the standard D-UCB, D-UCB^c with causal bounds from $n = 300$ samples (D-UCB^c−), and D-UCB seeded with samples of confounded observations (D-UCB$^-$). The cumulative regrets (CR) of each strategy are measured and compared in Fig. 4c. The analysis reveals a significant difference in CR experienced by D-UCB^c(CR = 0.07) compared to D-UCB (CR = 111.8) and D-UCB^c− (CR = 110.8). Unsurprisingly, D-UCB$^-$ (CR = 153.9) performs worst among all strategies.

These results corroborate with our findings: useful information could be extracted from the confounded, passively-collected data to improve the performance of a learning agent. The causal approaches (e.g, kl-UCB^c) dominate the standard, non-causal methods (kl-UCB) given sufficient observational samples. When the number of observations is not statistically significant, our approaches could still rival the standard methods, i.e., no negative transfer occurs.

Conclusions

In this paper, we investigated the problem of bounding causal effects from experimental studies in which treatment assignment is randomized but the subject compliance is imperfect. Under such conditions, the actual causal effects are not identifiable due to uncontrollable confounding. In particular, we derived informative bounds over the causal effect and accounted for challenging issues due to high-dimensional context and the lack of discreteness. We incorporated these bounds into UCB-like algorithms and proved that the causal approach, leveraging observational data, consistently dominates non-causal, state-of-the-art procedures. We hope that our framework can be useful in practical settings since, even though imperfect, observational data contains causal information and is abundantly available.

Acknowledge

The authors were partially supported by grants from NSF IIS-1704352 and IIS-1750807 (CAREER).

References

Angrist, J.; Imbens, G.; and Rubin, D. 1996. Identification of causal effects using instrumental variables (with Comments). *Journal of the American Statistical Association* 91(434): 444–472.

Audibert, J.-Y.; and Bubeck, S. 2010. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, 13–p.

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

Balke, A.; and Pearl, J. 1994a. Counterfactual Probabilities: Computational Methods, Bounds, and Applications. In de Mantaras, R. L.; and Poole, D., eds., *Uncertainty in Artificial Intelligence 10*, 46–54. San Mateo, CA: Morgan Kaufmann.

Balke, A.; and Pearl, J. 1994b. Counterfactual Probabilities: Computational Methods, Bounds and Applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, 46–54.

Balke, A.; and Pearl, J. 1995. Counterfactuals and policy analysis in structural models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 11–18.

Balke, A.; and Pearl, J. 1997. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association* 92(439): 1172–1176.

Bang, H.; and Robins, J. M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4): 962–973.

Bareinboim, E.; and Pearl, J. 2012. Causal inference by surrogate experiments: z-identifiability. In de Freitas, N.; and Murphy, K., eds., *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 113–120. Corvallis, OR: AUAI Press.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113: 7345–7352.

Cappé, O.; Garivier, A.; Maillard, O.-A.; Munos, R.; Stoltz, G.; et al. 2013. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* 41(3): 1516–1541.

Carolei, A.; et al. 1997. The International Stroke Trial (IST): a randomized trial of aspirin, subcutaneous heparin, both, or neither among 19435 patients with acute ischaemic stroke. *The Lancet* 349: 1569–1581.

Chickering, D.; and Pearl, J. 1996. A clinician's apprentice for analyzing non-compliance. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume Volume II, 1269–1276. Menlo Park, CA: MIT Press.

Cinelli, C.; Kumor, D.; Chen, B.; Pearl, J.; and Bareinboim, E. 2019. Sensitivity Analysis of Linear Structural Causal Models. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 1252–1261. Long Beach, CA: PMLR.

Correa, J.; and Bareinboim, E. 2019. From Statistical Transportability to Estimating the Effect of Stochastic Interventions. In Kraus, S., ed., *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1661–1667. Macao, China: International Joint Conferences on Artificial Intelligence Organization.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 1097–1104. Omnipress.

Fisher, R. 1935. *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Frangakis, C.; and Rubin, D. 2002. Principal Stratification in Causal Inference. *Biometrics* 1(58): 21–29.

Gittins, J. C. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)* 148–177.

Jamieson, K.; Malloy, M.; Nowak, R.; and Bubeck, S. 2014. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, 423–439.

Jamieson, K.; and Nowak, R. 2014. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS)*, 1–6. IEEE.

Kallus, N.; Puli, A. M.; and Shalit, U. 2018. Removing Hidden Confounding by Experimental Grounding. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 10911–10920. Curran Associates, Inc.

Kallus, N.; and Zhou, A. 2018. Confounding-Robust Policy Improvement. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 9289–9299. Curran Associates, Inc.

Kuleshov, V.; and Precup, D. 2014. Algorithms for multi-armed bandit problems. *arXiv preprint arXiv:1402.6028* .

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. ACM.

Li, L.; Munos, R.; and Szepesvari, C. 2015. Toward Minimax Off-policy Value Estimation. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*.

Mannor, S.; and Tsitsiklis, J. N. 2004. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5(Jun): 623–648.

Manski, C. 1990. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings* 80: 319–323.

Manski, C. F. 2003. *Partial identification of probability distributions*. Springer Science & Business Media.

Mattheiss, T. H. 1973. An algorithm for determining irrelevant constraints and all vertices in systems of linear inequalities. *Operations Research* 21(1): 247–260.

Munos, R.; Stepleton, T.; Harutyunyan, A.; and Bellemare, M. 2016. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*.

Neyman, J. 1923. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* 5(4): 465–480.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. 2nd edition, 2009.

Richardson, A.; Hudgens, M. G.; Gilbert, P. B.; and Fine, J. P. 2014. Nonparametric bounds and sensitivity analysis of treatment effects. *Statistical science: a review journal of the Institute of Mathematical Statistics* 29(4): 596.

Robins, J. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In Sechrest, L.; Freeman, H.; and Mulley, A., eds., *Health Service Research Methodology: A Focus on AIDS*, 113–159. Washington, D.C.: NCHSR, U.S. Public Health Service.

Rosenbaum, P.; and Rubin, D. 1983. The central role of propensity score in observational studies for causal effects. *Biometrika* 70: 41–55.

Sen, R.; Shanmugam, K.; and Shakkottai, S. 2018. Contextual Bandits with Stochastic Experts. In *International Conference on Artificial Intelligence and Statistics*, 852–861.

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

Thomas, P.; and Brunskill, E. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2139–2148.

Tian, J. 2008. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*.

Wright, P. 1928. *The Tariff on Animal and Vegetable Oils*. New York, NY: The MacMillan Company.

Zhang, J.; and Bareinboim, E. 2017. Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1340–1346. AAAI Press.

Zhang, J.; and Bareinboim, E. 2020. Bounding Causal Effects on Continuous Outcome. Technical report. URL https://causalai.net/r61.pdf.

Appendix 1. Algorithm Details

In this section, we provide details of the bandit algorithms proposed in the paper. We will denote $\hat{\mu}_x(t)$ the empirical mean estimator of the expected reward μ_x and $N_x(t)$ the number of times arm x is pulled till round t. Similarly, $N_\pi(t)$ represent the number times policy π has been invoked till round t.

Stochastic Multi-Armed Bandits We describe kl-UCB^c in Alg. 2. At Step 3, $f(t)$ is a non-increasing function regarding t, which we set as $f(t) = \log(t) + 3\log(\log(t))$ in the analysis. The clipped confidence bound $\overline{U}_x(t)$ is obtained from $U_x(t)$ and the causal bound $[l_x, h_x]$ following Eq. 1 in UCB^c (Alg. 1).

Best-Arm Identification The procedure of lil'LUCB^c is described in Alg. 3. At Step 2, the unclipped bounds $U_x(t), L_x(t)$ rely on the finite form of the law of iterated logarithm (Jamieson et al. 2014). Specifically, for any $\epsilon \in (0,1)$ and $\delta \in (0, \log(1+\epsilon)/e)$,

$$\begin{aligned} U_x(t) &= \hat{\mu}_x(t) + f(N_x(t), \delta), \\ L_x(t) &= \hat{\mu}_x(t) - f(N_x(t), \delta). \end{aligned} \tag{13}$$

where function $f(n, \delta)$ is equal to

$$(1+\sqrt{\epsilon})\sqrt{\frac{(1+\epsilon)\log(|\boldsymbol{\mathcal{X}}|\delta^{-1}\log((1+\delta/|\boldsymbol{\mathcal{X}}|)n))}{2n}}. \tag{14}$$

Contextual Bandits We will use the importance sampling estimators $\hat{\mu}_\pi(t)$ used in (Sen, Shanmugam, and Shakkottai 2018) for estimating the expected reward of a policy π. We first introduce the notion of log-divergence between two arbitrary policies $\pi_i, \pi_j \in \mathbf{\Pi}$.

Definition 2 (Log-Divergence)**.** Consider the function $f(x) = xe^{x-1}$. We define the log-divergence $M(\pi_i, \pi_j)$ between two arbitrary policies $\pi_i, \pi_j \in \mathbf{\Pi}$ as

$$M(\pi_i, \pi_j) = 1 + \log(1 + \sum_{x,z} f(\frac{\pi_i(x|z)}{\pi_j(x|z)})\pi_j(x|z)P(z)).$$

During the execution of D-UCB^c, we keep track of the experimental history as $\{X_t, Y_t, Z_t, \pi_t\}$ where π_t is the selected policy at time t. We set $Z_\pi(t) = \sum_{\pi' \in \mathbf{\Pi}} N_{\pi'}(t)/M(\pi, \pi')$. The importance sampling estimators $\hat{\mu}_\pi(t)$ of a policy π at time t is defined as:

$$\begin{aligned} \hat{\mu}_\pi(t) = \frac{1}{Z_\pi(t)} \sum_{i=1}^{t} \frac{1}{M(\pi, \pi_i)} Y_i \frac{\pi(X_i|Z_i)}{\pi_i(X_i|Z_i)} \\ \cdot I\{\frac{\pi(X_i|Z_i)}{\pi_i(X_i|Z_i)} \leq 2\log(\frac{2}{\epsilon(t)})M(\pi, \pi_i)\}, \end{aligned}$$

where $I\{\cdot\}$ is an indicator function. $\epsilon(t)$ is an adjustable terms that controls the bias-variance trade-off for the estimator, which we as $2/t$ in our implementation.

The procedure of D-UCB^c is described in Alg. 4. At Step 4, the upper confidence bound $U_\pi(t)$ for policy π is defined as:

$$U_\pi(t) = \hat{\mu}_\pi(t) + \frac{3}{2}\beta(t). \tag{15}$$

where $\beta(t)$ is chosen such that

$$\frac{\beta(t)}{\log(2/\beta(t))} = \frac{\sqrt{Ct\log(t)}}{Z_\pi(t)} \tag{16}$$

We pick $C = 16$ in the analysis.

In practice, however, it is difficult to derive a closed-form solution for $\beta(t)$ in Eq. 16. We follow the implementation of

Algorithm 2: kl-UCB^c

```
Input: A list of bounds over $\mu_x$: $\{[l_x, h_x]\}_{x \in \boldsymbol{\mathcal{X}}}$
Pull each arm of $x \in \boldsymbol{\mathcal{X}}$ once
for all $t = |\boldsymbol{\mathcal{X}}| + 1$ to $T$ do
    For each arm $x \in \boldsymbol{\mathcal{X}}$, let
        $U_x(t) = \sup\left\{\mu_x \in [0,1] : kl(\hat{\mu}_x(t), \mu_x) \leq f(t)/N_x(t)\right\}$
    Compute the clipped bound $\overline{U}_x(t)$ following Eq. 1, where
    Pick an arm $X_t = \arg\max_{x \in \boldsymbol{\mathcal{X}}} \overline{U}_x(t)$.
end for
```

Algorithm 3: lil'LUCB^c

```
Input: $\epsilon \in (0,1)$, $\delta \in (0, \log(1+\epsilon)/e)$, bounds over
  $\mu_x : \{[l_x, h_x]\}_{x \in \boldsymbol{\mathcal{X}}}$.
repeat
    For each arm $x \in \boldsymbol{\mathcal{X}}$, compute $U_x(t), L_x(t)$ following
      Eqs. 13-14.
    Compute the clipped bounds $\overline{U}_x(t), \overline{L}_x(t)$ following Eq. 1.
    Let $x_h, x_l$ be two arms with the largest $\overline{U}_x(t)$, i.e.,
      $x_h, x_l = arg\overset{2}{\underset{x}{\max}}\,\overline{U}_x(t)$. Sample each of $x_h, x_l$ once.
until $\overline{L}_{x_h}(t) > \overline{U}_{x_l}(t)$
return $x_h$
```

Algorithm 4: D-UCB^c

```
Input: A list of bounds over $\mu_\pi$: $\{[l_\pi, h_\pi]\}_{\pi \in \boldsymbol{\Pi}}$
For $t = 1$, play an arm following an arbitrary policy $\pi \in \boldsymbol{\Pi}$.
for all $t = 2$ to $T$ do
    Observe context $Z_t$.
    For each policy $\pi \in \boldsymbol{\Pi}$, compute $U_\pi(t)$ following Eq. 15.
    Compute the clipped bound $\overline{U}_\pi(t)$ following Eq. 1.
    Play $X_t \sim \pi_t(X_t|Z_t)$ where $\pi_t = \arg\max_{\pi \in \boldsymbol{\Pi}} \overline{U}_\pi(t)$.
end for
```

(Sen, Shanmugam, and Shakkottai 2018) and approximate $\beta(t)$ using

$$\beta(t) \approx \left(\frac{c_1 t \log(t)}{(Z_\pi(t))^2}\right)^{\frac{1}{2+\epsilon}},$$

where ϵ is small number in $(0,1)$, which we set as $\epsilon = 1 \times 10^{-5}$.

Appendix II. Proofs

In this section, we provide proofs of results in the paper. We first introduce some necessary notations. For all $n > 1$, let $\hat{\mu}_x(t)$ be the empirical estimation of μ_x, let $N_x(t)$ be the number of times arm x is pulled till round t, and let $\tau_{x,n}$ denote the round at which x was pulled for the n-th time, For reward samples from arm x, denoted by $\{Y_{x,0}, \ldots, Y_{x,n}\}$, define $\hat{\mu}_{x,n} = \frac{1}{n}\sum_{s=1}^{n} Y_{x,s}$. We of course have the writing $\hat{\mu}_x(t) = \hat{\mu}_{x,N_x(t)}$.

Proofs of Theorems 1-2

Proof of Theorem 1. Given a SCM $M_1 \in \mathcal{M}_{IV}[P(x,y|z)]$, we will translate it to a $M_2 \in \mathcal{M}_{RIV}[P(x,y|z)]$ while preserving the treatment effect $E[Y_x]$. We define a RIV model M_2 such that $P_{M_2}(x_{\boldsymbol{\mathcal{Z}}}) = P_{M_1}(x_{\boldsymbol{\mathcal{Z}}})$ and $P_{M_2}(y_{x_i}|x_{\boldsymbol{\mathcal{Z}}}) = P_{M_1}(y_{x_i}|x_{\boldsymbol{\mathcal{Z}}})$. $P_{M_1}(x,y|z)$ can be written as:

$$\begin{aligned}
P_{M_1}(x,y|z) &= \sum_{x_{\boldsymbol{\mathcal{Z}}}} I_{x_z = x} \sum_{\boldsymbol{u}} I_{Y_x(\boldsymbol{u}) = y, X_{\boldsymbol{\mathcal{Z}}}(\boldsymbol{u}) = x_{\boldsymbol{\mathcal{Z}}}} P_{M_1}(\boldsymbol{u}) \\
&= \sum_{x_{\boldsymbol{\mathcal{Z}}}} I_{x_z = x} P_{M_1}(y_x|x_{\boldsymbol{\mathcal{Z}}}) P_{M_1}(x_{\boldsymbol{\mathcal{Z}}}) \\
&= \sum_{x_{\boldsymbol{\mathcal{Z}}}} I_{x_z = x} P_{M_2}(y_x|x_{\boldsymbol{\mathcal{Z}}}) P_{M_2}(x_{\boldsymbol{\mathcal{Z}}}) = P_{M_2}(x,y|z)
\end{aligned}$$

Similarly, we have $E_{M_1}[Y_x] = E_{M_2}[Y_x]$.

Conversely, given a $M_2 \in \mathcal{M}_{RIV}[P(x,y|z)]$, since each RIV model is also an IV model, we have $M_1 = M_2 \in \mathcal{M}_{IV}[P(x,y|z)]$. □

Proof of Theorem 2. We will formulate LPs following the method in Sec. 4.2. let $q_{i,j} = P(X_{z_0} = i, X_{z_1} = j)$ and $e_{i,j,k} = E[Y_{x_i}|X_{z_0} = j, X_{z_1} = k] \cdot q_{i,j}$. Since function f_Y is bounded in $[0,1]$,

$$0 \leq e_{i,j,k} \leq q_{j,k}. \tag{17}$$

We could also have:

$$\begin{aligned}
p_{0,0} &= q_{0,0} + q_{0,1}, \quad p_{0,1} = q_{0,0} + q_{1,0} \\
p_{1,0} &= q_{1,0} + q_{1,1}, \quad p_{1,1} = q_{0,1} + q_{1,1}
\end{aligned} \tag{18}$$

and

$$\begin{aligned}
e_{0,0} &= e_{0,0,0} + e_{0,0,1}, \quad e_{0,1} = e_{0,0,0} + e_{0,1,0} \\
e_{1,0} &= e_{1,1,0} + e_{1,1,1}, \quad e_{1,1} = e_{1,0,1} + e_{1,1,1}.
\end{aligned} \tag{19}$$

The treatment effects $E[Y_{X=0}], E[Y_{X=1}]$ could be written as:

$$E[Y_{X=0}] = e_{0,0,0} + e_{0,0,1} + e_{0,1,0} + e_{0,1,1}. \tag{20}$$

$$E[Y_{X=1}] = e_{1,0,0} + e_{1,0,1} + e_{1,1,0} + e_{1,1,1}. \tag{21}$$

Optimizing Eq. (20) subject to Eqs. (17) to (19) gives the bounds $E[Y_{X=0}] \in [l_0, h_0]$. The solution could be derived by solving the dual problem using vertex enumeration algorithm in (Mattheiss 1973). The causal bound $E[Y_{X=1}] \in [l_1, h_1]$ is similarly obtained. When Y is binary, $[l_x, h_x]$ coincides with the the sharp bound of (Balke and Pearl 1995). □

Proofs of Theorems 3-5

Proof of Theorem 3. Given a $M_1 \in \mathcal{M}_{\text{IV}}[P(x,y,\boldsymbol{c}|z)]$, fix a context $\boldsymbol{c}$, we construct an IV model M_1' such that $P_{M_1'}(x_{\boldsymbol{\mathcal{Z}}}, y_{\boldsymbol{\mathcal{X}}}) = P_{M_1}(x_{\boldsymbol{\mathcal{Z}}}, y_{\boldsymbol{\mathcal{X}}}|\boldsymbol{c})$. It trivially holds that $P_{M_1'}(x,y|z) = P(x,y|\boldsymbol{c},z)$ and $E_{M_1'}[Y_x] = E_{M_1}[Y_x|\boldsymbol{c}]$. By Thm. 1, we can find a RIV M_2 such that $P_{M_2}(x,y,z) = P_{M_1'}(x,y,z)$ and $E_{M_2}[Y_x] = E_{M_1'}[Y_x]$.

Given a $M_2 \in \mathcal{M}_{\text{RIV}}[P(x,y|\boldsymbol{c},z)]$, let M denote the underlying CIV generating the observational data $P(x,y|z)$. Since in Fig. 3(a), $\boldsymbol{C} \perp\!\!\!\perp Z$, for any $z_1 \neq z_2$, we have $P(\boldsymbol{c}|z_1) = P(\boldsymbol{c}|z_2) = P(\boldsymbol{c})$. We will construct a CIV M_1 by combining M and M_2: (1) $P_{M_1}(\boldsymbol{c}) = P(\boldsymbol{c})$; (2) for context $\boldsymbol{c}$, values of X, Y are decided by the RIV M_2; (3) for any $\boldsymbol{c}' \neq \boldsymbol{c}$, values of X, Y are decided by the original model M. It thus trivially holds that $P_{M_1}(x,y,\boldsymbol{c}|z) = P(x,y,\boldsymbol{c}|z)$ and $E_{M_1}[Y_x|\boldsymbol{c}] = E_{M_2}[Y_x]$. □

Proof of Theorem 4. By Thm. 3, for each $\boldsymbol{c}$, there exists a RIV model $M_1(\boldsymbol{c})$ such that for each z, $P_{M_1(\boldsymbol{c})}(x,y|z) = P(x,y|\boldsymbol{c},z)$ and $E_{M_1(\boldsymbol{c})}[Y_x] = l_x(\boldsymbol{c})$. We now construct a CIV M_1 where $P_{M_1}(\boldsymbol{c}) = P(\boldsymbol{c})$; for each $\boldsymbol{c}$, values of X, Y are decided by the RIV $M_1(\boldsymbol{c})$. We thus have:

$$\begin{aligned}P_{M_1}(x,y,\boldsymbol{c}|z) &= P_{M_1}(x,y|\boldsymbol{c},z)P_{M_1}(\boldsymbol{c})\\ &= P_{M_l(\boldsymbol{c})}(x,y|z)P(\boldsymbol{c})\\ &= P(x,y,\boldsymbol{c}|z),\end{aligned}$$

and $E_{M_1}[Y_{\pi(x|\boldsymbol{c})}]$ can be written as:

$$\begin{aligned}E_{M_1}[Y_{\pi(x|\boldsymbol{c})}] &= \sum_{x,\boldsymbol{c}} E_{M_1(\boldsymbol{c})}[Y_x]\pi(x|\boldsymbol{c})P(\boldsymbol{c})\\ &= \sum_{x,\boldsymbol{c}} l_x(\boldsymbol{c})\pi(x|\boldsymbol{c})P(\boldsymbol{c}) = l_\pi.\end{aligned}$$

Similarly, we can construct an IV M_2 with $P_{M_2}(x,y,\boldsymbol{c}|z) = P(x,y,\boldsymbol{c}|z), E_{M_2}[Y_{\pi(x|\boldsymbol{c})}] = h_\pi$. □

Proof of Theorem 5. We first write $E[Y_{\pi(x|\boldsymbol{c})}]$ as:

$$\begin{aligned}E[Y_{\pi(x|\boldsymbol{c})}] &= \sum_{x,\boldsymbol{c}} E[Y_x|\boldsymbol{c}]\pi(x|\boldsymbol{c})P(\boldsymbol{c})\\ &= \sum_{x,\boldsymbol{c}} E[Y_x|\boldsymbol{c},z]\pi(x|\boldsymbol{c})P(\boldsymbol{c}|z).\end{aligned}$$

From the fact that $Y_x \in [0,1]$, we have

$$\begin{aligned}E[Y_{\pi(x|\boldsymbol{c})}] &\geq \sum_{x,\boldsymbol{c}} E[Y|x,\boldsymbol{c},z]P(x|\boldsymbol{c},z)\pi(x|\boldsymbol{c})P(\boldsymbol{c}|z)\\ &= \sum_{x,\boldsymbol{c}} E[Y|x,\boldsymbol{c},z]\pi(x|\boldsymbol{c})P(x,\boldsymbol{c}|z) = l_\pi(z).\end{aligned}$$

Similarly,

$$\begin{aligned}E[Y_{\pi(x|\boldsymbol{c})}] \leq &\sum_{x,\boldsymbol{c}} E[Y|x,\boldsymbol{c},z]P(x|\boldsymbol{c},z)\pi(x|\boldsymbol{c})P(\boldsymbol{c}|z)\\ &+ \sum_{x,\boldsymbol{c}} P(X \neq x|\boldsymbol{c},z)\pi(x|\boldsymbol{c})P(\boldsymbol{c}|z)\\ &= l_\pi(z) + \sum_{x,\boldsymbol{c}}(1 - P(x|\boldsymbol{c},z))\pi(x|\boldsymbol{c})P(\boldsymbol{c}|z)\\ &= l_\pi(z) + \sum_{x,\boldsymbol{c}} \pi(x|\boldsymbol{c})P(\boldsymbol{c}|z) - \sum_{x,\boldsymbol{c}} \pi(x|\boldsymbol{c})P(x,\boldsymbol{c}|z)\\ &= l_\pi(z) + 1 - \sum_{x,\boldsymbol{c}} \pi(x|\boldsymbol{c})P(x,\boldsymbol{c}|z)\\ &= l_\pi(z) + \sum_{x,\boldsymbol{c}} P(x,\boldsymbol{c}|z) - \sum_{x,\boldsymbol{c}} \pi(x|\boldsymbol{c})P(x,\boldsymbol{c}|z)\\ &= l_\pi(z) + \sum_{x,\boldsymbol{c}} \pi(X \neq x'|\boldsymbol{c})P(x,\boldsymbol{c}|z) = h_\pi(z).\end{aligned}$$

Since $E[Y_{\pi(x|\boldsymbol{c})}]$ is not a function of z, taking the maximum and minimum over $l_\pi(z)$ and $h_\pi(z)$ respectively completes the proof. □

Proofs of Lemmas 1-2

Proof of Lemma 1. Let $\{X_i, Y_i, \boldsymbol{C}_i, Z_i\}_{i=1}^n$ denote finite samples drawn from an observational distribution $P(x,y,\boldsymbol{c}|z)$. We define the empirical estimates $\hat{l}_\pi, \hat{h}_\pi$ of causal bounds $E[Y_{\pi(x|\boldsymbol{c})}] \in [l_\pi, h_\pi]$ as follows:

$$\begin{aligned}\hat{l}_\pi &= \frac{1}{n}\sum_{i=1}^n \sum_x \pi(x|\boldsymbol{C}_i)l_x(\boldsymbol{C}_i),\\ \hat{h}_\pi &= \frac{1}{n}\sum_{i=1}^n \sum_x \pi(x|\boldsymbol{C}_i)h_x(\boldsymbol{C}_i).\end{aligned}$$

By basic probabilistic operations,

$$E[Y_{\pi(x|\boldsymbol{c})}] = \sum_{x,\boldsymbol{c}} E[Y_x|\boldsymbol{c}]\pi(x|\boldsymbol{c})P(\boldsymbol{c}).$$

Since $\boldsymbol{C}$ and Z are independent in IV models, $P(\boldsymbol{c}) = P(\boldsymbol{c}|z)$. The consistency of $\hat{l}_\pi$ and $\hat{h}_\pi$ immediately follows. □

Proof. For any $Z = z$, let $n(z) = \sum_i^n Z_i = z$. The estimators $\hat{l}_\pi(z), \hat{h}_\pi(z)$ are defined as follows:

$$\begin{aligned}\hat{l}_\pi(z) &= \frac{1}{n(z)}\sum_{i=1}^n Y_i I_{Z_i=z}\pi(X_i|\boldsymbol{C}_i),\\ \hat{h}_\pi(z) &= \hat{l}_\pi(z) + \frac{1}{n(z)}\sum_{i=1}^n I_{Z_i=z}\pi(X \neq X_i|\boldsymbol{C}_i).\end{aligned}$$

The consistency if the above estimators follows immediately from Thm. 5. □

Proofs of Theorems 6

To prove Thm. 6, we first introduce following lemmas.

Lemma 4. *In kl-UCB^c, the term $\sum_{t=K}^{T-1} P\{\overline{U}_{x^*}(t) < \mu_{x^*}\}$ is bounded by:*

$$\sum_{t=K}^{T-1} P(\overline{U}_{x^*}(t) < \mu_{x^*}) \leq 3 + 4e\log(\log(T))$$

Proof. Let $h^* = h_{x^*}$ and $l^* = l_{x^*}$. Without loss of generality, assume that $l^* < \mu_{x^*} \leq h^*$. Recall that $\overline{U}_{x^*}(t) = (U_{x^*}(t) \wedge h^*) \vee l^*$. We thus have:

$$\begin{aligned}&\sum_{t=K}^{T-1} P(\overline{U}_{x^*}(t) < \mu_{x^*})\\ &= \sum_{t=K}^{T-1} P((U_{x^*}(t) \wedge h^*) \vee l^* < \mu_{x^*})\\ &\leq \sum_{t=K}^{T-1} P(U_{x^*}(t) \wedge h^* < \mu_{x^*}, l^* < \mu_{x^*})\\ &\leq \sum_{t=K}^{T-1} P(U_{x^*}(t) < \mu_{x^*}) + \sum_{t=K}^{T-1} P(h^* < \mu_{x^*})\end{aligned}$$

Since $\mu_{x^*} \leq h^*$, $P(h^* < \mu_{x^*}) = 0$. By (Cappé et al. 2013, Fact A.1), we could further bound $\sum_{t=K}^{T-1} P(U_{x^*}(t) < \mu_{x^*})$:

$$\begin{aligned}\sum_{t=K}^{T-1} P(\overline{U}_{x^*}(t) < \mu_{x^*}) &= \sum_{t=K}^{T-1} P(U_{x^*}(t) < \mu_{x^*})\\ &\leq 3 + 4e\log(\log(T)).\end{aligned}$$ □

Lemma 5. *In kl-UCB^c, the term $\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x)$ is bounded by:*

$$\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x) \le \begin{cases} 0 & \textit{if } h_x < \mu_{x^*} \\ \frac{\log(T)}{kl(\mu_x, \mu_{x^*})} + o(\log(T)) & \textit{otherwise} \end{cases}$$

Proof. We now bound the term by cases:
Case 1. $h_x < \mu_{x^*}$. Recall that $(U_x(t) \wedge h_x) \vee l_x$. $\mu_{x^*} \le \overline{U}_x(t)$ implies that $\mu_{x^*} \le h_x$ which contradicts the fact $h_x < \mu_{x^*}$. We thus have $\sum_{t=K'}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x) = 0$.
Case 2. $h_x \ge \mu_{x^*}$. Since $\overline{U}_x(t) = (U_x(t) \wedge h_x) \vee l_x$,

$$\begin{aligned} &\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x) \\ &\le \sum_{t=K}^{T-1} P(\mu_{x^*} \le U_x(t), \mu_{x^*} \le h_x, X_t = x) \\ &+ \sum_{t=K}^{T-1} P(\mu_{x^*} \le l_x). \end{aligned}$$

Since $l_x \le \mu_x < \mu_{x^*}$ for $x \ne x^*$, $P(\mu_{x^*} \le l_x) = 0$. We could further write $\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x)$ as:

$$\begin{aligned} &\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x) \\ &\le \sum_{t=K}^{T-1} P(\mu_{x^*} \le U_x(t), X_t = x) \\ &= \sum_{t=K}^{T-1} P(\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_x(t), \mu) \le \frac{f(t)}{N_x(t)}, X_t = x) \end{aligned}$$

To continue,

$$\begin{aligned} &\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x) \\ &= \sum_{n=1}^{T-K} \sum_{t=\tau_{x,n}+1}^{\tau_{x,n+1}} P(\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(t)}{n}, X_t = x) \\ &\le \sum_{n=1}^{T-K} \sum_{t=\tau_{x,n}+1}^{\tau_{x,n+1}} P(\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(T)}{n}, X_t = x) \\ &= \sum_{n=1}^{T-K} P(\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(T)}{n}) \\ &\le n_0 + \sum_{n=n_0+1}^{T-K} P(\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(T)}{n}) \end{aligned}$$

where $n_0 = \lceil \frac{f(T)}{kl(\mu_x, \mu_{x^*})} \rceil$. This implies

$$(\forall n \ge n_0 + 1) \quad kl(\mu_x, \mu_{x^*}) > \frac{f(T)}{n}$$

Since $kl(\cdot, \mu_{x^*})$ is continuous decreasing function on $[0, \mu_{x^*}]$, there must $\exists \mu_{\frac{f(T)}{n}} \in (\mu_x, \mu_{x^*}]$, such that:

$$kl(\mu_{\frac{f(T)}{n}}, \mu_{x^*}) \ge \frac{f(T)}{n}$$

We next show that:

$$\{\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(T)}{n}\} \Rightarrow \{\hat{\mu}_{x,n} \ge \mu_{\frac{f(T)}{n}}\}$$

This can be proved by contradiction. Suppose $\hat{\mu}_{x,n} < \mu_{\frac{f(T)}{n}}$, we then have for $\forall \mu \in [\mu_{x^*}, 1]$:

$$kl(\hat{\mu}_{x,n}, \mu) \ge kl(\hat{\mu}_{x,n}, \mu_{x^*}) > kl(\mu_{\frac{f(T)}{n}}, \mu_{x^*}) = \frac{f(T)}{n}$$

which contradicts $\{\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(T)}{n}\}$. Thus, $\forall \lambda > 0$, we have:

$$\begin{aligned} &P(\exists \mu \in [\mu_{x^*}, 1] : kl(\hat{\mu}_{x,n}, \mu) \le \frac{f(T)}{n}) \\ &\le P(\hat{\mu}_{x,n} \ge \mu_{\frac{f(T)}{n}}) \\ &\le e^{-\lambda \mu_{\frac{f(T)}{n}}} \mathbb{E}[e^{\lambda \hat{\mu}_{x,n}}] \end{aligned}$$

By (Cappé et al. 2013, Fact A.2), we have:

$$\sum_{t=K'}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t)) \le \frac{\log(T)}{kl(\mu_x, \mu_{x^*})} + o(\log(T)). \quad \square$$

Lemma 6. *In kl-UCB^c, the number of draws $\mathbb{E}[N_x(T)]$ for any sub-optimal arm a is upper bounded for any horizon $T \ge 3$ as:*

$$\mathbb{E}[N_x(T)] \le \begin{cases} 4 + 4e \log(\log(T)) & \textit{if } h_x < \mu_{x^*} \\ \frac{\log(T)}{kl(\mu_x, \mu_{x^*})} + o(\log(T)) & \textit{otherwise} \end{cases}$$

Proof. Following the same decomposition in (Cappé et al. 2013), we have

$$\begin{aligned} \mathbb{E}[N_x(T)] = 1 + &\underbrace{\sum_{t=K}^{T-1} P(\overline{U}_{x^*}(t) < \mu_{x^*})}_{\text{Term 1}} \\ &+ \underbrace{\sum_{t=K}^{T-1} P(\mu_{x^*} \le \overline{U}_x(t), X_t = x)}_{\text{Term 2}}. \end{aligned}$$

Term 1 and 2 are bounded by Lemma 4 and 5 respectively. Putting everything together, we prove the statement. $\square$

Proof of Theorem 6. The cumulative regret $E[R_T]$ can be written as

$$\begin{aligned} E[R_T] &= \sum_x \Delta_x E[N_x(T)] \\ &= \sum_{x: h_x < \mu_{x^*}} \Delta_x E[N_x(T)] + \sum_{x: h_x \ge \mu_{x^*}} \Delta_x E[N_x(T)]. \end{aligned}$$

From Lem. 6, we have:

$$\begin{aligned} E[R_T] \le &\sum_{x: h_x < \mu_{x^*}} \Delta_x (4 + 4e \log(\log(T))) \\ &+ \sum_{x: h_x \ge \mu_{x^*}} \Delta_x (\frac{\log(T)}{kl(\mu_x, \mu_{x^*})} + o(\log(T))) \\ = &\sum_{x \in \boldsymbol{\mathcal{X}}^-_{h_x \ge \mu_{x^*}}} \left(\frac{\Delta_x}{kl(\mu_x, \mu_{x^*})} \right) \log(T) + o(\log(T)) \, \square \end{aligned}$$

Proofs of Theorems 7

Proof of Theorem 7. We first prove that lil'LUCB^c stops with the optimal arm x^* with hight probability. Consider the event $\mathcal{C}$ defined as

$$\left\{\forall x \in \boldsymbol{\mathcal{X}}, t \in \{1, \dots, T\}, |\hat{\mu}_{x,N_x(t)} - \mu_x| < f(N_x(t), \delta)\right\},$$

where the function $f(n, \delta)$ is defined in Eq. 14. If $\mathcal{C}$ holds true, lil'LUCB^c must stop with the best arm. From (Jamieson et al. 2014, Lem. 1), we can show that the stopping condition is met only with the best arm with probability at least $1 - \frac{2+\epsilon}{\epsilon/2}(\log(1+\epsilon))^{-(1+\epsilon)}\delta$.

We next consider the stopping time of lil'LUCB^c if $\mathcal{C}$ holds true. Assume (without loss of generality) that arms are ordered such that $\mu_1 > \mu_2 \geq \cdots \geq \mu_N$ (so, $\mu^* = \mu_1$). We say arm 1 is BAD if $\overline{L}_1(t) \leq \mu_{\overline{1,2}}$ and an arm $i \neq 1$ is BAD if $\overline{U}_i(t) \geq \mu_{\overline{1,2}}$. We want to show that given event $\mathcal{C}$, if the stopping condition is not satisfied, then either x_l or x_h must be BAD. We will prove this statement by contradiction.

Suppose both x_l and x_h are not BAD, and lil'LUCB^c does not stop, i.e., $\overline{L}_{x_h}(t) \leq \overline{U}_{x_l}(t)$,

1. If $x_h = 1$, $x_l \neq 1$, by definition, $\overline{L}_{x_h}(t) > \mu_{\overline{1,2}} > \overline{U}_{x_l}(t)$. This means that lil'LUCB^c must stop. Contradiction.
2. If $x_h = x \neq 1$, $x_l = 1$, we have
$$((\hat{\mu}_{x,N_x(t)} + f(N_x(t), \delta)) \wedge h_x) \vee l_x < \mu_{\overline{1,2}},$$
and
$$((\hat{\mu}_{1,N_1(t)} - f(N_1(t), \delta)) \wedge h_1) \vee l_1 > \mu_{\overline{1,2}}$$
$$\Rightarrow ((\hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta)) \wedge h_1) \vee l_1 > \mu_{\overline{1,2}}$$
This contradicts that $x_h = x \neq 1$.
3. If $x_h = x \neq 1$, $x_l = x' \neq 1$, we then have
$$((\hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta)) \wedge h_1) \vee l_1 < \mu_{\overline{1,2}}$$
$$\Rightarrow (\hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta)) \wedge h_1 < \mu_{\overline{1,2}}$$
$$\Rightarrow \hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta) < \mu_{\overline{1,2}} \text{ or } h_1 < \mu_{\overline{1,2}}$$
$$\Leftrightarrow \hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta) < \mu_{\overline{1,2}}$$
The last follows from $h_1 \geq \mu_1 > \mu_{\overline{1,2}}$. We further have:
$$\hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta) < \mu_{\overline{1,2}}$$
$$\Rightarrow \hat{\mu}_{1,N_1(t)} + f(N_1(t), \delta) < \mu_1,$$
which contradicts event $\mathcal{C}$.

For arm x with $h_x < \overline{1,2}$, $\overline{U}_x(t) < \mu_{\overline{1,2}}$ for any t, i.e., arm i is never BAD.

For arm x with $h_x \geq \overline{1,2}$, define τ_x be the first integer such that $f(\tau_x, \delta) \leq \Delta_x/4$, and define $\tau_1 = \tau_2$. Assuming the event $\mathcal{C}$ holds, then for any $x \neq 1$ and $s \geq \tau_x$,

$$\hat{\mu}_{x,s} + f(s, \delta) \leq \mu_x + 2f(s, \delta/n)$$
$$= \mu_{\overline{1,2}} + 2f(s, \delta) + \frac{(\mu_x - \mu_1) + (\mu_x - \mu_2)}{2}$$
$$\leq \mu_{\overline{1,2}} + 2f(s, \delta) - \Delta_x/2 \leq \mu_{\overline{1,2}}$$

This implies that for $N_x(t) > \tau_x$, arm x is not bad.

Let $x_l(t), x_h(t)$ denote arms x_l, x_h at round t. By the above arguments, we observe that the total number of rounds does not exceed

$$\sum_{t=1}^{\infty} I\{x_h(t) \text{ is BAD or } x_l(t) \text{ is BAD}\}$$
$$= \sum_{t=1}^{\infty} \sum_{x \in \boldsymbol{\mathcal{X}}} I\{\{x_h(t) = x \text{ or } x_l(t) = x\} \cap \{x \text{ is BAD}\}\}$$
$$= \sum_{t=1}^{\infty} \sum_{x \in \boldsymbol{\mathcal{X}}_{h_x \geq \mu_{\overline{1,2}}}} I\{\{x_h(t) = x \text{ or } x_l(t) = x\} \cap \{x \text{ is BAD}\}\}$$

The last step holds since for arm $x \notin \boldsymbol{\mathcal{X}}_{h_x \geq \mu_{\overline{1,2}}}$, i.e., $h_x < \boldsymbol{\mathcal{X}}_{h_x \geq \mu_{\overline{1,2}}}$, it is never BAD. We can further write the above equation as:

$$\sum_{t=1}^{\infty} \sum_{x \in \boldsymbol{\mathcal{X}}_{h_x \geq \mu_{\overline{1,2}}}} I\{\{x_h(t) = x \text{ or } x_l(t) = x\} \cap \{x \text{ is BAD}\}\}$$
$$\leq \sum_{t=1}^{\infty} \sum_{x \in \boldsymbol{\mathcal{X}}_{h_x \geq \mu_{\overline{1,2}}}} I\{\{x_h(t) = x \text{ or } x_l(t) = x\} \cap \{N_x(t) \leq \tau_x\}\}$$
$$\leq \sum_{x \in \boldsymbol{\mathcal{X}}_{h_x \geq \mu_{\overline{1,2}}}} \tau_x$$

where the last inequality holds by the fact that if $\{x_h(t) = 1 \text{ or } x_l(t) = x\}$, then $N_x(t+1) = N_x(t) + 1$, and this can oly occur τ_x times before $N_x(t) > \tau_x$.

Solving for $f(\tau_i, \delta) \leq \Delta_x/4$ gives:

$$\tau_x \leq \frac{2\gamma}{\Delta_x^2} \log\left(\frac{2\log(\gamma(1+\epsilon)\Delta_x^{-2})}{\delta/|\boldsymbol{\mathcal{X}}|}\right).$$

where $\gamma = 8(1+\sqrt{\epsilon})^2(1+\epsilon)$. Recalling that each round we sample two times, we observe that with probability at least $1 - \frac{2+\epsilon}{\epsilon/2}(\log(1+\epsilon))^{-(1+\epsilon)}\delta$, the algorithm obtains a sample complexity of order $\mathcal{O}(\sum_{x \in \boldsymbol{\mathcal{X}}^-_{h_x \geq \mu_{\overline{1,2}}}} \Delta_x^{-2} \log(\delta^{-1}|\boldsymbol{\mathcal{X}}|\log(\Delta_x^{-2})))$. □

Proofs of Theorem 8 and Lemma 3

To prove Thm. 8, we need to introduce following lemmas.

Lemma 7. *In D-UCB^c, the term $P(\overline{U}_{\pi^*}(t) < \mu_{\pi^*})$ at time t is bounded by*

$$P(\overline{U}_{\pi^*}(t) < \mu_{\pi^*}) \leq t^{-2}.$$

Proof. Recall that $\overline{U}_{\pi^*}(t) = (U_{\pi^*}(t) \wedge h_{\pi^*}) \vee l_{\pi^*}$. Without loss of generality, assume $\mu_{\pi^*} \in (l_{\pi^*}, h_{\pi^*}]$. We have

$$\begin{aligned} P(\overline{U}_{\pi^*}(t) < \mu_{\pi^*}) &= P((U_{\pi^*}(t) \wedge h_{\pi^*}) \vee l_{\pi^*} < \mu_{\pi^*}) \\ &\leq P((U_{\pi^*}(t) \wedge h_{\pi^*}) < \mu_{\pi^*}, l_{\pi^*} < \mu_{\pi^*}) \\ &= P(U_{\pi^*}(t) \wedge h_{\pi^*} < \mu_{\pi^*}) \\ &\leq P(U_{\pi^*}(t) < \mu_{\pi^*}) + P(h_{\pi^*} < \mu_{\pi^*}) \\ &= P(U_{\pi^*}(t) < \mu_{\pi^*}). \end{aligned}$$

The last step follows from the fact $h_{\pi^*} \geq \mu_{\pi^*}$. By (Sen, Shanmugam, and Shakkottai 2018, Lem. 3), we have

$$P(\overline{U}_{\pi^*}(t) < \mu_{\pi^*}) \leq P(U_{\pi^*}(t) < \mu_{\pi^*}) < t^{-2}. \qquad \square$$

Lemma 8. *In D-UCB^c, the term $P(\overline{U}_\pi(t) \geq \mu_{\pi^*}) \leq t^{-2}$ at time*

$$\begin{cases} \forall t > 1 & \text{if } h_\pi < \mu_{\pi^*} \\ t > CM^2\gamma(\Delta_\pi)\log(T) & \text{otherwise} \end{cases}$$

where $C = 144$ and $\gamma(\Delta_\pi) = \frac{\log^2(6/\Delta_\pi)}{\Delta_\pi^2}$.

Proof. If $h_\pi < \mu_{\pi^*}$, since $\overline{U}_\pi(t) = (U_\pi(t) \wedge h_\pi) \vee l_\pi$, we must have $P(\overline{U}_\pi(t) \geq \mu_{\pi^*}) = 0$.

As for $h_\pi \geq \mu_{\pi^*}$, we have

$$P(\overline{U}_\pi(t) \geq \mu_{\pi^*}) \leq P(U_\pi(t) \wedge h_\pi \geq \mu_{\pi^*}) + P(l_\pi \geq \mu_{\pi^*})$$

Since $l_\pi \leq \mu_\pi < \mu_{\pi^*}$, $P(l_\pi \geq \mu_{\pi^*}) = 0$. The above equation can thus be further written as:

$$P(\overline{U}_\pi(t) \geq \mu_{\pi^*}) \leq P(U_\pi(t) \wedge h_\pi \geq \mu_{\pi^*}) \leq P(U_\pi(t) \geq \mu_{\pi^*})$$

Let $t > CM^2\gamma(\Delta_\pi)\log(T)$, by (Sen, Shanmugam, and Shakkottai 2018, Lem. 4), we have

$$P(\overline{U}_\pi(t) \geq \mu_{\pi^*}) \leq P(U_\pi(t) \geq \mu_{\pi^*}) \leq t^{-2}. \qquad \square$$

Lemma 9. *In D-UCB^c algorithm, let π_t denote the policy selected at time $t > 1$. For $\pi \neq \pi^*$, the term $P(\pi_t = \pi) \leq 2t^{-2}$ at time*

$$\begin{cases} \forall t > 1 & \text{if } h_\pi < \mu_{\pi^*} \\ t > CM^2\gamma(\Delta_\pi)\log(T) & \text{otherwise} \end{cases}$$

where $C = 144$ and $\gamma(\Delta_\pi) = \frac{\log^2(6/\Delta_\pi)}{\Delta_\pi^2}$.

Proof. We can decompose $P(\pi_t = \pi)$ as:

$$\begin{aligned} P(\pi_t = \pi) &= P(\overline{U}_{\pi^*}(t) < \mu_{\pi^*}, \pi_t = \pi) \\ &\quad + P(\overline{U}_{\pi^*}(t) \geq \mu_{\pi^*}, \pi_t = \pi) \\ &\leq P(\overline{U}_{\pi^*}(t) < \mu_{\pi^*}) + P(\overline{U}_\pi(t) \geq \mu_{\pi^*}) \end{aligned}$$

The rest proof follows from Lems. 7-8. $\square$

Proof of Theorem 9. Recall that policies in $\mathbf{\Pi}' \subseteq \mathbf{\Pi}$ are ordered such that $\mu_{\pi_1} \geq \mu_{\pi_2} \geq \cdots \geq \mu_{\pi_{|\mathbf{\Pi}'|}}$. Let $d(i)$ denote the index of a policy $\pi_i \in \mathbf{\Pi}^-_{h_\pi \geq \mu_{\pi^*}}$ in set $\mathbf{\Pi}$. For a policy $\pi_k \in \mathbf{\Pi}^-$, let

$$T_k = \lceil CM^2\gamma(\Delta_\pi)\log(T) \rceil.$$

Let $N = |\mathbf{\Pi}^-_{h_\pi \geq \mu_{\pi^*}}|$. The regret of D-UCB^c can be bounded as:

$$\begin{aligned} E[R_T^{\mathbf{\Pi}}] \leq & \underbrace{\sum_{t=1}^{T_{d(N)}-1} \Delta_{\pi_t} P(\pi_t \neq \pi^*)}_{Term1} \\ & + \underbrace{\sum_{t=T_{d(1)}}^{T} \Delta_{\pi_t} P(\pi_t \neq \pi^*)}_{Term2} \\ & + \underbrace{\sum_{k=0}^{N-1} \sum_{t=T_{d(N-k)}}^{T_{d(N-k-1)}-1} \Delta_{\pi_t} P(\pi_t \neq \pi^*)}_{Term3} \end{aligned}$$

Term 1 can be bounded as:

$$\begin{aligned} \sum_{t=1}^{T_{d(N)}-1} \Delta_{\pi_t} P(\pi_t \neq \pi^*) \leq & \sum_{t=1}^{T_{d(N)}-1} \Delta_{\pi_t} P(\pi_t \in \{\pi_2, \ldots, \pi_{d(N)}\}) \\ & + \sum_{t=1}^{T_{d(N)}-1} \sum_{i=d(N)+1}^{|\boldsymbol{\mathcal{X}}|} \Delta_{\pi_i} P(\pi_t = \pi_i) \\ & \leq \Delta_{\pi_{d(N)}} T_{d(N)} + \sum_{t=1}^{T_{d(N)}-1} \sum_{\pi_i \in \mathbf{\Pi}^-} \frac{2\Delta_{\pi_i}}{t^2}. \end{aligned}$$

Term 2 can be bounded as:

$$\begin{aligned} \sum_{t=T_{d(1)}}^{T} \Delta_{\pi_t} P(\pi_t \neq \pi^*) &= \sum_{t=T_{d(1)}}^{T} \sum_{i=2}^{|\boldsymbol{\mathcal{X}}|} \Delta_{\pi_i} P(\pi_t = \pi_i) \\ &\leq \sum_{t=T_{d(1)}}^{T} \sum_{\pi_i \in \mathbf{\Pi}^-} \frac{2\Delta_{\pi_i}}{t^2}. \end{aligned}$$

The last follows from Lem. 9. Term 3 can be bounded as:

$$\begin{aligned} & \sum_{k=0}^{N-2} \sum_{t=T_{d(N-k)}}^{T_{d(N-k-1)}-1} \Delta_{\pi_t} P(\pi_t \neq \pi^*) \\ & \leq \sum_{k=0}^{N-2} \sum_{t=T_{d(N-k)}}^{T_{d(N-k-1)}-1} \Bigg(\Delta_{\pi_{d(N-k-1)}} P(\pi_t \in \{\pi_i\}_{i \leq d(N-k-1)}) \\ & + \sum_{i=d(N-k-1)+1}^{N} \Delta_{\pi_i} P(\pi_t = \pi_i) \Bigg) \\ & \leq \sum_{k=0}^{N-2} \Bigg(\Delta_{\pi_{d(N-k-1)}} (T_{d(N-k-1)} - T_{d(N-k)}) \\ & + \sum_{t=T_{d(N-k)}}^{T_{d(N-k-1)-1}} \sum_{\pi_i \in \mathbf{\Pi}^-} \frac{2\Delta_{\pi_i}}{t^2} \Bigg) \end{aligned}$$

Together, we can bound $E[R_T^{\mathbf{\Pi}}]$ as

$$\begin{aligned} E[R_T^{\mathbf{\Pi}}] \leq & \Delta_{\pi_{d(N)}} T_{d(N)} \\ & + \sum_{t=1}^{T} \sum_{\pi \in \mathbf{\Pi}^-} \frac{2\Delta_\pi}{t^2} \\ & + \sum_{k=0}^{N-2} \Delta_{\pi_{d(N-k-1)}} (T_{d(N-k-1)} - T_{d(N-k)}). \end{aligned}$$

Since $T_k \in [CM^2\gamma(\Delta_\pi)\log(T), CM^2\gamma(\Delta_\pi)\log(T) + 1)$, $E[R_T^{\mathbf{\Pi}}]$

$$\begin{aligned} & \leq CM^2 \Delta_{\pi_{d(N)}} \gamma(\Delta_\pi) \log(T) + \sum_{t=1}^{T} \sum_{\pi \in \mathbf{\Pi}^-} \frac{2\Delta_\pi}{t^2} + \sum_{i=1}^{N} \Delta_{\pi_{d(i)}} \\ & + CM^2 \log(T) \sum_{k=0}^{N-2} \Delta_{\pi_{d(N-k-1)}} (\gamma(\Delta_{\pi_{d(N-k-1)}}) - \gamma(\Delta_{\pi_{d(N-k)}})) \\ & \leq CM^2 \log(T) \Delta_{d(N)}^{-1} + (\pi^2/3) \sum_{\pi_i \in \mathbf{\Pi}^-} \Delta_{\pi_i} + \sum_{i=1}^{N} \Delta_{\pi_{d(i)}} \\ & + CM^2 \log(T) \sum_{k=0}^{N-2} \Delta_{\pi_{d(N-k-1)}} (\gamma(\Delta_{\pi_{d(N-k-1)}}) - \gamma(\Delta_{\pi_{d(N-k)}})). \end{aligned}$$

which gives

$$E[R_T^{\mathbf{\Pi}}] \leq C\lambda(\mathbf{\Pi}^-_{h_x \geq \mu_{\pi^*}}) M^2 \log(T) + o(\log(T)). \qquad \square$$

Proof of Lemma 3. For $\mathbf{\Pi}_1 \subset \mathbf{\Pi_2}$, let $d(i)$ denote the index of a policy $\pi_i \in \mathbf{\Pi}_1$ in set $\mathbf{\Pi}_2$. Let $N = |\mathbf{\Pi}_1|$, $\lambda(\mathbf{\Pi}_1)$ can be written as:

$$
\begin{aligned}
&\lambda(\mathbf{\Pi}_1) \\
&= \Delta_{\pi_{d(N)}} \gamma(\Delta_{\pi_{d(N)}}) + \sum_{i=1}^{N-1} \Delta_{\pi_{d(i)}} (\gamma(\Delta_{\pi_i}) - \gamma(\Delta_{\pi_{i+1}})) \\
&= \Delta_{\pi_{d(N)}} \gamma(\Delta_{\pi_{|\mathbf{\Pi}_2|}}) + \sum_{j=d(N)}^{|\mathbf{\Pi}_2|-1} \Delta_{\pi_{d(N)}} (\gamma(\Delta_{\pi_j}) - \gamma(\Delta_{\pi_{j+1}})) \\
&+ \sum_{i=1}^{N-1} \sum_{j=d(i)}^{d(i+1)-1} \Delta_{\pi_{d(i)}} (\gamma(\Delta_{\pi_j}) - \gamma(\Delta_{\pi_{j+1}})) \\
&\leq \Delta_{\pi_{|\mathbf{\Pi}_2|}} \gamma(\Delta_{\pi_{|\mathbf{\Pi}_2|}}) + \sum_{j=d(N)}^{|\mathbf{\Pi}_2|-1} \Delta_{\pi_j} (\gamma(\Delta_{\pi_j}) - \gamma(\Delta_{\pi_{j+1}})) \\
&+ \sum_{i=1}^{N-1} \sum_{j=d(i)}^{d(i+1)-1} \Delta_{\pi_j} (\gamma(\Delta_{\pi_j}) - \gamma(\Delta_{\pi_{j+1}})) \\
&+ \sum_{j=1}^{d(1)} \Delta_{\pi_j} (\gamma(\Delta_{\pi_j}) - \gamma(\Delta_{\pi_{j+1}})) \\
&\leq \Delta_{\pi_{|\mathbf{\Pi}_2|}} \gamma(\Delta_{\pi_{|\mathbf{\Pi}_2|}}) + \sum_{j=1}^{|\mathbf{\Pi}_2|-1} \Delta_{\pi_j} (\gamma(\Delta_{\pi_j}) - \gamma(\Delta_{\pi_{j+1}})) \\
&\leq \lambda(\mathbf{\Pi}_2)
\end{aligned}
$$

The above inequality strictly holds if there exists a policy $\pi_i \in \mathbf{\Pi}_2 \backslash \mathbf{\Pi}_1$ such that $\Delta_{\pi_i} < \Delta_{\pi_{i+1}}$. □

Appendix III. Experimental Setup

In this section, we describe the International Stroke Trial dataset and details of the experimental setup. Our evaluation framework contains two main components: (1) a censoring procedure creating confounded observational data from perfectly randomized clinical trial records; (2) a simulator modeling the bandit process using logged clinical trial data. We will describe the general procedures of these components, followed by the details of each experiment (if not included in the main text). Finally, we provide further experiment results for a more involved contextual bandit task.

International Stroke Trials

International Stroke Trial (IST) is a randomized clinical trial assessing the treatment effect of aspirin, subcutaneous heparin, both, or neither among $19,435$ patients with acute ischemic stroke (Carolei et al. 1997). Previous studies found that a significant reduction in death or non-fatal recurrent stroke with aspirin. During the trial, a set of pre-treatment attributes U of the patients are collected, including the gender S (0 for male, 1 for female), conscious state C at randomization (0 for unconscious, 1 for drowsy and 2 for awake) and age A (0 if younger than 73 years old, 1 otherwise). We estimate from data the joint distribution $P(s, c, a)$. summarized in Table 2.

	$S = 0$		$S = 1$	
	$A = 0$	$A = 1$	$A = 0$	$A = 1$
$C = 0$	0.002	0.004	0.003	0.006
$C = 1$	0.051	0.051	0.036	0.08
$C = 2$	0.267	0.165	0.146	0.19

Table 2: The probability table for the distribution $P(s, c, a)$ where S, C, A stands for the gender, conscious state and age of the patient.

We focus on the treatment effect of the aspirin allocation X (1 for yes, 0 for no). One of the main contribution of this paper is the relaxation of the key assumption in (Zhang and Bareinboim 2017) that the derivation of causal bounds requires the outcome variable Y to be finite. Unfortunately, the primary clinical outcomes of IST (e.g., death, recurrent stroke, recovery) are all categorical variables. To illustrate the efficiency of our approach, we thus consider the treatment effect of aspirin on an artificial score Y that is a continuous variable in $[0, 1]$. Specifically, the value of the composite score Y is decided by the logistic function defined as:

$$
y = \frac{1}{1 + e^{-g(x,s,c,a)}}, \tag{22}
$$

where the function $g(x, s, c, a)$ is equal to

$$
x + s + 0.2(c - 1) - 2xa(1 - s) + u_y. \tag{23}
$$

U_y is an independent gaussian noise with 0 mean and variance unity. In the above equation, the coefficients represent the direction of the causal effects of the associated variable on the score Y. Specifically, we penalize the unconscious state and the aspirin usage on male patients who are older than 73 years while recognizing the benefits of aspirin allocation on the general population. Throughout our experiments, the form of the score function is never revealed. The agents could only infer about statistical features of Y through finite samples of the environment.

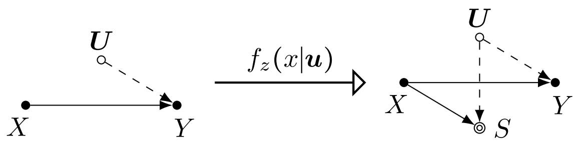
Creating Confounding Bias

In this section, we discuss methods to create a confounded dataset from IST data since it's the very goal of the randomization procedure to remove the confounding bias.

Given the IST dataset $\mathcal{D} = \{X_i, \boldsymbol{U}_i, Y_i\}_{i=1}^{N}$ where the treatment allocation X is fully randomized, independent of the context U, our goal is to obtain a dataset $\mathcal{D}^c = \{X_i, \boldsymbol{U}_i, Y_i\}_{i=1}^{N_c}$ where X and U are associated. To create this association, we censor the $\mathcal{D}$ following an inclusion rule $f_z(x|\boldsymbol{u})$, similar to the procedure used in (Kallus and Zhou 2018).

The censoring procedure goes as follows. For each sample X_i, U_i, Y_i, we draw an independent variable $\tilde{X}_i$ following the selection rule $f_z(x|\boldsymbol{u})$ given U_i, i.e., $\tilde{X}_i \sim f_z(X|\boldsymbol{U}_i)$. We then compare the values of X_i and $\tilde{X}_i$. If $X_i = \tilde{X}_i$, we include the sample X_i, U_i, Y_i in $\mathcal{D}^c$. Otherwise, the sample $X_i, \boldsymbol{U}_i, Y_i$ is dropped.

We describe in Fig. 5(a-b) the graphical representation of this censoring procedure. Given the perfectly randomized data $\mathcal{D}$ of Fig. 6(a), we censor $\mathcal{D}$ using the inclusion



(a) $\mathcal{D}$ (b) $\mathcal{D}^c$

Figure 5: Causal diagrams the censoring procedure. $\mathcal{D}$ is a perfectly randomized data and in $\mathcal{D}^c$, variables X and $\boldsymbol{U}$ are associated through the selection bias introduced by the inclusion rule $f_z(x|\boldsymbol{u})$.

Rules		$S = 0$		$S = 1$	
		$A = 0$	$A = 1$	$A = 0$	$A = 1$
f_{z_1}	$C = 0$	1	0	1	0
	$C = 1$	0	1	1	0
	$C = 2$	0	0.2	0	1
f_{z_2}	$C = 0$	1	0	1	0
	$C = 1$	1	1	1	0
	$C = 2$	1	1	0	1
f_{z_3}	$C = 0$	0	1	0	0.1
	$C = 1$	0	0.3	0	0
	$C = 2$	0	0.0	0	0
f_{z_1}	$C = 1$	1	0	1	0
	$C = 1$	0	1	1	0
	$C = 2$	0	0.2	0	1

Table 3: The inclusion rules $\{f_{z_i}(x|\boldsymbol{u})\}_{i=1,\dots,4}$.

$f_z(x|\boldsymbol{u})$. This censoring procedure introduces the selection bias to the resulting data $\mathcal{D}^c$, creating association between X and U. Since the treatment X in $\mathcal{D}$ is fully randomized, it is verifiable that the conditional distribution $P(x|\boldsymbol{u})$ computed from $\mathcal{D}^c$ coincides with the selection rule, i.e., $P(x|\boldsymbol{u}) = f_z(x|\boldsymbol{u})$.

Note that the composite score Y is affected by the context $\boldsymbol{U}$ (Eq. 22). We thus obtain a dataset $\mathcal{D}^c = \{X_i, U_i, Y_i\}_{i=1}^{N_c}$ where the treatment X and outcome Y are both associated with the covariate $\boldsymbol{U}$. Since our methods only assume that $\boldsymbol{U}$ is non-descendant of the treatment X, it suffices to use the created dataset $\mathcal{D}^c$ as the confounded observational data. To emulate the unobserved confounding, one could simply drop the column of $\boldsymbol{U}$.

Following the procedure described above, we generate confounded observational samples using 4 different includes rules $\{f_{z_i}(x|\boldsymbol{u})\}_{i=1,\dots,4}$, which are defined in Table 3.

The Simulation Framework

We evaluate the performance of each algorithm with a simulation framework modeling the bandit process from the logged clinical trial data (Li et al. 2010; Kuleshov and Precup 2014). For each bandit algorithm, 100 simulations were performed. For the regret minimization task (in both MABs and contextual bandits), each simulation lasts for 5000 trials. In the best-arm identification task, the simulation continues until the bandit procedure stopped with the optimal treatment. All the results presented in the paper form an average over these 100 simulations.

Each simulation proceeds as follows. At each trial t, a patient $\boldsymbol{U}_t$ is randomly selected with replacement from $19,435$ logged trial data. The bandit strategy then picks a treatment X_t for the patient, possibly depending on the context $\boldsymbol{U}_t$. A reward Y_t is then decided following the score function Eq. 22. Note that we make bootstrap estimates of the pretreatment attributes $\boldsymbol{U}$, which allows us to preserve the natural relations among variables.

Details about Contextual Bandits

In contextual bandit settings, we aim to minimize the cumulative regret over two candidate policies $\boldsymbol{\Pi} = \{\pi_0, \pi_1\}$. Table 5 provides the details about these candidate policies.

	$S = 0$	$S = 1$	$E[Y_\pi]$	$[l_\pi, h_\pi]$
π_0	0.05	0.1	0.6277	$[0.5436, 0.6747]$
π_1	0.97	0.99	$\ast$0.6950	$[0.3663, 0.8441]$

Table 5: Parameterizations of the candidate policies $\boldsymbol{\Pi}$ for the contextual bandit task report in Fig. 4(c).

Columns $S = 0$ and $S = 1$ correspond to, respectively, the probabilities $\pi_i(X = 1|S = 0)$ and $\pi_i(X = 1|S = 1)$. Column $E[Y_\pi]$ reports the expected reward of each policy, where the optimal policy π_1 with the largest reward is marked with $\ast$. Column $[\hat{l}_\pi, \hat{h}_\pi]$ represents the causal bounds of each policy. Each empirical bound is computed using all $|\boldsymbol{\mathcal{Z}}| = 4$ sets of confounded observational samples.

We note that in this experiment, for the sub-optimal policy $\pi_0 \neq \pi_1$, its causal bound $\hat{h}_{\pi_0} < \mu_{\pi_1}$, satisfying the improvement condition of Thm. 13. This explains the dominating performance of D-UCB^c reported in the paper. To illustrate the efficiency of our approach, we also include a more involve contextual bandit learning scenario where the improvement condition $\hat{h}_\pi < \mu_{\pi^*}$ is not always satisfied.

A More Involved Contextual Bandits Experiment

Suppose now that the conscious state C of each patient is also observed. We consider the regret minimization in contextual bandits over a set of candidate policies $\boldsymbol{\Pi} = \{\pi_i\}_{i=1}^{6}$. Table 4 provides the detailed descriptions of these candidate policies. Row $S = i, C = j$ shows the probabilities $\pi(X = 1|S = i, C = j)$ for each $\pi \in \boldsymbol{\Pi}$. Row $E[Y_\pi]$ and $[l_\pi, h_\pi]$ represent, respectively, the expected reward and the causal bounds of each policy π. The optimal policy π_4 is marked with $\ast$. Row $h_\pi < \mu_{\pi^*}$ indicates the improvement condition of Thm. 13 (marked ✓ if satisfied).

Similarly, we apply strategies D-UCB^c, D-UCB^c–, D-UCB and D-UCB$^-$ to this contextual bandit instance. Their cumulative regrets (CR) are measured, shown in Fig. 6. We also measure the average regret, which is the ratio between cumulative regret and the number of trials. The analysis reveals a significant difference in CR experienced by D-UCB^c (CR = 44.4) compared to D-UCB (CR = 230.52) and D-UCB^c– (CR = 187.06). D-UCB$^-$ (CR = 318.21) performs worst among all due to the confounding bias. This experiment corroborate with our findings in the paper, showing the

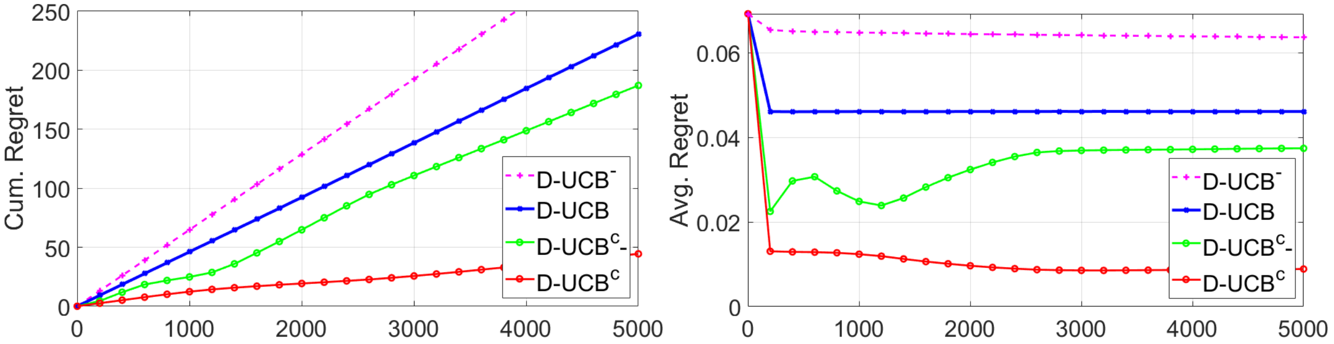


Figure 6: Simulations results comparing solvers that are causally enhanced contextual bandit algorithms (D-UCB^C) and standard (D-UCB), warm-started from observations (D-UCB$^-$) and D-UCB^C with the causal bounds estimated with $n = 200$ observational samples (D-UCB$^C-$).

		π_1	π_2	π_3	π_4	π_5	π_6
$S = 0$	$C = 0$	0.99	0.99	0.01	0.01	0.01	0.99
	$C = 1$	0.01	0.99	0.99	0.01	0.01	0.99
	$C = 2$	0.01	0.01	0.01	0.01	0.01	0.99
$S = 1$	$C = 0$	0.99	0.01	0.01	0.01	0.01	0.01
	$C = 1$	0.01	0.01	0.99	0.99	0.01	0.01
	$C = 2$	0.01	0.01	0.01	0.99	0.01	0.99
$E[Y_\pi]$		0.621	0.619	0.638	$\star 0.691$	0.621	0.678
$[l_\pi, h_\pi]$		$[0.579, 0.655]$	$[0.547, 0.684]$	$[0.467, 0.719]$	$[0.262, 0.775]$	$[0.584, 0.654]$	$[0.07, 0.966]$
$h_\pi < \mu_{\pi^*}$		$\checkmark$	$\checkmark$			$\checkmark$	

Table 4: Details of the candidate policies for the more involved contextual bandit task.

efficiency of the proposed causal approach in more complicated contextual bandit settings.