Structural Causal Bandits with Non-Manipulable Variables

Sanghack Lee, Elias Bareinboim
Department of Computer Science
Purdue University
West Lafayette, IN 47907
{lee2995,eb}@purdue.edu

Abstract

Causal knowledge is sought after throughout data-driven fields due to its explanatory power and potential value to inform decision-making. If the targeted system is well-understood in terms of its causal components, one is able to design more precise and surgical interventions so as to bring certain desired outcomes about. The idea of leveraging the causal understanding of a system to improve decision-making has been studied in the literature under the rubric of *structural causal bandits* (Lee and Bareinboim, 2018). In this setting, (1) pulling an arm corresponds to performing a causal intervention on a set of variables, while (2) the associated rewards are governed by the underlying causal mechanisms. One key assumption of this work is that any observed variable (X) in the system is manipulable, which means that intervening and making $do(X = x)$ is always realizable. In many real-world scenarios, however, this is a too stringent requirement. For instance, while scientific evidence may support that obesity shortens life, it's not feasible to manipulate obesity directly, but, for example, by decreasing the amount of soda consumption (Pearl, 2018). In this paper, we study a relaxed version of the structural causal bandit problem when not all variables are manipulable. Specifically, we develop a procedure that takes as argument partially specified causal knowledge and identifies the possibly-optimal arms in structural bandits with non-manipulable variables. We further introduce an algorithm that uncovers non-trivial dependence structure among the possibly-optimal arms. Finally, we corroborate our findings with simulations, which shows that MAB solvers enhanced with causal knowledge and leveraging the newly discovered dependence structure among arms consistently outperform causal-insensitive solvers.

Introduction

Causal knowledge is deemed highly valuable and desirable across a wide range of disciplines, including in the sciences, engineering, and businesses. The reasons for such a prominence are various. First, causal knowledge entails explanatory power, which epitomizes the very goal of science — i.e., opening Nature's "black box" and explaining how reality "works," or how it could be understood in terms of more elementary, and interpretable components. Second, and more pragmatically, purposeful agents (AI systems, policy-makers, physicians) are better equipped for designing cleaner, more precise, and more effective interventions once the underlying causal mechanisms are well-understood (Bareinboim and Pearl 2016; Pearl and Mackenzie 2018; Lee and Bareinboim 2018).

For instance, economists strive to understand the *root causes* of poverty, which could allow the design of new policies (i.e., causal interventions) to improve the population's socioeconomic status (SES). A considerable body of evidence was accumulated for many decades, notably by the University of Chicago's Professor, and Nobel Prize laureate, James Heckman, who demonstrated the effects of early education on families' SES, among other indicators (Heckman 2006). The understanding following this causal link translated to the larger support of early childhood education, and a push for new policies; see, for an example, Obama's one billion dollar investment (The White House, Office of the Press Secretary 2014). There are abundant of such cases in public health as well — e.g., evidence supports that tobacco smoking is one of the determinant factors of lung cancer (Cornfield 1951; U.S. Department of Health and Human Services 2014), or obesity is responsible to shortening life expectancy (Flegal, Graubard, and Williamson 2005; Pearl 2018).

Despite all the impressive results achieved so far, the treatment of how to use causal knowledge to support decision-making is still in its infancy. If our goal is to build a more automated, data-driven society, and intelligent systems that can reason and act autonomously, we need to move from a heuristic understanding of the interplay between causal knowledge and decision-making (which currently relies on the insights of highly skilled scientists) to a more fundamental understanding of the principles that allow one to translate causal evidence into a more robust decision-making process.

In order to realize this goal, we start by highlighting the rich literature on automated decision-making in AI, including the setting known as *multi-armed bandits* (MABs, for short) (Robbins 1952; Lai and Robbins 1985). We'll use the MAB framework as the baseline of our analysis. In a typical MAB instance, there are multiple arms (or actions) to play at each time step. Pulling an arm at each step returns a stochastic reward from the underlying, unknown distribution. Algorithms for the bandit problem attempt to minimize the cumulative regret, while having to cope with an exploration and exploitation trade-off. This setting has been popular in experimental circles since many real-world situations can be cast as MAB instances, including clinical trials, online advertisement placement, network routing, news article recommendation, just to cite a few. Remarkably, many algorithms with attractive properties have been developed, including the celebrated *Upper Confidence Bound* (Garivier and Cappé 2011; Auer, Cesa-Bianchi, and Fischer 2002; Cappé et al. 2013) and *Thompson sampling* (Thompson 1933; Chapelle and Li 2011; Kaufmann, Korda, and Munos 2012). One critical assumption used throughout these algorithms, and their multiple extensions, is that the arms are usually considered independent. Researchers have started to explore the dependence across arms to make these solvers applicable to a broader set of applications where the independence across arms is clearly unattainable (Dani, Hayes, and Kakade 2008; Magureanu, Combes, and Proutiere 2014).

Even though not explicitly acknowledged until recently, MAB solvers inherently estimate causal effects, or properties of these effects. To witness, note that (1) pulling an arm corresponds to intervening on a set of variables and setting them to specific values (from whatever *natural* regime the decision used to be based on, to a new compulsory policy determined by the solver); and, (2) the rewards associated with these arms are governed by the underlying causal mechanisms (Bottou et al. 2013; Bareinboim, Forney, and Pearl 2015; Lattimore, Lattimore, and Reid 2016; Zhang and Bareinboim 2017; Forney, Pearl, and Bareinboim 2017; Sen et al. 2017).

Recently, Lee and Bareinboim (2018) combined these observations (i.e., that MABs are inherently causal and that the dependence across arms carries decision-making value) and introduced what they called the *structural causal bandits* problem (SCM-MAB), where a Structural Causal Model (SCM) (Pearl 2000) is used to describe the underlying causal system. The dependency among arms naturally arises due to the causal relationships among the endogenous and exogenous variables. Given partial knowledge about the underlying SCM, the authors characterized two structural properties computable from any SCM-MAB, i.e., i) arms with the same reward distribution using *do*-calculus constraints, and ii) under what topological conditions one arm can be optimal.

For concreteness, we consider the SCM-MAB instance shown in Fig. 1, where Y represents the reward, X and Z represent two variables that can be manipulated, and the dashed-bidirected arrow, following standard notation, represents the existence of a latent common cause (exogenous) affecting both X and Y. There are four sets of variables that can be intervened upon in this case (intervention sets), i.e., $\{\emptyset, \{X\}, \{Z\}, \{X, Z\}\}$, which totals 9 arms if we assume binary variables. The first property reveals that $P(y|do(z)) = P(y|do(x, z))$, which means that intervening on X and Z simultaneously is regarded as wasteful — one should always prefer intervening on $do(Z)$ over $do(X, Z)$. Intuitively, all effects of X on Y are mediated by Z in this case, which means X has no effect on Y whenever we intervene on Z. The second property characterizes that $\max_z \mu_z \geq \max_x \mu_x$ where $\mu_x = \mathbb{E}[Y|do(x)]$ holds true in *any* model, which implies that $do(Z)$ should be *always* preferred over $do(X)$. This implies that three arms, $do(\emptyset)$, $do(z = 0)$, and $do(z = 1)$, relating to the non-dominated, minimal intervention sets, $\emptyset$ and $\{Z\}$, are contenders to be optimal relative to the underlying SCM compatible with $\mathcal{G}$.[1] Such a minimal set that can be optimal when intervened upon was termed a *possibly-optimal minimal intervention set*, or POMIS, in (Lee and Bareinboim 2018).

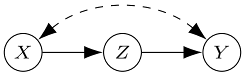


Figure 1: A causal model where Y is the outcome variable.

In many settings, however, it is not the case that every observed variable is manipulable. For the graph discussed earlier, imagine that X, Z, and Y correspond to diet, cholesterol level, and heart failure. Changing the cholesterol directly is not a conceivable physical manipulation, despite the obvious effect of cholesterol in heart failure. Given that Z is not manipulable, then intervening on X, e.g., diet, may lead to the desired outcome. Clearly, POMISs under non-manipulable variables will not be, in general, a subset of the POMISs with only manipulable variables. We'll show in the paper that $do(X)$, which was dominated by $do(Z)$, should be regarded as a POMIS whenever Z is not manipulable. It is critical, therefore, to be able to identify the POMISs from a causal graph under an arbitrary set of non-manipulable variables.

Previous work considered identifying the POMISs based on the graph structure, which allowed a more precise decision-making by the corresponding MAB solver. Still, the dependence structure across arms was not fully exploited. More sophisticated relations among arms can be learned through the *do*-calculus (Pearl 2000, Ch. 3). For instance, the identity $P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x')$ holds in Fig. 1, which is known as the 'front-door' formula. The front-door implies that the distribution of the arm $do(X = x)$ can be learned from $do(\emptyset)$, i.e., without having to directly intervene on X. More generally, an interventional distribution can be expressed in terms of the other distributions, which implies new learning opportunities we aim to leverage.

Specifically, our contributions are as follow:

1. We start by formalizing the SCM-MAB problem under non-manipulability constraints using the language of structural causality. We formally characterize POMIS with manipulability constraints, i.e., the collection of interventions that are possibly optimal.
2. We develop a new algorithm that derives an expression for the arm's distribution given an arbitrary set of interventional distributions (other arms). We use this result to enhance standard MAB algorithms with causal capabilities, which will now exploit the structural properties of the underlying causal graph and the relations across interventional distributions, improving their efficiency.

Simulations corroborate our intuition — solvers that leverage causal knowledge can operate orders of magnitude more data-efficiently than their non-causal counterparts.

[1] Note that this implies that not intervening, and letting the system operate in its natural state, can yield an optimal reward.

Preliminaries

We follow the convention in the field and use a capital letter as a variable, X, and the corresponding lower case, x, as its realization. Bold face is used for a set of variables or values, $\mathbf{X}$ or $\mathbf{x}$, blackboard bold for a set of sets, $\mathbb{X}$. We denote by $\mathfrak{X}_X$ the domain of X. We also use caligraphies for graphs and models. The basic semantical framework of our analysis relies on Structural Causal Models (SCM, Pearl, 2000), i.e.,

Definition 1 (SCM)**.** A structural causal model (SCM) $\mathcal{M}$ is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where:

- $\mathbf{U}$ is a set of exogenous (unobserved) variables, which are determined by factors outside of the model;
- $\mathbf{V}$ is a set $\{V_i\}_{i=1}^{n}$ of endogenous (observed) variables that are determined by variables in $\mathbf{U} \cup \mathbf{V}$;
- $\mathbf{F}$ is a set of structural functions $\{f_i\}_{i=1}^{n}$ where each f_i is a process by which V_i is assigned a value $v_i \leftarrow f_i(\mathbf{pa}^i, \mathbf{u}^i)$ in response to the current values of $\mathbf{PA}^i \subset \mathbf{V}$ and $\mathbf{U}^i \subseteq \mathbf{U}$;
- $P(\mathbf{U})$ is a distribution over the exogenous variables $\mathbf{U}$.

A SCM induces observational and interventional probability distributions, $P(\mathbf{v}) = \sum_{\mathbf{u}} \prod_i P(v_i | \mathbf{pa}^i, \mathbf{u}^i) P(\mathbf{u})$, and $P(\mathbf{v} \backslash \mathbf{t} | do(\mathbf{t})) = \sum_{\mathbf{u}} \prod_{i | V_i \notin \mathbf{T}} P(v_i | \mathbf{pa}^i, \mathbf{u}^i) P(\mathbf{u})$. For short, we write $P(\mathbf{y} | do(\mathbf{x}))$ as $P_{\mathbf{x}}(\mathbf{y})$.

Each SCM is associated with a causal graph $\mathcal{G} = \langle \mathbf{V}, \mathbf{E} \rangle$, where there are two types of edges in $\mathbf{E}$ — directed edges such as $V_i \rightarrow V_j$ indicating direct functional dependence, i.e., V_i is used to define $f_j \in \mathbf{F}$; bidirected edges such as $V_i \leftrightarrow V_j$ indicating the existence of an unobserved confounder (UC, for short), i.e., there exists $U \in \mathbf{U}$ which appears in both f_i and f_j. We use family relations, pa, ch, an, de to denote parents, children, ancestors, and descendants of its argument. With Pa, Ch, An, De, we include the arguments, e.g., $An(W) = an(W) \cup \{W\}$. Given a set of variables as argument, we use the union of individual outputs, e.g., $An(\mathbf{W}) = \bigcup_{W \in \mathbf{W}} An(W)$. Note that $pa(V_i) = \mathbf{PA}^i$. We denote by $\mathcal{G}_{\overline{\mathbf{X}}}$ a subgraph where edges onto $\mathbf{X}$ are removed.

The *do*-calculus (Pearl 1995) consists of three rules relating observational and interventional distributions that obey the topology of the causal graph $\mathcal{G}$. For example, $P_{x,z}(y) = P_z(y)$ (Fig. 1) can be inferred due to $(Y \perp\!\!\!\perp X | Z)_{\mathcal{G}_{\overline{X,Z}}}$, which is a *d*-separation statement between Y and X given Z in the subgraph $\mathcal{G}_{\overline{X,Z}}$. We refer readers to (Pearl 2000, Sec. 3.4) for a more detailed discussion on SCMs and *do*-calculus.

A K-armed bandit problem consists of K *independent* arms such that their reward distributions are independent to each other. In this setting, the task is often minimizing Reg_T, the cumulative regret after T rounds. This is the difference between a maximum expected cumulative reward and an expected cumulative reward of a MAB algorithm, $\text{Reg}_T = T\mu^* - \sum_{t=1}^{T} \mathbb{E}[Y_{a_t}]$, where a_t is the arm played at time t, Y_{a_t} is a random variable associated with the arm, and μ^* is the optimal expected reward. In a SCM-MAB setting, each arm corresponds to intervening on a set of variables. Given $\mathcal{G}$ and Y, *all* arms are $\{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}} \mid \mathbf{X} \subseteq \mathbf{V} \backslash \{Y\}\}$ with $P(Y_{\mathbf{x}})$, the distribution of a reward variable with arm $\mathbf{x}$, is equivalent to an interventional distribution $P_{\mathbf{x}}(Y)$, and $\mu_{\mathbf{x}} = \mathbb{E}\left[Y | do(\mathbf{x})\right]$. We denote by $\mu_{\mathbf{x}^*}$ the best expected reward by intervening on $\mathbf{X}$, i.e., $\mu_{\mathbf{x}^*} = \max_{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}}} \mu_{\mathbf{x}}$.

SCM-MAB with Non-manipulability

In this section, we generalize SCM-MABs by allowing non-manipulability constraints, and then characterize the POMIS under such constraints. We start by denoting a set of non-manipulable variables by $\mathbf{N} \subseteq \mathbf{V} \setminus \{Y\}$, where the reward variable Y is assumed to be non-manipulable. For simplicity, we override the definitions of MIS and POMIS by incorporating $\mathbf{N}$ into them, and limiting the intervention set $\mathbf{X}$ to be $\mathbf{X} \subseteq \mathbf{V} \backslash \{Y\} \backslash \mathbf{N}$ instead of $\mathbf{X} \subseteq \mathbf{V} \backslash \{Y\}$. First, we define Minimal Intervention Sets, which represent a non-redundant set of intervention sets.

Definition 2 (Minimal Intervention Set (MIS))**.** Given $\langle \mathcal{G}, Y, \mathbf{N} \rangle$, a set of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\} \setminus \mathbf{N}$ is said to be a *minimal intervention set* if there is no $\mathbf{X}' \subset \mathbf{X}$ such that $\mu_{\mathbf{x}'} = \mu_{\mathbf{x}}$ for every SCM conforming to $\mathcal{G}$ where $\mathbf{x}' \in \mathfrak{X}_{\mathbf{X}'}$ that is consistent with $\mathbf{x}$.

We denote by $\mathbb{M}_{\mathcal{G},Y}^{\mathbf{N}}$ a set of MISs given $\langle \mathcal{G}, Y, \mathbf{N} \rangle$ where we omit $\mathbf{N}$ if $\mathbf{N} = \emptyset$. The following relationship can be easily derived from the definition, $\mathbb{M}_{\mathcal{G},Y}^{\mathbf{N}} = \{\mathbf{W} \in \mathbb{M}_{\mathcal{G},Y} \mid \mathbf{W} \cap \mathbf{N} = \emptyset\}$. MISs can be identified by checking whether any two probability distributions are equal (through Rule 3 of *do*-calculus). Next, we define a subset of the MISs that can lead to an optimal reward.

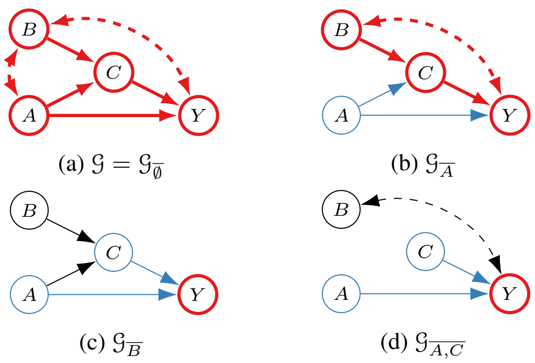
Definition 3 (Possibly-Optimal Minimal Intervention Set (POMIS))**.** Given $\langle \mathcal{G}, Y, \mathbf{N} \rangle$, let $\mathbf{X} \in \mathbb{M}_{\mathcal{G},Y}^{\mathbf{N}}$. If there exists a SCM conforming to $\mathcal{G}$ such that $\mu_{\mathbf{x}^*} > \forall_{\mathbf{W} \in \mathbb{M}_{\mathcal{G},Y}^{\mathbf{N}} \backslash \{\mathbf{X}\}} \mu_{\mathbf{w}^*}$, then $\mathbf{X}$ is a *possibly-optimal minimal intervention set* with respect to $\langle \mathcal{G}, Y, \mathbf{N} \rangle$.

We similarly denote by $\mathbb{P}_{\mathcal{G},Y}^{\mathbf{N}}$, a set of POMISs given $\langle \mathcal{G}, Y, \mathbf{N} \rangle$. Lee and Bareinboim (2018) introduced two key concepts that will be instrumental to our analysis, i.e., *minimal unobserved confounders' territory* (MUCT) and *interventional border* (IB). MUCT is a minimal set of variables in $An(Y)$, which includes Y and is closed under descendants and unobserved confounders given a causal graph. For example, the MUCT in Fig. 2a includes all the variables as B is confounded with Y, A is confounded with B, and C is a descendant of B (or A). The IB is defined as the parents of the MUCT excluding the MUCT itself, which is an empty set in this case. With $\mathcal{G}_{\overline{A}}$, the MUCT is $\{B, C, Y\}$ and the IB is $\{A\}$ as shown in Fig. 2b. It was shown that $\mathbf{X} \subseteq \mathbf{V} \backslash \{Y\}$ is a POMIS if and only if IB in $\mathcal{G}_{\overline{\mathbf{X}}}$ is $\mathbf{X}$. For example, $\{A, C\}$ is a POMIS in Fig. 2d while $\{B\}$ is not in Fig. 2c, which indicates that intervening on $\{A, C\}$ is preferred to $\{B\}$.

Characterizing POMIS with Non-manipulability

Identifying all the POMISs given non-manipulable variables is non-trivial as the unconstrained POMIS (i.e., $\mathbf{N} = \emptyset$) cannot be, in general, applicable by filtering out intervention sets containing only manipulable variables, i.e., $\exists_{\mathcal{G},Y,\mathbf{N}}\ \mathbb{P}_{\mathcal{G},Y}^{\mathbf{N}} \neq \{\mathbf{X} \in \mathbb{P}_{\mathcal{G},Y}^{\emptyset} \mid \mathbf{X} \cap \mathbf{N} = \emptyset\}$. This contrasts with MISs in which a set of constrained MISs is a mere subset of unconstrained MISs. Although the equality does not hold in general, the feasible subset of unconstrained POMISs is related to POMIS with constraints as shown below:

Proposition 1. *If* $\mathbf{X} \in \mathbb{P}_{\mathcal{G},Y}$ *and* $\mathbf{X} \cap \mathbf{N} = \emptyset$*, then* $\mathbf{X} \in \mathbb{P}_{\mathcal{G},Y}^{\mathbf{N}}$.



(a) $\mathcal{G} = \mathcal{G}_{\overline{\emptyset}}$ (b) $\mathcal{G}_{\overline{A}}$ (c) $\mathcal{G}_{\overline{B}}$ (d) $\mathcal{G}_{\overline{A,C}}$

Figure 2: MUCT (thick, red) and IB (blue) on manipulated graphs. Given $\mathcal{G}$ and Y, POMISs are $\emptyset$, $\{A\}$, and $\{A, C\}$.

Simply put, if an arm $do(\mathbf{x})$ is optimal among all arms without constraints, then the arm is still optimal among the manipulable subset of arms. However, this proposition only answers the intersection between $\mathbb{P}_{\mathcal{G},Y}$ and $\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}$. At this point, the natural question is how to systematically find the POMISs that are not unconstrained POMISs. Recall that in the front-door graph, we noted that $\{X\}$ is a POMIS when $\{Z\}$ is non-manipulable, while this isn't the case in the unconstrained setting. To explain this phenomenon, we note that X, which affects Y through Z, can be seen as directly connected to Y in the constrained setting; clearly, Z should be disregarded since it no longer can disturb the pathway from X to Y.

Focusing only on a subset of variables in a causal graph and their causal relationship lies at the heart of causal modeling. Latent projection (projection, for short) (Verma and Pearl 1990; Verma 1992) is an operation that induces a causal graph over a subset of the variables from the original graph while maintaining some key topological properties of the original causal structure (Tian 2002). In general, the causal graph at hand is already the result of the projection of some unknown, more involved causal structure. For example, the front-door graph (Fig. 1) with the cholesterol example might be the projection of a larger graph such as in Fig. 3, where the shaded nodes were marginalized out. At an intuitive level, without constraints, POMISs in the front-door graph can be understood as POMISs in the larger graph where the shaded nodes are not manipulable. It would be natural to ask whether we can obtain $\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}$ indirectly from $\mathbb{P}_{\mathcal{H},Y}$, where $\mathcal{H}$ is the projection of $\mathcal{G}$ onto $\mathbf{V}\setminus\mathbf{N}$.

We formally describe how a projection of a causal graph $\mathcal{G}$ onto $\mathbf{V}\setminus\mathbf{N}$ is constructed next. Let $\hat{\mathcal{G}}$ be a DAG that explicitly represents the unobserved confounders. We initialize a graph $\mathcal{H} = \langle \mathbf{V}\setminus\mathbf{N}, \emptyset\rangle$, then proceed to add

1. a directed edge between V_i and V_j if $V_i \rightarrow V_j \in \mathcal{G}$ or there exists a directed path from V_i to V_j where all non-end nodes in the path between them are in $\mathbf{N}$.
2. a bidirected edge between V_i and V_j if $V_i \leftrightarrow V_j \in \mathcal{G}$; or there exists directed paths from an unobserved confounder to V_i and V_j in $\hat{\mathcal{G}}$ where all non-end nodes are in $\mathbf{N}$.

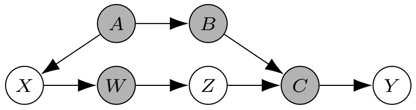


Figure 3: A causal graph where its projection onto $\{X, Y, Z\}$ yields the front-door graph (Fig. 1).

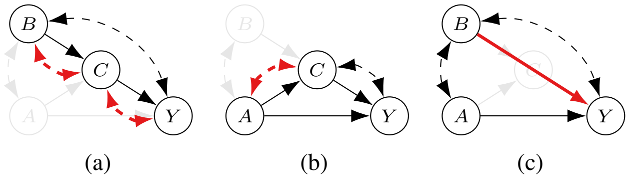


Figure 4: Causal graphs after marginalizing out A, B, and C from Fig. 2a, respectively, with added edges in thick, red.

Let $\mathcal{G}_{[\mathbf{V}']}$ be a causal graph obtained by projecting $\mathcal{G}$ onto $\mathbf{V}'$. We will show $\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y} = \mathbb{P}_{\mathcal{H},Y}$ through two propositions that guarantees that the optimality of an arm will be preserved during i) the projection from $\mathcal{G}$ to $\mathcal{H}$, and ii) the other way around.

Proposition 2. *Given a SCM $\mathcal{M}^1 = \mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{U})\rangle$, there exists a SCM $\mathcal{M}^2 = \mathcal{M}_{[\mathbf{V}\setminus\mathbf{N}]} = \langle \mathbf{V}\setminus\mathbf{N}, \mathbf{U}, \mathbf{F}', P(\mathbf{U})\rangle$ such that $P^1_{\mathbf{x}}(\mathbf{y}){=}P^2_{\mathbf{x}}(\mathbf{y})$ for any $\mathbf{X}, \mathbf{Y}{\subseteq}\mathbf{V}\setminus\mathbf{N}$ and $\mathbf{Y} \neq \emptyset$.*

Proof. Let $\mathcal{M}' = \mathcal{M}_{[\mathbf{V}\setminus\{W\}]}$ where $W \in \mathbf{N}$. The functions $\{f_X\}_{X\notin ch(W)_{\mathcal{G}}}$ can be used without modifications in $\mathcal{M}'$. We modify functions of the children of W. For every $Q \in ch(W)_{\mathcal{G}}$, we devise f'_Q with f_Q and f_W. By the projection's construction, $\mathbf{PA}'^Q = (\mathbf{PA}^Q\setminus\{W\})\cup\mathbf{PA}^W$ and let $\mathbf{U}'^Q = \mathbf{U}^Q \cup \mathbf{U}^W$ so that f'_Q takes $\mathbf{pa}^W$ and $\mathbf{u}^W$ as argument in addition to $\mathbf{pa}^Q\setminus\{W\}$ and $\mathbf{u}^Q$, which conforms to $\mathcal{G}_{[\mathbf{V}\setminus\{W\}]}$. Hence, f'_Q is simply f_Q with w computed inside,

$$f'_Q\left(\mathbf{pa}'^Q, \mathbf{u}'^Q\right) = f_Q\left(\mathbf{pa}^Q\setminus\{W\}, f_W(\mathbf{pa}^W, \mathbf{u}^W), \mathbf{u}^Q\right).$$

This guarantees that $\mathcal{M}_{[\mathbf{V}\setminus\{W\}]}$ and $\mathcal{M}$ yield the same observational and interventional distributions over $\mathbf{V}\setminus\{W\}$. The above procedure can be iteratively applied to the rest of variables in $\mathbf{N}$ to prove the equivalence of $\mathcal{M}_{[\mathbf{V}\setminus\mathbf{N}]}$ and $\mathcal{M}$ with respect to the distributions they yield over $\mathbf{V}\setminus\mathbf{N}$. □

Fig. 4a is an example where marginalizing out A from Fig. 2a induces bidirected edges between B and C (due to the UC between A and B and $A \rightarrow C$), between C and Y since $C{\leftarrow}A{\rightarrow}Y$.[2] Based on Prop. 2, C and Y take the UC between A and B to compute a (the value A would take) inside. This explains all three bidirected edges in Fig. 4a.

Proposition 3. *Given a causal diagram $\mathcal{G}$, let $\mathcal{H} = \mathcal{G}_{[\mathbf{V}\setminus\mathbf{N}]}$. For a SCM $\mathcal{M}^1 = \mathcal{M}_{[\mathbf{V}\setminus\mathbf{N}]} = \langle \mathbf{V}\setminus\mathbf{N}, \mathbf{U}, \mathbf{F}', P(\mathbf{U})\rangle$ conforming to $\mathcal{H}$, there exists a SCM $\mathcal{M}^2 = \mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{U})\rangle$ that conforms to $\mathcal{G}$ such that $P^1_{\mathbf{x}}(\mathbf{y}) = P^2_{\mathbf{x}}(\mathbf{y})$, for any $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}\setminus\mathbf{N}$ and $\mathbf{Y} \neq \emptyset$.*

[2] Although $B{\leftrightarrow}A{\rightarrow}Y$ implies a bidirected edge between B and Y, we do not represent multiple edges of the same type.

Proof. Consider a simple case where constructing $\mathcal{M}$ of $\mathcal{G}$ from $\mathcal{M}'$ of $\mathcal{G}_{[\mathbf{V}\setminus\{W\}]}$. Without loss of generality, assume that $U \in \mathbf{U}$ appears in at most two functions. Otherwise, we can decompose U into multiple U's and change $\mathbf{U}$ and $P(\mathbf{U})$ accordingly. We similarly reuse f'_X for any $X \notin ch(W)_{\mathcal{G}}$ to define $\mathbf{F}$. We need to define f_W and f_Q for $Q \in \mathbf{Q} = ch(W)_{\mathcal{G}}$. By definition, endogenous variables to be used in f_W and $\{f_Q\}_{Q\in\mathbf{Q}}$ are $pa(W)_{\mathcal{G}}$ and $pa(Q)_{\mathcal{G}}$, respectively. Next, we determine $\mathbf{U}^i$ for $V \in \mathbf{Q} \cup \{W\}$ as follows. First, any $U \in \mathbf{U}$ that appears only in a variable V (i.e., $U \in \mathbf{U}'^V$) will be in $\mathbf{U}^V$. Now consider U used in two variables, V_i and V_j, in $\mathcal{G}_{[\mathbf{V}\setminus\{W\}]}$. If a bidirected edge between V_i and V_j exists in $\mathcal{G}$, every $U \in \mathbf{U}'^i \cap \mathbf{U}'^j$ will also be included in $\mathbf{U}^i$ and $\mathbf{U}^j$. Otherwise, there are two (non-mutually exclusive) cases: i) both V_i and V_j are children of W in $\mathcal{G}$ or ii) V_i is confounded with W and V_j is a child of W in $\mathcal{G}$ (or vice versa). In both cases, simply put U into $\mathbf{U}^W$. This construction guarantees that all $\mathbf{U}$ are modeled in $\mathcal{M}$ conforming to $\mathcal{G}$.

Then, we define f_W be any one-to-one function so that its argument can be passed to its children through $f_W^{-1}(w)$. Finally, f_Q is devised to reuse f'_Q with f_W^{-1} as

$$f_Q\left(\mathbf{pa}^Q, \mathbf{u}^Q\right) = f'_Q\left(\mathbf{pa}^Q \setminus \{W\}, f_W^{-1}(w), \mathbf{u}^Q\right).$$

The procedure can be sequentially applied to construct $\mathcal{M}$ for $\mathcal{G}$ from $\mathcal{M}'$ for $\mathcal{H}$. □

We now revisit Fig. 4a for another example. We focus on how the unobserved variables will be assigned. Let UCs be U_{BC}, U_{CY}, and U_{BY}. Unobserved variables other than UCs will be reused as described above. First, $f_B = f'_B$ which already takes U_{BC} and U_{BY}. f_Y takes U_{BY} since $B \leftrightarrow Y$ in $\mathcal{G}$. Since two red bidirected edges do not appear in $\mathcal{G}$, f_A will take both U_{BC} and U_{CY}, and f_C and f_Y become irrelevant to two red UCs. Then, we can observe that A is confounded with B via U_{BC}, and B and Y are confounded with U_{BY}. However, U_{CY} will be served as a variable-specific hidden variable for A, which we do not explicitly represent. The construction conforms to $\mathcal{G}$.

Theorem 4. *Given a causal diagram $\mathcal{G} = \langle\mathbf{V}, \mathbf{E}\rangle$, a reward variable $Y \in \mathbf{V}$, and a set of non-manipulable variables $\mathbf{N} \subseteq \mathbf{V}\setminus\{Y\}$, let $\mathcal{H}$ be the projection of $\mathcal{G}$ onto $\mathbf{V}\setminus\mathbf{N}$. Then,*

$$\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y} = \mathbb{P}_{\mathcal{H},Y}$$

Proof. This follows from Prop. 2 and Prop. 3. □

In words, the projection onto the manipulable variables preserves the causal structure among them with respect to POMIS. Fig. 4 illustrates three projected graphs of $\mathcal{G}$ in Fig. 2a with $\{A\}$, $\{B\}$, and $\{C\}$ marginalized out, respectively. Applying the (unconstrained) POMIS identification procedure (Lee and Bareinboim 2018) on the graphs yields: $\mathbb{P}^{\{A\}}_{\mathcal{G},Y} = \{\emptyset, \{B\}, \{C\}\}$, $\mathbb{P}^{\{B\}}_{\mathcal{G},Y} = \{\emptyset, \{A\}, \{A, C\}\}$, and $\mathbb{P}^{\{C\}}_{\mathcal{G},Y} = \{\emptyset, \{A\}, \{A, B\}\}$.

This section studied the problem of which intervention sets can be optimal. Once the POMIS under constraints are formally characterized, the arms to be played by a MAB algorithm can be refined and its performance improved.

Exploiting the POMISs Structure

In this section, we note that arms exhibited interesting dependence structure that we will try to exploit, which follows from the underlying causal structure. The previous section investigated two types of dependence among arms (equality and partial-orderedness), which helped identifying POMIS arms, $\{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}} | \mathbf{X} \in \mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}\}$. In this section, we explore more sophisticated dependence among arms that allows one arm's reward distribution to be inferred from pulling another arm. Following the convention, we note that when an arm $\mathbf{x}$ is pulled, a full realization of the other variables in the system is observed, e.g., $\mathbf{v} = \langle v_1, v_2, \ldots, v_n\rangle$, which is sampled from the joint distribution $P_{\mathbf{x}}(\mathbf{v})$ associated with the pulled arm.

Recall the causal model in Fig. 1 when Z is non-manipulable. There are three POMIS arms $do(\emptyset)$, $do(x{=}0)$, and $do(x{=}1)$, where the two $do(x)$ arms' reward distributions can be expressed as $P_x(y) = \sum_z P(z|x) \sum_{x'} P(y|z, x')P(x')$, which, as noted, is called the front-door formula. Given such an expression, $P_x(y)$ can be estimated not only with a trivial expression $P_x(y) = \sum_{\mathbf{v}\setminus\{y\}} P_x(\mathbf{v})$, from the experimental samples obtained by pulling the arm $do(x)$, but also with the front-door formula with observational samples from $P(\mathbf{v})$, in this case, by playing the do-nothing arm. This possibility of leveraging samples from other arms is not an exclusive phenomenon to the model in Fig. 1, but can be cast in much broader terms. Given a general graph $\mathcal{G}$, under what conditions can a probability term (a reward distribution, in particular) be turned into an expression with terms of other distributions?

In the causal inference literature, the problem of *identifiability* (Pearl 2000; Tian and Pearl 2002) asks whether the experimental distribution of interest can be uniquely estimated from the observational distribution (do-nothing intervention). Given a causal graph and $P(\mathbf{V})$, there exists a complete algorithm which outputs an expression, if exists, for a quantity of interest, $P_{\mathbf{x}}(\mathbf{y})$ in this case. A more general problem, called z-identifiability (Bareinboim and Pearl 2012), asks whether one can identify $P_{\mathbf{x}}(\mathbf{y})$ when a set of experiments (i.e., distributions) are available, $\{P_{\mathbf{Z}'} \mid \mathbf{Z}' \subseteq \mathbf{Z}\}$, for some $\mathbf{Z} \subseteq \mathbf{V}$. That is, there exists a set of manipulable variables $\mathbf{Z}$, and experiments on every subset of manipulable variables are available.

In the SCM-MAB setting with non-manipulability constraints, it is natural to take advantage of z-identifiability where $\mathbf{Z} = \mathbf{V}\setminus\{Y\}\setminus\mathbf{N}$. Although it is feasible, in a sense, that the resulting expression will not include any of $\mathbf{N}$, the expression may include terms related to experiments on non-POMISs. Consider Fig. 4a for an example: an expression containing $P_{b,c}$ will not be useful since arms $\{do(b, c)\}_{b\in\mathfrak{X}_B, c\in\mathfrak{X}_C}$ will not be played at all since $\{B, C\} \notin \mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}$. Hence, there exists a need to extend the treatment given to z-identifiability to account for an arbitrary set of experiments, $\{P_{\mathbf{Z}}\}_{\mathbf{Z}\in\mathbb{Z}}$, where $\mathbb{Z}$ is a subset of the superset of $\mathbf{V}$. We developed a generalized z-identifiability algorithm (Alg. 1), which we call $\mathsf{z}^2\mathsf{ID}$. Using this algorithm, we can identify whether an arm's reward distribution can be estimated from the other arms' information. Before explaining the algorithm, we introduce a few notations. We denote

Algorithm 1 z^{2}ID

```
1:  function z$^2$ID($\mathcal{G}, \mathbf{y}, \mathbf{x}, \mathbb{Z}$)
2:      if $\emptyset \notin \mathbb{Z}$ and $\mathbf{X} = \emptyset$ then
3:          return $\sum_{\mathbf{v}\setminus\mathbf{y}} \prod_{\mathbf{S}_i \in \mathsf{CC}(\mathcal{G})}$ sub-z$^2$ID($\mathcal{G}, \mathbf{s}_i, \mathbf{v}\setminus\mathbf{s}_i, \mathbb{Z}$)
4:      else return sub-z$^2$ID($\mathcal{G}, \mathbf{y}, \mathbf{x}, \mathbb{Z}$)
5:  function SUB-Z$^2$ID($\mathcal{G}, \mathbf{y}, \mathbf{x}, \mathbb{Z}$)
6:      for $\mathbf{Z} \in \mathbb{Z}$ s.t. $\mathbf{Z} \subseteq \mathbf{X}$ and $\neg\exists_{\mathbf{Z}' \in \mathbb{Z}} \mathbf{Z} \subset \mathbf{Z}' \subseteq \mathbf{X}$ do
7:          return ID($\mathcal{G} \setminus \mathbf{Z}, \mathbf{y}, \mathbf{x} \setminus \mathbf{Z}, P_{\mathbf{z}}$) if not FAIL
8:      if $\mathbf{V} \setminus An(\mathbf{Y})_{\mathcal{G}} \neq \emptyset$ then
9:          return sub-z$^2$ID($\mathcal{G}[An(\mathbf{Y})_{\mathcal{G}}], \mathbf{y}, \mathbf{x} \cap An(\mathbf{Y})_{\mathcal{G}}$,
                  $\{\mathbf{Z} \cap An(\mathbf{Y})_{\mathcal{G}}\}_{\mathbf{Z} \in \mathbb{Z}}$)
10:     if $\left(\mathbf{W} \leftarrow (\mathbf{V}\setminus\mathbf{X})\setminus An(\mathbf{Y})_{\mathcal{G}_{\overline{\mathbf{X}}}}\right) \neq \emptyset$ then
11:         return sub-z$^2$ID($\mathcal{G}, \mathbf{y}, \mathbf{x} \cup \mathbf{w}, \mathbb{Z}$)
12:     if $|\mathsf{CC}(\mathcal{G} \setminus \mathbf{X})| > 1$ then
13:         return $\sum_{\mathbf{v}\setminus(\mathbf{x}\cup\mathbf{y})} \prod_{\mathbf{S}_i \in \mathsf{CC}(\mathcal{G}\setminus\mathbf{X})}$ sub-z$^2$ID($\mathcal{G}, \mathbf{s}_i, \mathbf{v}\setminus\mathbf{s}_i, \mathbb{Z}$)
14:     return FAIL
```

by $\mathcal{G}[\mathbf{W}]$, a vertex-induced subgraph of $\mathcal{G}$ by $\mathbf{W}$ (not to confuse with $\mathcal{G}_{[\mathbf{V}']}$). Also, $\mathcal{G}\setminus\mathbf{W} = \mathcal{G}[\mathbf{V}\setminus\mathbf{W}]$. $\mathsf{CC}(\mathcal{G})$ represents C-components (Tian and Pearl 2002) of $\mathcal{G}$, which is a family of sets of endogenous variables where each set consists of variables that are maximally connected via bidirected edges.

z^{2}ID first examines whether the query of interest is an observational quantity where only interventional distributions are available (Line 2). Then, the procedure tries to identify the quantity by decomposing it into multiple interventional quantities using C-component factorization (Tian and Pearl 2002) (Line 3). This part, especially, differs from variants of identifiability algorithms where an observational distribution $P(\mathbf{V})$ is taken for granted. Then, z^{2}ID recursively transforms the query $P_{\mathbf{x}}(y)$ into different forms using well-known equalities that can be derived from basic probability operations and *do*-calculus, and it tries to reduce $P_{\mathbf{x}'}(y)$ to an expression only made of $P_{\mathbf{z}}$ whenever $\mathbf{Z} \subseteq \mathbf{X}'$ and $\mathbf{Z} \in \mathbb{Z}$. This can be understood as solving an identifiability problem where the query is $Q_{\mathbf{x}'\setminus\mathbf{z}}(y) = P_{\mathbf{x}'}(y)$, and $Q(\mathbf{v}) = P_{\mathbf{z}}(\mathbf{v})$ is regarded as an available observational distribution.

We show next that the procedure is indeed sound, i.e., any expression that it returns is a valid equality with respect to the underlying structural causal model.

Theorem 5 (soundness)**.** *Whenever z^{2}ID returns an expression for $P_{\mathbf{x}}(\mathbf{y})$, it is correct.*

Proof. This follows from the previous identifiability results (Tian and Pearl 2002; Shpitser and Pearl 2006; Bareinboim and Pearl 2012). Specifically, Lines 2–3 correspond to C-factorization (Tian and Pearl 2002). Lines 7–8 follow from soundness of ID since $P_{\mathbf{x}}(\mathbf{y}) = P_{\mathbf{x}\setminus\mathbf{z},\mathbf{z}}(\mathbf{y})$ which is identifying $Q_{\mathbf{x}\setminus\mathbf{z}}(\mathbf{y})$ with $Q = P_{\mathbf{z}}$ in $\mathcal{G}\setminus\mathbf{Z}$ (Bareinboim and Pearl 2012). Lines 9–12 are based on *do*-calculus. Finally, Line 13–14 follows from Lemma 3 of (Shpitser and Pearl 2006). □

For concreteness, consider a causal graph $\mathcal{G}$ in Fig. 2a. With $\mathbf{N} = \{A\}$, the POMISs are $\{\emptyset, \{B\}, \{C\}\}$ (as shown in Fig. 4a above). The algorithm z^{2}ID will returns expressions for each arm's distribution as follows:

Algorithm 2 bMVWA

```
function BMVWA($\mathbf{D}$, $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m\}$, $B$)
    Input: $\mathbf{D}$: samples $\{D_{\mathbf{x}}\}_{\mathbf{x}\in\mathbf{A}}$; $\{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m\}$: $m$ estimators; $B$: the number of bootstraps
    for $b \in 1, \ldots, B$ do
        $\mathbf{D}^{(b)} \leftarrow \{D_{\mathbf{x}}^{(b)}\}_{\mathbf{x}\in\mathbf{A}}$                ▷ bootstrap samples
        $(\forall_{i=1}^{m})\ \hat{\theta}_i^{(b)} \leftarrow \hat{\theta}_i(\mathbf{D}^{(b)})$          ▷ evaluate expressions
    $\hat{\Sigma} \leftarrow \mathrm{Cov}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_m)$
    $\mathbf{w} \leftarrow \arg\min_{\mathbf{w}} \mathbf{w}^{\mathsf{T}}\hat{\Sigma}\mathbf{w}$ such that $\sum_{i=1}^{m} w_i = 1, w_i \geq 0$
    return $\sum_{i=1}^{m} w_i\hat{\theta}_i(\mathbf{D}), \mathbf{w}^{\mathsf{T}}\hat{\Sigma}\mathbf{w}$
```

$$P(y) = \textstyle\sum_{a,b,c} P_b(c|a) P_c(a, b, y) \tag{1}$$

$$P_b(y) = \textstyle\sum_{a,c} P(c|a, b) \sum_{b'} P(y|a, b', c) P(a, b') \tag{2}$$

$$P_c(y) = \textstyle\sum_{a,b} P(y|a, b, c) P(a, b) \tag{3}$$

$$P_c(y) = \textstyle\sum_{a} P_b(y|a, c) P_b(a) \tag{4}$$

Note that there are two equations for $P_c(y)$, Eqs. (3) and (4) — the former suggests that P_c can be estimated with samples from the do-nothing intervention, while the second says that it can be estimated from samples from P_b. Moreover, b in Eq. (4) can be any realization within $\mathfrak{X}_B$. This means that such a mapping can be viewed as two separate sources of data, i.e., $\sum_a P_{b=0}(y|a, c) P_{b=0}(a)$ and $\sum_a P_{b=1}(y|a, c) P_{b=1}(a)$.

The failing condition of z^{2}ID implies that it encountered a (decomposed) probability term that cannot be reduced to any of the available distributions *individually*. It is an open problem, and outside of the scope of this work, to prove the necessity of z^{2}ID, which requires to show how the available distributions cannot answer the query *collectively*.

Integrating Expressions into MAB Algorithms

Given that we have multiple expressions for each arm's distribution, it is crucial to find a principled way of combining them. Let $\theta = P_{\mathbf{x}}(y)$ be a quantity of interest, and there be m *dependent* estimators $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_i\}_{i=1}^{m}$ corresponding to expressions relating to θ. They can be combined using a weighted sum, $\hat{\theta} = \sum_{i=1}^{m} w_i\hat{\theta}_i$ with $\sum_{i=1}^{m} w_i = 1$ and $(\forall_{1\leq i\leq m})\ w_i \geq 0$ (i.e., convex combination). Given that every $\hat{\theta}_i$ is unbiased, we may focus on minimizing the variance of $\hat{\theta}$, $\mathrm{Var}(\hat{\theta}) = \mathbf{w}^{\mathsf{T}}\Sigma\mathbf{w}$, where $\Sigma = \mathrm{Cov}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}})$. Weights resulting minimum variance can be obtained through solving a constrained quadratic program. Bootstrap is a well-known technique to estimate the variance of an estimator (Efron and Tibshirani 1993). Given data D of size n, sampling n instances from the data, with replacement, yields a bootstrap sample. By repeating the procedure B times, we obtain bootstrap samples $\{D^{(b)}\}_{b=1}^{B}$, which allow us to estimate sample variance for an estimator.

Given multiple datasets, we bootstrap them simultaneously, and obtain bootstrap estimates for each estimator so as to compute a covariance matrix Σ. We then are able to compute weights that yield a minimum variance estimator. A pseudo-code (Alg. 2) describes the aforementioned procedure, named bootstrap-based minimum variance weighted average (bMVWA).

We introduce two algorithms z^2-TS and z^2-KL-UCB (Alg. 3), generalizations of the classic TS and KL-UCB algorithms, respectively, that incorporate bMVWA so that information about other arms' rewards can be exploited to infer an arm's expected reward more precisely via expressions retrieved from z^{2}ID. We first explain z^2-TS. Given a SCM-MAB instance with non-manipulable variables ($\mathcal{G}, Y$, and $\mathbf{N}$), the algorithm first computes corresponding POMISs (Line 2) and obtain estimators for arms relevant to the POMISs (Line 4). For each arm $\mathbf{x}$, bMVWA returns a minimum variance estimate and its variance. Then, it finds a pair of parameters for a beta distribution whose mean and variance correspond to $\hat{\theta}_{\mathbf{x}}$ and $\hat{s}^2_{\mathbf{x}}$, and a sample is drawn from this distribution as an approximate posterior sample.[3] Similarly to TS, an arm is chosen based on posterior samples.

Next we describe z^2-KL-UCB. The algorithm follows the same initialization steps, and computes each arm's mean payout and its variance through bMVWA. Conventional KL-UCB computes the upper-confidence bound of the expected reward of an arm $\mathbf{x}$ with a non-decreasing function f and how many times the arm is pulled, $N_{\mathbf{x}}$. It infers $\hat{N}_{\mathbf{x}}$ (Line 18), the number of 'effective' samples for each arm $\mathbf{x}$, rather than using the actual number of arms played, $N_{\mathbf{x}}$.

One property of the bootstrap is that if the number of samples is too small, bootstrap will fail, in the sense that the bootstrap distribution will not approximate the target distribution. In practice, the aforementioned approach will be deferred until enough number of samples is observed. As a rule of thumb, we do not use an expression (i.e., estimator) unless its support is covered (i.e., every possible value with respect to each term in the expression has positive mass). For example, Eq. (1) will be evaluated if all 8 and 16 values of $P_b(c|a)$ and $P_c(a, b, y)$ exist, respectively (assuming that the variables are binary). Furthermore, with a binary reward variable, we utilize a beta distribution, Beta$(\alpha_{\mathbf{x}} + 1, \beta_{\mathbf{x}} + 1)$, for the mean and variance of a trivial estimator, e.g. $\hat{\theta}_i = P_{\mathbf{x}}(y)$, where $\alpha_{\mathbf{x}}$ and $\beta_{\mathbf{x}}$ are the number of $Y_{\mathbf{x}}$ being 1 and 0, respectively, at round t.

Empirical Evaluation

In this section, we evaluate the performance (cumulative regret, CR for short) of SCM-MAB algorithms under different strategies so as to assess the effect of employing POMIS-based arm refinement as well as the dependence structure among arms. In particular, we compare the following strategies: Brute-force (BF), which intervenes on every combination of manipulable variables, MIS $\mathbb{M}^{\mathbf{N}}_{\mathcal{G},Y}$; POMIS $\mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}$; POMIS+, which is POMIS enhanced with z^{2}ID. Formally, note that the following relationships hold: BF $\supseteq$ MIS $\supseteq$ POMIS $=$ POMIS+ relative to their arms. Combined with two base MAB algorithms, TS and KL-UCB, we compared total eight settings, where only POMIS+ employs z^2-TS and z^2-KL-UCB.

Algorithm 3 Bernoulli TS and KL-UCB with z^{2}ID

```
1: function z^2-TS($\mathcal{G}, Y, \mathbf{N}, T$)
2:   $\mathbb{Z} \leftarrow \mathbb{P}^{\mathbf{N}}_{\mathcal{G},Y}$
3:   $\mathbf{A} \leftarrow \{\mathbf{x} \in \mathfrak{X}_{\mathbf{X}} \mid \mathbf{X} \in \mathbb{Z}\}$
4:   $\hat{\boldsymbol{\theta}}_{\mathbf{x}} \leftarrow \{P_{\mathbf{x}}(y)\} \cup \{\mathsf{z}^2\mathsf{ID}(\mathcal{G}, y, \mathbf{x}, \mathbb{Z}')\}_{\mathbb{Z}' \subseteq \mathbb{Z} \setminus \{\mathbf{X}\}}$ for $\mathbf{x} \in \mathbf{A}$
5:   $\mathbf{D} \leftarrow \{D_{\mathbf{x}} = \emptyset\}_{\mathbf{x} \in \mathbf{A}}$
6:   for $t$ in $1, \ldots, T$ do
7:     for $\mathbf{x} \in \mathbf{A}$ do
8:       $\hat{\theta}_{\mathbf{x}}, \hat{s}^2_{\mathbf{x}} \leftarrow \mathsf{bMVWA}(\mathbf{D}, \hat{\boldsymbol{\theta}}_{\mathbf{x}})$
9:       Find $\hat{\alpha}_{\mathbf{x}}, \hat{\beta}_{\mathbf{x}}$ such that Beta$(\hat{\alpha}_{\mathbf{x}}, \hat{\beta}_{\mathbf{x}})$ matching $\hat{\theta}_{\mathbf{x}}, \hat{s}^2_{\mathbf{x}}$
10:      $\theta_{\mathbf{x}} \sim$ Beta$(\hat{\alpha}_{\mathbf{x}}, \hat{\beta}_{\mathbf{x}})$
11:    $\mathbf{x}' \leftarrow \arg\max_{\mathbf{x} \in \mathbf{A}} \theta_{\mathbf{x}}$
12:    Sample $\mathbf{v}$ by $do(\mathbf{x}')$ and append $\mathbf{v}$ to $D_{\mathbf{x}'}$

13: function z^2-KL-UCB($\mathcal{G}, Y, \mathbf{N}, T, f \leftarrow \ln(t) + 3\ln(\ln(t))$)
14:   Initialize $\mathbb{Z}, \mathbf{A}, \{\hat{\boldsymbol{\theta}}_{\mathbf{x}}\}_{\mathbf{x} \in \mathbf{A}}, \mathbf{D}$
15:   $(\forall_{\mathbf{x} \in \mathbf{A}})$ Sample $\mathbf{v}$ by $do(\mathbf{x})$, and append $\mathbf{v}$ to $D_{\mathbf{x}}$
16:   for $t$ in $|\mathbf{A}|, \ldots, T$ do
17:     $\hat{\theta}_{\mathbf{x}}, \hat{s}^2_{\mathbf{x}} \leftarrow \mathsf{bMVWA}(\mathbf{D}, \hat{\boldsymbol{\theta}}_{\mathbf{x}})$ for $\mathbf{x} \in \mathbf{A}$
18:     $\hat{N}_{\mathbf{x}} \leftarrow \hat{\theta}_{\mathbf{x}}(1 - \hat{\theta}_{\mathbf{x}})/\hat{s}^2_{\mathbf{x}}; \quad \hat{t} \leftarrow \sum_{\mathbf{x}} \hat{N}_{\mathbf{x}}$
19:     $\boldsymbol{\mu} = \left\{ \sup \left\{ \mu \in [0,1] : \mathrm{KL}(\hat{\theta}_{\mathbf{x}}, \mu) \leq \frac{f(\hat{t})}{\hat{N}_{\mathbf{x}}} \right\} \right\}_{\mathbf{x} \in \mathbf{A}}$
20:     $\mathbf{x}' \leftarrow \arg\max_{\mathbf{x} \in \mathbf{A}} \mu_{\mathbf{x}}$
21:     Sample $\mathbf{v}$ by $do(\mathbf{x}')$, and append $\mathbf{v}$ to $D_{\mathbf{x}'}$
```

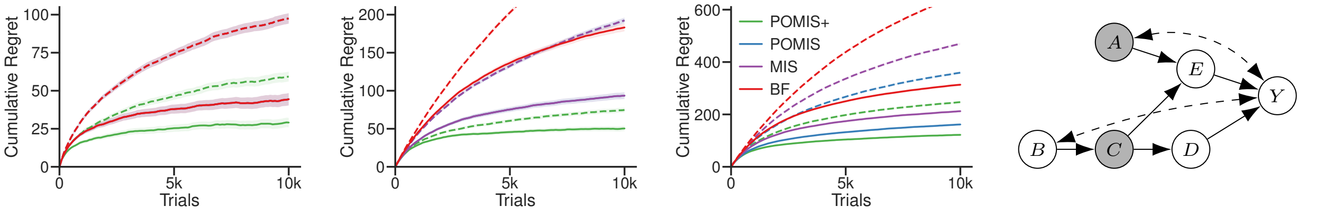
We simulated three SCM-MAB instances, i.e., the front-door setting (Fig. 1) and two models following Fig. 2a and Fig. 5d. The time horizon T is set to 10,000, which is enough to observe the performance difference; simulations are repeated 1,000 times, and the number of bootstraps B is set to 500. Detailed descriptions and expressions generated by z^{2}ID are provided in the full technical report (Lee and Bareinboim 2019). Fig. 5 illustrates experimental results and Table 1 summarizes the average cumulative regrets. We discuss these results in the sequel.

(**The Front-door Graph** with $\mathbf{N}{=}\{Z\}$) The results are shown in Fig. 5a. Due to its simplicity, BF, MIS, and POMIS are all the same in this setting.[4] A gap between POMIS and POMIS+ indicates how well expressions reduce uncertainty of arms' expected rewards so that the underlying MAB algorithm plays the optimal arm more often (and suboptimal arms less often). POMIS+ reduces CRs 34.5% (TS) and 39.2% (KL-UCB) compared to POMIS.

(**Causal Model for Fig. 2a** with $\mathbf{N}{=}\{A\}$) Results are shown in Fig. 5b. MIS and POMIS have the same performance as their arms are the same. The gain for (PO)MIS compared to BF is due to the fact that there are four less arms corresponding to intervening on $\{B, C\}$. z^{2}ID helps gain to reduce cumulative regret for both MAB algorithms. Reduction in CRs by employing POMIS+ compared to POMIS are 46.2% and 61.0% for TS and KL-UCB, respectively. Such a high reduction rate is achieved in part because expressions construct reciprocal relationships among arms (see Eqs. (1)

[3] One might simply use a weighted average of one of bootstrap estimates, $\mathbf{w}^{\intercal}\hat{\boldsymbol{\theta}}^{(b)}$ for some b.

[4] If two strategies share the same set of arms, then their results will be exactly the same since we assigned a unique random seed for each of 1,000 Monte Carlo simulations so as to reproduce the same experimental results.



(a) (b) (c) (d)

Figure 5: (a,b,c) Simulated results for settings in Figs. 1, 2a, and 5d, where the average cumulative regret (95% confidence interval) is shown relative to the four strategies discussed in the paper (BF, MIS, POMIS, POMIS+). The results are also shown for the corresponding TS (solid) and KL-UCB (dashed) solvers. (d) The causal graph used in (c) with $\mathbf{N} = \{A, C\}$.

		Fig. 5a	Fig. 5b	Fig. 5c
TS	POMIS+	29.08	50.29	121.54
	POMIS	44.40	93.53	161.54
	MIS	″	″	212.24
	BF	″	183.15	313.39
UCB	POMIS+	59.30	74.93	247.05
	POMIS	97.52	192.39	359.40
	MIS	″	″	469.30
	BF	″	328.68	657.11

Table 1: Average cumulative regret at $T = 10,000$

to (4)) except that $P_b(y)$ is not expressed with $P_c(\mathbf{v})$.

(**Causal Model for Fig. 5d**) The results are shown in Fig. 5c. This graph clearly demonstrates the advantage of having a smaller number of arms (BF 27, MIS 19, POMIS 15). In this setting, POMIS+ makes use of 24 expressions. However, the larger number of arms (compared to the previous tasks) and larger number of variables involved in formulas, e.g., a probability term $P(y|a, b, d, e)$ in some formula, make POMIS+ difficult to take full advantage of every expression in a given time horizon $T = 10,000$. The gap between POMIS and POMIS+ will be more visible with a larger time horizon. POMIS+ reduced CRs 24.8% and 31.3% for TS and KL-UCB, respectively, compared to those for POMIS.

In sum, our experiments corroborate our theoretical findings — MAB algorithms are benefited from playing a smaller number of qualified arms (POMIS) and precise estimations of arms' rewards (z^2ID). As per our observations, z^2ID will be less rewarding if the gap between rewards of the optimal and sub-optimal arms is big enough so that they can easily be distinguished before the MAB algorithm takes advantage of available expressions. Further, the performance gain will not be outstanding if the aforementioned gap is relatively small and the given time horizon is limited, or if there are many arms with low rewards that are not sufficiently played to utilize related expressions.

Conclusions

We studied the problem of *structural causal bandits* that asks whether (and how) a decision-maker should intervene in a causal system so as to optimize a particular measure. Our results generalize the previous treatment given to the problem (Lee and Bareinboim 2018), which assumed that all observed variables in the system are manipulable (i.e., could be intervened upon). For example, while cholesterol has an indisputable effect on heart failure, it's not the case that one could physically manipulate cholesterol, but perhaps she could intervene on another variable, say, maybe diet, which could accomplish the same goal of decreasing the likelihood of heart failure. Formally, we characterized two crucial properties in structural bandits with manipulability constraints, i.e.: (1) the possibly-optimal arms (POMIS), and (2) the topological relationship between such arms. We further developed an identification algorithm that outputs an expression for an (interventional) distribution providing a way to fuse an arbitrary set of experiments so that each arm's reward distribution can be estimated from samples of other arms. We equipped existing MAB algorithms with such capabilities, which are now able to play only a qualified subset of the arms while more accurately estimating their expected rewards. Following the current debate on the topic (Pearl 2018), we hope that our results will help to move the discussion from whether a non-manipulable variable causes another to how a decision-maker in the real world should intervene (or not intervene) in the system and be able to optimize a socially agreed target outcome (e.g., survival, happiness, wealth).

Acknowledgments

This research is supported in parts by grants from NSF IIS-1704352, and IIS-1750807 (CAREER), IBM Research, and Adobe Research.

References

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

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.

Bareinboim, E.; Forney, A.; and Pearl, J. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, 1342–1350.

Bottou, L.; Peters, J.; Charles, D. X.; Chickering, M.; Portugaly, E.; Ray, D.; Simard, P. Y.; and Snelson, E. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research* 14(1):3207–3260.

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.

Chapelle, O., and Li, L. 2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, 2249–2257.

Cornfield, J. 1951. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* 11:1269–1275.

Dani, V.; Hayes, T. P.; and Kakade, S. M. 2008. Stochastic linear optimization under bandit feedback. In *Proceedings of Conference On Learning Theory (COLT)*, 355–366.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Flegal, K. M.; Graubard, B. I.; and Williamson, D. F. 2005. Excess deaths associated with underweight, overweight, and obesity. 293(15):1861–1867.

Forney, A.; Pearl, J.; and Bareinboim, E. 2017. Counterfactual data-fusion for online reinforcement learners. In *International Conference on Machine Learning*, 1156–1164.

Garivier, A., and Cappé, O. 2011. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, 359–376.

Heckman, J. J. 2006. Skill formation and the economics of investing in disadvantaged children. *Science* 312(5782):1900–1902.

Kaufmann, E.; Korda, N.; and Munos, R. 2012. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, 199–213.

Lai, T., and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4 – 22.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems 29*. 1181–1189.

Lee, S., and Bareinboim, E. 2018. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, forthcoming.

Lee, S., and Bareinboim, E. 2019. Structural causal bandits with non-manipulable variables. Technical Report R-40, Purdue AI Lab, Department of Computer Science, Purdue University.

Magureanu, S.; Combes, R.; and Proutiere, A. 2014. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, 975–999. Barcelona, Spain: PMLR.

Pearl, J., and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–688.

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

Pearl, J. 2018. Does obesity shorten life? or is it the soda? on non-manipulable causes. Technical Report R-483, Forthcoming, Journal of Causal Inference, Department of Computer Science, University of California, Los Angeles, CA.

Robbins, H. 1952. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58(5):527–535.

Sen, R.; Shanmugam, K.; Dimakis, A. G.; and Shakkottai, S. 2017. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, 3057–3066.

Shpitser, I., and Pearl, J. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of The Twenty-First National Conference on Artificial Intelligence*, 1219–1226. AAAI Press.

The White House, Office of the Press Secretary. 2014. Fact sheet: Invest in us: The white house summit on early childhood education. Press Release.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Tian, J., and Pearl, J. 2002. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, 567–573. Menlo Park, CA: AAAI Press/The MIT Press.

Tian, J. 2002. *Studies in Causal Reasoning and Learning*. Ph.D. Dissertation, Computer Science Department, University of California, Los Angeles, CA.

U.S. Department of Health and Human Services. 2014. The health consequences of smoking — 50 years of progress: A report of the surgeon general. Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA.

Verma, T., and Pearl, J. 1990. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence (UAI 1990)*, 220–227.

Verma, T. 1992. Invariant properties of causal models. Technical Report R-134, Cognitive Systems Laboratory, Department of Computer Science, UCLA.

Zhang, J., and Bareinboim, E. 2017. Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1340–1346.