TECHNICAL REPORT
R-36-L
First version: September, 2018
Revised: October, 2018

Structural Causal Bandits: Where to Intervene?

Sanghack Lee
Department of Computer Science
Purdue University
lee2995@purdue.edu

Elias Bareinboim
Department of Computer Science
Purdue University
eb@purdue.edu

Abstract

We study the problem of identifying the best action in a sequential decision-making setting when the reward distributions of the arms exhibit a non-trivial dependence structure, which is governed by the underlying causal model of the domain where the agent is deployed. In this setting, playing an arm corresponds to intervening on a set of variables and setting them to specific values. In this paper, we first note that whenever the underlying causal model is not taken into account during the decision-making process, the standard strategies of simultaneously intervening on all variables or on all the subsets of the variables may, in general, lead to suboptimal policies, regardless of the number of interventions performed in the environment. In order to explain this phenomenon, we investigate a number of structural properties implied by the underlying (possibly unobserved) causal model, which include a complete characterization of the relationships between the arms' distributions. We then propose an efficient strategy that takes as input a causal structure and finds a minimal, sound, and complete set of qualified arms that an agent should play to maximize its expected reward. We empirically demonstrate that the new strategy learns an optimal policy and leads to orders of magnitude faster convergence rates when compared with its causal-insensitive counterparts.

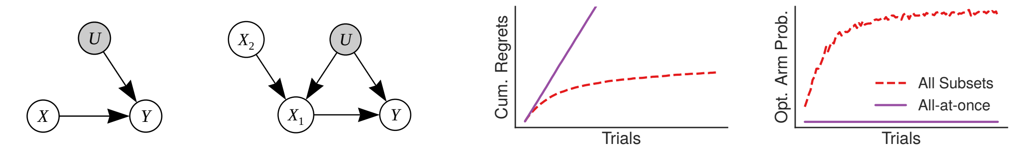
1 Introduction

The multi-armed bandit (MAB) problem is one of the prototypical settings studied in the sequential decision-making literature [Lai and Robbins, 1985, Even-Dar et al., 2006, Bubeck and Cesa-Bianchi, 2012]. In this setting, an agent pulls an arm and receives a corresponding reward at each time step, while its goal is to maximize the cumulative reward in the long run. The challenge is the inherent trade-off between exploiting known arms versus exploring new reward opportunities [Sutton and Barto, 1998, Szepesvári, 2010]. There is a wide range of assumptions underlying MABs, but in most of the traditional settings, the arms' rewards are assumed to be independent, which means that knowing the reward distribution of one arm has no implication to the reward of the other arms. Many strategies were developed to solve this problem, including classic algorithms such as ϵ-greedy, variants of UCB (Auer et al., 2002, Cappé et al., 2013), and Thompson sampling [Thompson, 1933].

Recently, the existence of some non-trivial dependencies among arms has been acknowledged in the literature and studied under the rubric of *structured bandits*, which include settings such as linear [Dani et al., 2008], combinatorial [Cesa-Bianchi and Lugosi, 2012], unimodal [Combes and Proutiere, 2014], and Lipschitz [Magureanu et al., 2014], just to name a few. For example, a linear (or combinatorial) bandit imposes that an action $x_t \in \mathbb{R}^d$ (or $\{0,1\}^d$) at a time step t incurs a cost $\ell_t^\top x_t$, where ℓ_t is a loss vector chosen by, e.g., an adversary. In this case, an *index-based* MAB algorithm, oblivious to the structural properties, can be suboptimal.

In another line of investigation, rich environments with complex dependency structures are modeled explicitly through the use of causal graphs, where nodes represent decisions and outcome variables, and direct edges represent direct influence of one variable on another [Pearl, 2000]. Despite the

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



(a) Standard MAB (b) IV-MAB (c) Cumulative Regrets (d) Probability

Figure 1: MAB problems as directed acyclic graphs where U is an unobserved variable. Plots of cumulative regrets and probability selecting an optimal arm when a MAB algorithm intervenes X_1 and X_2 simultaneously (All-at-once) or all subsets of $\{X_1, X_2\}$ for IV-MAB. The IV-MAB is also used in the experimental section (see Appendix D [Lee and Bareinboim, 2018] for its parametrization).

apparent connection between MABs and causality, only recently has the use of causal reasoning been incorporated into the design of MAB algorithms. For instance, Bareinboim et al. [2015] explored the connection between causal models with unobserved confounders (UCs) and reinforcement learning, where the UCs affect both the reward distribution and the player's intuition. The critical observation leveraged in this work is that while standard MAB algorithms optimize based on the experimental distribution (formally written as the do-distribution, $\mathbb{P}[Y|do(X)]$, and are dismissive about the UCs, it's possible to take the UCs into account through the use of other counterfactuals, in particular, the quantity $\mathbb{P}[Y_x|X = x']$. It was then showed that a counterfactual-based strategy dominates the traditional ones and, therefore, should be preferred as the target of the optimization process whenever UCs cannot be ruled out a priori. This strategy was later generalized to handle counterfactual distributions of higher dimensionality by Forney et al. [2017]. Lattimore et al. [2016] and Sen et al. [2017] studied the problem of best arm identification through importance weighting, where information on how playing arms influences the direct causes (parents, in causal terminology) of a reward variable is available. Zhang and Bareinboim [2017] noted that off-policy evaluation methods can be arbitrarily biased and lead to linear regret whenever UCs are present. It was then shown how causal bounds can be derived and used as prior of the arms' distributions, which led to provably, orders-of-magnitude more efficient off-policy evaluation. [1] Overall, these works showed different aspects of a deeper phenomenon – the expected guarantees provided by standard methods are no longer valid whenever UCs are present in the real world, which translates to an inability to converge to any reasonable policy. They then showed that convergence can be restored once the causal structure is acknowledged and used during the decision-making process.

In this paper, we focus on the challenge of identifying the best action in MABs where the arms correspond to interventions on an arbitrary causal graph, including when UCs affect the observed relations (also known in causal terminology as Semi-Markovian setting). To understand the subtlety of this problem, we first note that a standard MAB can be seen as the simple causal model shown in Fig. 1a, where X represents an arm (with K different values), Y the reward variable, and U the unobserved variable that generates the randomness of Y.[2] After a sufficiently large number of pulls of X (chosen by the specific algorithm), Y's average reward can be determined with high confidence.

On the other hand, if UCs affect more than one observed variable, non-trivial challenges arise. To witness, consider a more involved MAB structure shown in Fig. 1b, where an unobserved confounder U affects both the action variable X_1 and the reward Y. A naive approach for an algorithm to play such a bandit would be to pull arms in a combinatorial manner, i.e., combining both variables ($X_1 \times X_2$) so that arms are $D(X_1) \times D(X_2)$, where $D(X)$ is the domain of X. One may surmise that this is a valid strategy, albeit not the most efficient one. Somewhat unexpectedly, however, Fig. 1c shows that this is not the case — the optimal action comes from pulling X_2 and ignoring X_1, while pulling $\{X_1, X_2\}$ together would lead to subpar cumulative rewards, regardless of the number of interventions performed (see also Fig. 1d). After all, if one is oblivious to the causal structure, dismissing the topological relation of X_1 and X_2, and decides to take all intervenable variables as one (in this case, $X_1 \times X_2$), indiscriminately, he may be doomed to learn a suboptimal policy.

[1] On another line of investigation, Ortega and Braun [2014] introduced a generalized version of Thompson sampling applied to the problem of adaptive control.

[2] In causal notation, $Y \leftarrow f_Y(U, X)$, which means that Y's value is determined by X and the realization of the latent variable U. If f_Y is linear, we would have a (stochastic) linear bandit. Our results do not constrain the types of structural functions, which is usually within nonparametric causal inference [Pearl, 2000, Ch. 7].

In this paper, we investigate this phenomenon and, more broadly, causal MABs with non-trivial dependency structure between the arms. Specifically, our contributions are as follows: (1) We formulate a SCM-MAB problem that is a structured multi-armed bandit instance within the causal framework. We then derive some key structural properties of a SCM-MAB, which are computable from any causal model, including arms' equivalence based on *do*-calculus and partial orderedness among sets of variables associated with arms in regards to the maximum rewards achievable. (2) We characterize a special set of variables called POMIS (possibly-optimal minimal intervention set), which is worth intervening based on the aforementioned partial orders. We develop an algorithm that identifies a complete set of POMISs so that only the subset of the arms associated with them can be explored in a MAB instance. Simulations corroborate our findings.

Big picture. The multi-armed bandit is a rich setting in which a large number of variants has been proposed. Different aspects of the decision-making process have been contemplated and successfully analyzed in the last decades, which include a variety of functional forms (e.g., linear, Lipschitz, Gaussian process), different types of feedback (bandit, semi-bandit, full), the adversarial or i.i.d. nature of the interactions, just to cite some of the most popular ones. Our study of SCM-MABs puts the causal dimension front and center in this landscape. In particular, we fully acknowledge the existence of a causal structure among the underlying variables (if not known a priori, see footnote 3), and leverage the topological relations among them. This is in clear contrast with the prevailing practice that, almost invariably, is oblivious to the causal structure, implicitly assuming a non-informative model such as the one shown in Fig. 1a. We draft in Fig. 2 an initial map that shows the relationship between these qualitatively different dimensions. Our goal is to start building some intuition about these class of problems (but neither exhaustive, nor prescriptive). In particular, we study in the sequel bandits with no constraints over the underlying functional form (non-parametric, in causality terminology), i.i.d. stochastic rewards, and with an explicit causal structure available (or learnable) by the agent.

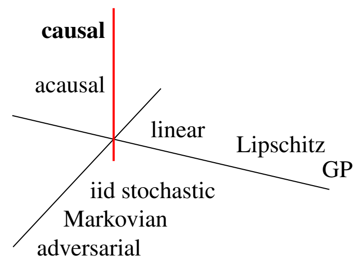
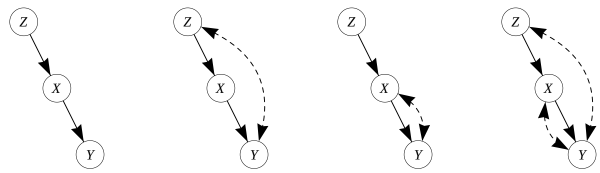


Figure 2: A bandit space with various dimensions (not all dimensions are shown)

Preliminaries: notations and structural causal models

We follow the notation used in the causal inference literature. A capital letter is used for a variable or a mathematical object. The domain of X is denoted by $D\left(X\right)$. A bold capital letter is for a set of variables, e.g., $\mathbf{X}=\left\{X_i\right\}_{i=1}^{n}$, while a lowercase letter $x \in D\left(X\right)$ is a value assigned to X, and $\mathbf{x} \in D\left(\mathbf{X}\right)=\times_{X\in\mathbf{X}}\left(D\left(X\right)\right)$. We denote by $\mathbf{x}\left[\mathbf{W}\right]$, values of $\mathbf{x}$ corresponding to $\mathbf{W}\cap\mathbf{X}$. A graph $G=\langle\mathbf{V},\mathbf{E}\rangle$ is a pair of vertices $\mathbf{V}$ and edges $\mathbf{E}$. We adopt family relationships — pa, ch, an, and de to denote parents, children, ancestors, and descendants of a given variable; Pa, Ch, An, and De extends pa, ch, an, and de by including the argument as the result, e.g., $Pa\left(X\right)_G = pa\left(X\right)_G \cup \{X\}$. With a set of variables as argument, $pa\left(\mathbf{X}\right)_G=\bigcup_{X\in\mathbf{X}} pa\left(X\right)_G$ and similarly defined for other relations. We denote by $\mathbf{V}\left(G\right)$ the set of variables in G. $G\left[\mathbf{V}'\right]$ for $\mathbf{V}'\subseteq\mathbf{V}\left(G\right)$ is a vertex-induced subgraph where all edges among $\mathbf{V}'$ are preserved. We define $G\backslash\mathbf{X}$ as $G\left[\mathbf{V}\left(G\right)\backslash\mathbf{X}\right]$ for $\mathbf{X}\subseteq\mathbf{V}\left(G\right)$.

We adopt the language of Structural Causal Models (SCM) [Pearl, 2000, Ch. 7]. An SCM M is a tuple $\langle\mathbf{U},\mathbf{V},\mathbf{F},P\left(\mathbf{U}\right)\rangle$, where $\mathbf{U}$ is a set of exogenous (unobserved or latent) variables and $\mathbf{V}$ is a set of endogenous (observed) variables. $\mathbf{F}$ is a set of deterministic functions $\mathbf{F}=\{f_i\}$, where f_i determines the value of $V_i\in\mathbf{V}$ based on endogenous variables $\mathbf{PA}_i\subseteq\mathbf{V}\backslash\left\{V_i\right\}$ and exogenous variables $\mathbf{U}^i\subseteq\mathbf{U}$, that is, e.g., $v_i\leftarrow f_i(\mathbf{pa}_i,\mathbf{u}^i)$. $P(\mathbf{U})$ is a joint distribution over the exogenous variables. A causal diagram $G=\langle\mathbf{V},\mathbf{E}\rangle$, associated with M, is a tuple of vertices $\mathbf{V}$ (the endogenous variables) and edges $\mathbf{E}$, where a directed edge $V_i\rightarrow V_j\in\mathbf{E}$ if $V_i\in\mathbf{PA}_j$, and a bidirected edge between V_i and V_j if they share an unobserved confounder, i.e., $\mathbf{U}^i\cap\mathbf{U}^j\neq\emptyset$. Note that $pa(V_i)_G$ corresponds to $\mathbf{PA}_i$. Probability of $Y=y$ when $\mathbf{X}$ is held fixed at $\mathbf{x}$ (i.e., intervened) is denoted by $P\left(y|do(\mathbf{x})\right)$, where intervention on $\mathbf{X}$ is graphically represented by $G_{\overline{\mathbf{X}}}$, the graph G with incoming edges onto $\mathbf{X}$ removed. We denote by $\mathsf{CC}\left(X\right)_G$ the *c-component* of G that contains X where a c-component is a maximal set of vertices connected with bidirected edges [Tian and Pearl, 2002]. We define $\mathsf{CC}\left(\mathbf{X}\right)_G=\bigcup_{X\in\mathbf{X}}\mathsf{CC}\left(X\right)_G$. For a more detailed discussion on the properties of SCMs, we refer readers to [Pearl, 2000, Bareinboim and Pearl, 2016]. For all the proofs and appendices, please refer to the full technical report [Lee and Bareinboim, 2018].



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

	$\emptyset$	$D(X)$	$D(Z)$
(a)		✓	
(b)	✓	✓	
(c)		✓	✓
(d)	✓	✓	✓

(e)

Figure 3: (a–d) Causal graphs such that $\mu_x = \mu_{x,z}$, and (e) non-dominated arms

2 Multi-armed bandits with structural causal models

We recall that MABs consider a sequential decision-making setting where pulling one of the K available arms at each round gives the player a stochastic reward from an unknown distribution associated with the corresponding arm. The goal is to minimize (maximize) the cumulative regret (reward) after T rounds. The mean reward of an arm a is denoted by μ_a and the maximal reward is $\mu^* = \max_{1 \leq a \leq K} \mu_a$. We focus on the cumulative regret, $\text{Reg}_T = T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[Y_{A_t}\right] = \sum_{a=1}^{K} \Delta_a \mathbb{E}\left[T_a\left(T\right)\right]$, where A_t is the arm played at time t, $T_a\left(t\right)$ is the number of arm a has been played after t rounds, and $\Delta_a = \mu^* - \mu_a$.

We now can explicitly connect a MAB instance to its SCM counterpart. Let M be a SCM $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P\left(\mathbf{U}\right) \rangle$ and $Y \in \mathbf{V}$ be a reward variable, where $D\left(Y\right) \subseteq \mathbb{R}$. The bandit contains arms $\{\mathbf{x} \in D\left(\mathbf{X}\right) \mid \mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}\}$, a set of all possible interventions on endogenous variables except the reward variable. Each arm $A_{\mathbf{x}}$ (or simply $\mathbf{x}$) associates with a reward distribution $P\left(Y|do(\mathbf{x})\right)$ where its mean reward $\mu_{\mathbf{x}}$ is $\mathbb{E}\left[Y|do(\mathbf{x})\right]$. We call this setting a SCM-MAB, which is fully represented by the pair $\langle M, Y \rangle$. Throughout this paper, we assume that the causal graph G of M is fully accessible to the agent,[3] although its parametrization is unknown: that is, an agent facing a SCM-MAB $\langle M, Y \rangle$ plays arms with knowledge of G and Y, but not of $\mathbf{F}$ and $P\left(\mathbf{U}\right)$. For simplicity, we denote information provided to an agent playing a SCM-MAB by $[\![G, Y]\!]$. We now investigate some key structural properties that follow from the causal structure G of the SCM-MAB.

Property 1. Equivalence among arms

We start by noting that *do*-calculus [Pearl, 1995] provides rules to evaluate invariances in the interventional space. In particular, we focus here on the Rule 3, which ascertains the condition such that a set of interventions does not have an effect on the outcome variable, i.e., $P(y|do(\mathbf{x}, \mathbf{z}), \mathbf{w}) = P(y|do(\mathbf{x}), \mathbf{w})$. Since arms correspond to interventions (including the *null* intervention) and there is no contextual information, we consider examining $P(y|do(\mathbf{x}, \mathbf{z})) = P(y|do(\mathbf{x}))$ through $Y \perp\!\!\!\perp \mathbf{Z} \mid \mathbf{X}$ in $G_{\overline{\mathbf{X} \cup \mathbf{Z}}}$, which implies $\mu_{\mathbf{x},\mathbf{z}} = \mu_{\mathbf{x}}$. If valid, this condition implies that it is sufficient to play only one arm among arms in the equivalence class.

Definition 1 (Minimal Intervention Set (MIS))**.** A set of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$ is said to be a *minimal intervention set* relative to $[\![G, Y]\!]$ if there is no $\mathbf{X}' \subset \mathbf{X}$ such that $\mu_{\mathbf{x}[\mathbf{X}']} = \mu_{\mathbf{x}}$ for every SCM conforming to the G.

For instance, the MISs corresponding to the causal graphs in Fig. 3 are $\{\emptyset, \{X\}, \{Z\}\}$, which do not include $\{X, Z\}$ since $\mu_x = \mu_{x,z}$. The MISs are determined without considering the UCs in a causal graph. The empty set and all singletons in $an\left(Y\right)_G$ are MISs for G with respect to Y. The task of finding the best arm among all possible arms can be reduced to a search within the MISs.

Proposition 1 (Minimality)**.** *A set of variables $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$ is a minimal intervention set for G with respect to Y if and only if $\mathbf{X} \subseteq an\left(Y\right)_{G_{\overline{\mathbf{X}}}}$.*

All the MISs given $[\![G, Y]\!]$ can be determined without explicitly enumerating $2^{\mathbf{V} \setminus \{Y\}}$ while checking the condition in Prop. 1. We provide an efficient recursive algorithm enumerating the complete set of MISs given G and Y (Appendix A), which runs in $O(mn^2)$ where m is the number of MISs.

[3]In settings where this is not the case, one can spend the first interactions with the environment to learn the causal graph G from observational [Spirtes et al., 2001] or experimental data [Kocaoglu et al., 2017].

Property 2. Partial-orders among arms

We now explore the partial-orders among subsets of $\mathbf{V}\backslash\{Y\}$ within the MISs. Given the causal diagram G, it is possible that intervening on some variables is *always* as good as intervening on another set of variables (regardless of the parametrization of the underlying model). Formally, there can be two different sets of variables $\mathbf{W}, \mathbf{Z} \subseteq \mathbf{V}\backslash\{Y\}$ such that

$$\max_{\mathbf{w}\in D(\mathbf{W})} \mu_{\mathbf{w}} \leq \max_{\mathbf{z}\in D(\mathbf{Z})} \mu_{\mathbf{z}}$$

in every possible SCM conforming to G. If that is the case, it would be unnecessary (and possibly harmful in terms of sample efficiency) to play arms $D(\mathbf{W})$. We next define Possibly-Optimal MIS, which incorporates the partial-orderedness among subsets of $\mathbf{V}\backslash\{Y\}$ into MIS denoting the optimal value for a $\mathbf{X} \subseteq \mathbf{V}\backslash\{Y\}$ given a SCM by $\mathbf{x}^*$.

Definition 2 (Possibly-Optimal Minimal Intervention Set (POMIS))**.** Given information $[\![G, Y]\!]$, let $\mathbf{X}$ be a MIS. If there exists a SCM conforming to G such that $\mu_{\mathbf{x}^*} > \forall_{\mathbf{Z}\in\mathbb{Z}\backslash\{\mathbf{X}\}} \mu_{\mathbf{z}^*}$, where $\mathbb{Z}$ is the set of MISs with respect to G and Y, then $\mathbf{X}$ is a *possibly-optimal minimal intervention set* with respect to the information $[\![G, Y]\!]$.

Intuitively, one may believe that the best action will be to intervene on the direct causes (parents) of the reward variable Y, since this would entail a higher degree of "controllability" of Y within the system. This, in fact, holds true if Y is not confounded with any of its ancestors, which includes the case where no unobserved confounders are present in the system (i.e., Markovian models).

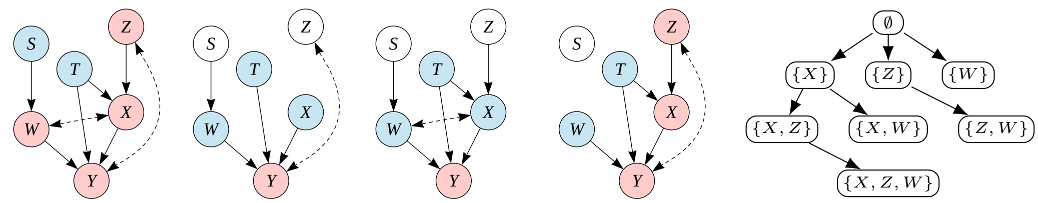
Proposition 2. *Given information $[\![G, Y]\!]$, if Y is not confounded with $an(Y)_G$ via unobserved confounders, then $pa(Y)_G$ is the only POMIS.*

Corollary 3 (Markovian POMIS)**.** *Given $[\![G, Y]\!]$, if G is Markovian, then $pa(Y)_G$ is the only POMIS.*

For instance, in Fig. 3a, $\{\{X\}\}$ is the set of POMISs. Whenever unobserved confounders (UCs) are present,[4] on the other hand, the analysis becomes more involved. To witness, let us analyze the maximum achievable rewards of the MISs in the other causal diagrams in Fig. 3. We start with Fig. 3b and note that $\mu_{z^*} \leq \mu_{x^*}$ since $\mu_{z^*} = \sum_x \mu_x P(x|do(z^*)) \leq \sum_x \mu_x^* P(x|do(z^*)) = \mu_{x^*}$. On the other hand, $\mu_\emptyset$ is not comparable to μ_{x^*}. For a concrete example, consider a SCM where the domains of variables are $\{0, 1\}$. Let U be the UC between Y and Z where $P(U = 1) = 0.5$. Let $f_Z(u) = 1 - u$, $f_X(z) = z$, and $f_Y(x, u) = x \oplus u$, where $\oplus$ is the exclusive-or function. If X is not intervened on, x will be $1 - u$ yielding $y = 1$ for both cases $u = 0$ or $u = 1$ so that $\mu_\emptyset = 1$. However, if X is intervened to either 0 or 1, y will be 1 only half the time since $P(U = 1) = 0.5$, which results in $\mu_{x^*} = 0.5$. We also provide a SCM in Appendix A such that $\mu_\emptyset < \mu_{x^*}$ holds true. This model ($\mu_\emptyset > \mu_{x^*}$) illustrates an interesting phenomenon — allowing an UC to affect Y freely may lead to a higher reward, which may be broken upon interventions. We now consider the different confounding structure shown in Fig. 3c (similar to Fig. 1b), where the variable Z lies outside of the influence of the UC associated with Y. In this case, intervening on Z leads to a higher reward, $\mu_{z^*} \geq \mu_\emptyset$. To witness, note that $\mu_\emptyset = \sum_z \mathbb{E}[Y|z] P(z) = \sum_z \mu_z P(z) \leq \sum_z \mu_{z^*} P(z) = \mu_{z^*}$. However, μ_{z^*} and μ_{x^*} are incomparable, which is shown through two models provided in Appendix A. Finally, we can add the confounders of the two previous models, which is shown in Fig. 3d. In this case, all three μ_{x^*}, μ_{z^*}, and $\mu_\emptyset$ are incomparable. One can imagine scenarios where the influence of the UCs are weak enough so that corresponding models produce results similar to Figs. 3a to 3c.

It's clear that the interplay between the location of the intervened variable, the outcome variable, and the UCs entails non-trivial interactions and consequences in terms of the reward. The table in Fig. 3e highlights the arms that are contenders to generate the highest rewards in each model (i.e., each arm intervenes a POMIS to specific values), while intervening on a non-POMIS represents a waste of resources. Interestingly, the only parent of Y, i.e., X, is not dominated by any other arms in any of the scenarios discussed. In words, this suggests that the intuition that controlling variables closer to Y is not entirely lost even when UCs are present; they are not the only POMIS, but certainly one of them. Given that more complex mechanisms cannot be, in general, ruled out, performing experiments would be required to identify the best arm. Still, the results of the table guarantee that the search can be refined so that MAB solvers can discard arms that cannot lead to profitable outcomes, and converge faster to playing the optimal arm.

[4] Recall that unobserved confounders are represented in the graph as bidirected dashed edges.



(a) G (b) $G_{\overline{X}}$ (c) $G_{\overline{Z}}$ (d) $G_{\overline{W}}$ (e)

Figure 4: Causal graphs where pink and blue nodes are MUCT and IB, respectively. (Right most) A schematic showing an exploration order of subsets of variables.

3 Graphical characterization of POMIS

Our goal in this section is to graphically characterize POMISs. We will leverage the discussion in the previous section and note that UCs connected to a reward variable affect the reward distributions in a way that intervening on a variable outside the coverage of such UCs (including no UC) can be optimal — e.g., $\{X\}$ for Fig. 3a, $\emptyset$ for Figs. 3b and 3d, and $\{Z\}$ for Fig. 3c. We introduce two graphical concepts to help characterizing this property.

Definition 3 (Unobserved-Confounders' Territory)**.** Given information $\llbracket G, Y \rrbracket$, let H be $G\left[An\left(Y\right)_G\right]$. A set of variables $\mathbf{T} \subseteq \mathbf{V}\left(H\right)$ containing Y is called an *UC-territory* on G with respect to Y if $De\left(\mathbf{T}\right)_H = \mathbf{T}$ and $\mathsf{CC}\left(\mathbf{T}\right)_H = \mathbf{T}$.

An UC-territory $\mathbf{T}$ is said to be *minimal* if no $\mathbf{T}' \subset \mathbf{T}$ is an UC-territory. A minimal UC-Territory (MUCT) for G and Y can be constructed by extending a set of variables, starting from $\{Y\}$, alternatively updating the set with the c-component and descendants of the set.

Definition 4 (Interventional Border)**.** Let $\mathbf{T}$ be a minimal UC-territory on G with respect to Y. Then, $\mathbf{X} = pa\left(\mathbf{T}\right)_G \setminus \mathbf{T}$ is called an *interventional border* for G with respect to Y.

The interventional border (IB) encompasses essentially the parents of the MUCT. For concreteness, consider Fig. 4a, and note that $\{W, X, Y, Z\}$ is the MUCT for the causal graph with respect to Y, and the IB is $\{S, T\}$ (marked in pink and blue in the graph, respectively). As its name suggests, MUCT is a set of endogenous variables governed by a set of UCs where at least one UC is adjacent to a reward variable. Specifically, the reward is determined by values of: (1) the UCs governing the MUCT; (2) a set of unobserved variables (other than the UCs) where each affects an endogenous variable in the MUCT; and (3) the IB. In other words, there is no UC interplaying *across* MUCT and its outside so that $\mu_{\mathbf{x}} = \mathbb{E}[Y|\mathbf{x}]$ where $\mathbf{x}$ is a value assigned to the IB $\mathbf{X}$. We now connect MUCT and IB with POMIS. Let $\mathsf{MUCT}(G, Y)$ and $\mathsf{IB}(G, Y)$ be, respectively, the MUCT and IB given $\llbracket G, Y \rrbracket$.

Proposition 4. *$\mathsf{IB}(G, Y)$ is a POMIS given $\llbracket G, Y \rrbracket$.*

The main strategy of the proof is to construct a SCM M where intervening on any variable in $\mathsf{MUCT}(G, Y)$ causes significant loss of reward. It seems that MUCT and IB can only identify a single POMIS given $\llbracket G, Y \rrbracket$. However, they, in fact, serve as basic units to identify all POMISs.

Proposition 5. *Given $\llbracket G, Y \rrbracket$, $\mathsf{IB}(G_{\overline{\mathbf{W}}}, Y)$ is a POMIS, for any $\mathbf{W} \subseteq \mathbf{V} \setminus \{Y\}$.*

Prop. 5 generalizes Prop. 4 for when $\mathbf{W} \neq \emptyset$ while taking care of UCs across $\mathsf{MUCT}(G_{\overline{\mathbf{W}}}, Y)$, and its outside in the original causal graph G. See Fig. 4d, for an instance, where $\mathsf{IB}(G_{\overline{W}}, Y) = \{W, T\}$. Intervening on W cuts the influence of S and the UC between W and X, while still allowing the UC to affect X.[5] Similarly, one can see in Fig. 4b that $\mathsf{IB}(G_{\overline{X}}, Y) = \{T, W, X\}$ where intervening on X lets Y be the only element of MUCT making its parents an interventional border, hence, a POMIS. Note that $pa(Y)_G$ is always a POMIS since $\mathsf{MUCT}(G_{\overline{pa(Y)_G}}, Y) = \{Y\}$ and $\mathsf{IB}(G_{\overline{pa(Y)_G}}, Y) = pa(Y)_G$. With Prop. 5, one can enumerate the POMISs given $\llbracket G, Y \rrbracket$ considering all subsets of $\mathbf{V} \setminus \{Y\}$. We show in the sequel that this strategy encompasses all the POMISs.

Theorem 6. *Given $\llbracket G, Y \rrbracket$, $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$ is a POMIS if and only if $\mathsf{IB}(G_{\overline{\mathbf{X}}}, Y) = \mathbf{X}$.*

[5]Note that exogenous variables that do not affect more than one endogenous variable (i.e., non-UCs) are not explicitly represented in the graph.

Algorithm 1 Algorithm enumerating all POMISs with $[\![G, Y]\!]$

1: **function** POMISs($G,\ Y$)
2: $\quad \mathbf{T}, \mathbf{X} = \mathsf{MUCT}\,(G, Y)\,, \mathsf{IB}\,(G, Y);\ H = G_{\overline{\mathbf{X}}}\,[\mathbf{T} \cup \mathbf{X}]$
3: $\quad$ **return** $\{\mathbf{X}\} \cup \mathsf{subPOMISs}\,(H,\ Y,\ \mathsf{reversed}\,(\mathsf{topological\text{-}sort}\,(H)) \cap (\mathbf{T} \setminus \{Y\})\,,\ \emptyset)$
4: **function** subPOMISs($G,\ Y,\ \boldsymbol{\pi},\ \mathbf{O}$)
5: $\quad \mathbf{P} = \emptyset$
6: $\quad$ **for** $\pi_i \in \boldsymbol{\pi}$ **do**
7: $\quad\quad \mathbf{T},\ \mathbf{X},\ \boldsymbol{\pi}',\ \mathbf{O}' = \mathsf{MUCT}(G_{\overline{\pi_i}}, Y),\ \mathsf{IB}(G_{\overline{\pi_i}}, Y),\ \boldsymbol{\pi}^{i+1:|\boldsymbol{\pi}|} \cap \mathbf{T},\ \mathbf{O} \cup \boldsymbol{\pi}^{1:i-1}$
8: $\quad\quad$ **if** $\mathbf{X} \cap \mathbf{O}' = \emptyset$ **then**
9: $\quad\quad\quad \mathbf{P} = \mathbf{P} \cup \{\mathbf{X}\} \cup (\mathsf{subPOMISs}\,(G_{\overline{\mathbf{X}}}\,[\mathbf{T} \cup \mathbf{X}]\,,\ Y,\ \boldsymbol{\pi}',\ \mathbf{O}')$ **if** $\boldsymbol{\pi}' \neq \emptyset$ **else** $\emptyset)$
10: $\quad$ **return** $\mathbf{P}$

Algorithm 2 POMIS-based kl-UCB

function POMIS-KL-UCB(B, G, Y, f, T)
$\quad$ **Input**: B, a SCM-MAB, G, a causal diagram; Y, a reward variable
$\quad \mathbf{A} = \bigcup_{\mathbf{X} \in \mathsf{POMISs}(G,\ Y)} D(\mathbf{X})$
$\quad \mathsf{kl\text{-}UCB}(B,\ \mathbf{A},\ f,\ T)$

Thm. 6 provides a graphical necessary and sufficient condition for a set of variables being a POMIS given $[\![G, Y]\!]$. This characterization allows one to determine all possible arms in a SCM-MAB that are worth intervening on, and, therefore, being free from pulling the other unnecessary arms.

4 Algorithmic characterization of POMIS

Although the graphical characterization provides a means to enumerate the complete set of POMISs given $[\![G, Y]\!]$, a naively implemented algorithm requires time exponential in $|\mathbf{V}|$. We construct an efficient algorithm (Alg. 1) that enumerates all the POMISs based on Props. 7 and 8 below and the graphical characterization introduced in the previous section (Thm. 6).

Proposition 7. *Let* $\mathbf{T}$ *and* $\mathbf{X}$ *be the* $\mathsf{MUCT}(G_{\overline{\mathbf{W}}}, Y)$ *and* $\mathsf{IB}(G_{\overline{\mathbf{W}}}, Y)$*, respectively, relative to* G *and* Y*. Then, for any* $\mathbf{Z} \subseteq \mathbf{V} \backslash \mathbf{T}$*,* $\mathsf{MUCT}(G_{\overline{\mathbf{X} \cup \mathbf{Z}}}, Y) = \mathbf{T}$ *and* $\mathsf{IB}(G_{\overline{\mathbf{X} \cup \mathbf{Z}}}, Y) = \mathbf{X}$.

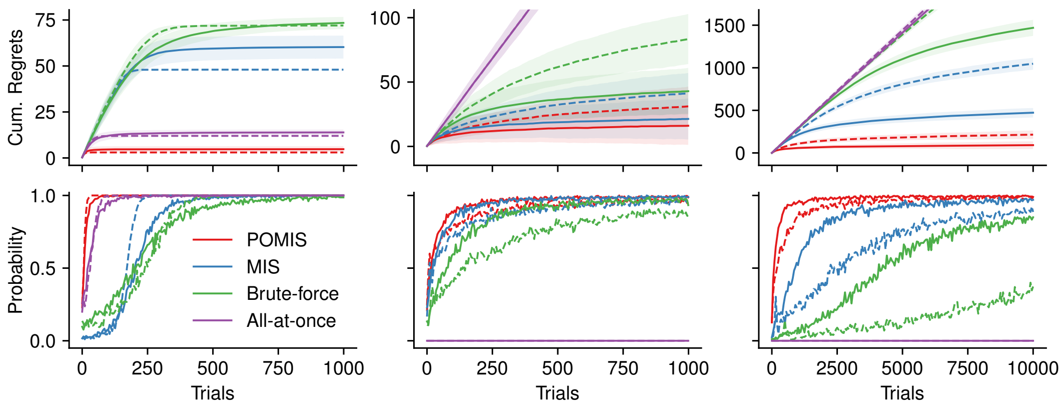
Proposition 8. *Let* $H{=}G_{\overline{\mathbf{X}}}\,[\mathbf{T} \cup \mathbf{X}]$ *where* $\mathbf{T}$ *and* $\mathbf{X}$ *are MUCT and IB given* $[\![G_{\overline{\mathbf{W}}}, Y]\!]$*, respectively. Then, for any* $\mathbf{W}' \subseteq \mathbf{T} \backslash \{Y\}$*,* $H_{\overline{\mathbf{W}'}}$ *and* $G_{\overline{\mathbf{W} \cup \mathbf{W}'}}$ *yield the same MUCT and IB with respect to* Y.

Prop. 7 allows one to avoid having to examine $G_{\overline{\mathbf{W}}}$ for every $\mathbf{W} \subseteq \mathbf{V} \backslash \{Y\}$. Prop. 8 characterizes the recursive nature of MUCT and IB, where identification of POMISs can be evaluated by subgraphs. Based on these results, we design a recursive algorithm (Alg. 1) to explore subsets of $\mathbf{V} \backslash \{Y\}$ with a certain order. See Fig. 4e for an example where subsets of $\{X, Z, W\}$ are connected based on set inclusion relationship and an order of variables, e.g., (X, Z, W). That is, there exists a directed edge between two sets if (i) one set is larger than the other by a variable and (ii) the variable's index (as in the order) is larger than other variable's index in the smaller set. The diagram traces how the algorithm will explore the subsets following the edges, while effectively skipping nodes.

Given G and Y, POMISs (Alg. 1) computes a POMIS, i.e., $\mathsf{IB}(G, Y)$. Then, a recursive procedure subPOMISs is called with an order of variables (Line 3). Then subPOMISs examines POMISs by intervening on a single variable against the given graph (Line 6–9). If the IB ($\mathbf{X}$ in Line 7) of such an intervened graph intersects with $\mathbf{O}'$ (a set of variables that should be considered in other branch), then no subsequent call is made (Line 8). Otherwise, a subsequent subPOMISs call will take as arguments an MUCT-IB induced subgraph (Prop. 8), a refined order, and a set of variables not to be intervened in the given branch. For clarity, we provide a detailed working example in Appendix C with Fig. 4a where the algorithm explores only four intervened graphs (G, $G_{\overline{\{X\}}}$, $G_{\overline{\{Z\}}}$, $G_{\overline{\{W\}}}$) and generates the complete set of POMISs $\{\{S, T\}, \{T, W\}, \{T, W, X\}\}$.

Theorem 9 (Soundness and Completeness)**.** *Given information* $[\![G, Y]\!]$*, the algorithm POMISs (Alg. 1) returns all, and only POMISs.*

The POMISs algorithm can be combined with a MAB algorithm, such as the kl-UCB, creating a simple yet effective SCM-MAB solver (see Alg. 2). kl-UCB satisfies $\limsup_{n \to \infty} \frac{\mathbb{E}[\mathrm{Reg}_n]}{\log(n)} \leq$



(a) Task 1 (b) Task 2 (c) Task 3

Figure 5: Comparisons across tasks (columns) with cumulative regrets (top) and optimal arm selection probability (bottom) with TS for solid and kl-UCB for dashed lines. Best viewed in color.

$\sum_{\mathbf{x}:\mu_{\mathbf{x}}<\mu^*} \frac{\mu^*-\mu_{\mathbf{x}}}{KL(\mu_{\mathbf{x}},\mu^*)}$ where KL is Kullback-Leibler divergence between two Bernoulli distributions [Garivier and Cappé, 2011]. It is clear that the reduction in the size of arms will lower the upper bounds of the corresponding cumulative regrets.

5 Experiments

In this section, we present empirical results demonstrating that the selection of arms based on POMISs makes standard MAB solvers converge faster to an optimal arm. We employ two popular MAB solvers, kl-UCB, which enjoys cumulative regret growing logarithmically with the number of rounds [Cappé et al., 2013], and Thompson sampling (TS, Thompson [1933]), which has strong empirical performance [Kaufmann et al., 2012]. We considered four strategies for selecting arms, including POMISs, MISs, Brute-force, and All-at-once, where Brute-force evaluates all combinations of arms $\bigcup_{\mathbf{X}\subseteq\mathbf{V}\setminus\{Y\}} D(\mathbf{X})$, and All-at-once considers intervening in all variables simultaneously, $D(\mathbf{V}\setminus\{Y\})$, oblivious to the causal structure and any knowledge about the action space. The performance of the eight (4×2) algorithms are evaluated relative to three different SCM-MAB instances (the detailed parametrizations are provided in Appendix D). We set the horizon large enough so as to observe near convergence, and repeat each simulation 300 times. We plot (i) the average cumulative regrets (CR) along with their respective standard deviations and (ii) the probability of an optimal arm being selected averaged over the repeated tests (OAP).[6,7]

Task 1: We start by analyzing a Markovian model. We note that by Cor. 3, searching for the arms within the parent set is sufficient in this case. The number of arms for POMISs, MISs, Brute-force, and All-at-once are 4, 49, 81, and 16, respectively. Note that there are 4 optimal arms within All-at-once arms — for instance, if the parent configuration is $X_1 = x_1, X_2 = x_2$, this strategy will also include combinations of $Z_1 = z_1, Z_2 = z_2, \forall z_1, z_2$. The simulated results are shown in Fig. 5a. CR at round 1000 with kl-UCB are 3.0, 48.0, 72, and 12 (in the order), and all strategies were able to find the optimal arms at this time. POMIS and All-at-once first reached 95% OAP at round 20 and 66, respectively. There are two interesting observations at this point. First, at an

[6] All the code is available at `https://github.com/sanghack81/SCMMAB-NIPS2018`

[7] One may surmise that combinatorial bandit (CB) algorithms can be used to solve SCM-MAB instances by noting that an intervention can be encoded as a binary vector, where each dimension in the vector corresponds to intervening on a single variable with a specific value. However, the two settings invoke a very different set of assumptions, which makes their solvers somewhat difficult to compare in some reasonably fair way. For instance, the current generation of CB algorithms is oblivious to the underlying causal structure, which makes them resemble very closely the Brute-force strategy, the worst possible method for SCM-MABs. Further, the assumption of linearity is arguably one of the most popular considered by CB solvers. The corresponding algorithms, however, will be unable to learn the arms' rewards properly since a SCM-MAB is nonparametric, making no assumption about the underlying structural mechanisms. These are just a few immediate examples of the mismatches between the current generation of algorithms for both causal and combinatorial bandits.

early stage, OAP for MISs is smaller than Brute-force since it has only 1 optimal arm among 49 arms, while Brute-force has 9 among 81. The advantage of employing MIS over Brute-force is only observed after a sufficiently large number of plays. More interestingly, POMIS and All-at-once both have the common optimal to non-optimal arms-ratio (1:3 versus 4:12), however, POMIS dominates All-at-once since the agent can learn better about the mean reward of the optimal arm while playing non-optimal arms less. Naturally, this translates into less variability and additional certainty about the optimal arm even in Markovian settings.

Task 2: We consider the setting known as instrumental variable (IV), which was shown in Fig. 3c. The optimal arm in this simulation is setting $Z = 0$. The number of arms for the four strategies is 4, 5, 9, and 4, respectively. The results are shown in Fig. 5b. Since the All-at-once strategy only considers non-optimal arms (i.e., pulling Z, X together), it incurs in a linear regret without selecting an optimal arm (0%). CR (and OAP) at round 1000 with TS are POMIS 16.1 (98.67%), MIS 21.4 (99.00%), Brute-force 42.9 (93.33%), and All-at-once 272.1 (0%). At round 5000, where Brute-force nearly converged, the ratio of CRs for POMIS and Brute-force is $\frac{54.2}{18.1} = 2.99 \gtrapprox 2.67 = \frac{9-1}{4-1}$. POMIS, MIS, and Brute-force first hits 95% OAP at 172, 214, and 435.

Task 3: Finally, we study the more involved scenario shown in Fig. 4a. In this case, the optimal arm is intervening on $\{S, T\}$, which means that the system should follow its natural flow of UCs, which All-at-once is unable to "pull." There are 16, 75, 243, and 32 arms for the strategies (in the order). The results are shown in Fig. 5c. The CR (and OAP) at round 10000 with TS are POMIS 91.4 (99.0%), MIS 472.4 (97.0%), Brute-force 1469.0 (85.0%), and All-at-once 2784.8 (0%). Similarly, the ratio (in round 10000) is $\frac{1469.0}{91.4} = 16.07 \approx 16.13 = \frac{243-1}{16-1}$ which is expected to increase since Brute-force is not yet converged at the moment. Only POMIS and MIS achieved OAP of 95% first in 684 and 3544 steps, respectively.

We start by noticing that the reduction in the CRs is approximately proportional to the reduction in the number of non-optimal arms pulled by (PO)MIS by the corresponding algorithm, which makes the POMIS-based solver the clear winner throughout the simulations. It's still not inconceivable that the number of arms examined by All-at-once is smaller than for POMIS in a specific SCM-MAB instance, which would entail a lower CR to the former. However, such a lower CR in some instances does not constitute any sort of assurance since arms excluded from All-at-once, but included in POMIS, can be optimal in some SCM-MAB instance conforming to $[\![G, Y]\!]$. Furthermore, a POMIS-based strategy always dominates the corresponding MIS and Brute-force ones. These observations together suggest that, in practice, a POMIS-based strategy should be preferred given that it will always converge and will usually be faster than its counterparts. Remarkably, there is an interesting trade-off between having knowledge of the causal structure versus not knowing the corresponding dependency structure among arms, and potentially incurring in linear regret (All-at-once) or exponential slow-down (Brute-force). In practice, for the cases in which the causal structure is unknown, the pull of the arms themselves can be used as experiments and could be coupled with efficient strategies to simultaneously learn the causal structure [Kocaoglu et al., 2017].

6 Conclusions

We studied the problem of deciding whether an agent should perform a causal intervention and, if so, which variables it should intervene upon. The problem was formalized using the logic of structural causal models (SCMs) and formalized through a new type of multi-armed bandit called SCM-MABs. We started by noting that whenever the agent cannot measure all the variables in the environment (i.e., unobserved confounders exist), standard MAB algorithms that are oblivious to the underlying causal structure may not converge, regardless of the number of interventions performed in the environment. (We note that the causal structure can easily be learned in a typical MAB setting since the agent always has interventional capabilities.) We introduced a novel decision-making strategy based on properties following the *do*-calculus, which allowed the removal of redundant arms, and the partial-orders among the sets of variables existent in the underlying causal system, which led to the understanding of the maximum achievable reward of each interventional set. Leveraging this new strategy based on the possibly-optimal minimal intervention sets (called POMIS), we developed an algorithm that decides whether (and if so, where) interventions should be performed in the underlying system. Finally, we showed by simulations that this causally-sensible strategy performs more efficiently and more robustly than their non-causal counterparts. We hope that formal machinery and the algorithms developed here can help decision-makers to make more principled and efficient decisions.

Supplementary Material – "Structural Causal Bandits: Where to Intervene?"

Appendix A Multi-armed bandits with structural causal models

Proposition 1 (Minimality). *A set of variables $\mathbf{X} \subseteq \mathbf{V}\backslash\{Y\}$ is a minimal intervention set for G with respect to Y if and only if $\mathbf{X} \subseteq an\left(Y\right)_{G_{\overline{\mathbf{X}}}}$.*

Proof. (If) Assume that there exists $\mathbf{X}' \subset \mathbf{X}$ such that $\mu_{\mathbf{x}'} = \mu_{\mathbf{x}}$ with $\mathbf{x}' = \mathbf{x}\left[\mathbf{X}'\right]$ so that $\mathbf{X}$ is not minimal. Create a SCM with all variables real-valued, where each variable $V_i \in \mathbf{V}$ associates with its own binary exogenous variable $P\left(u_i = 1\right) = 0.5$. Let the function of an endogenous variable be the sum of values of its parents. For the sake of contradiction assume that $\mathbf{X} \subseteq an(Y)_{G_{\overline{X}}}$. Then, there exists directed paths from $\mathbf{X} \setminus \mathbf{X}'$ to Y without passing $\mathbf{X}'$. Hence, setting $\mathbf{W} = \mathbf{X} \setminus \mathbf{X}'$ to $\mathbb{E}[\mathbf{w}|do(\mathbf{x}')] + 1$ will yield a larger outcome, i.e., $\mu_{\mathbf{w},\mathbf{x}'} > \mu_{\mathbf{x}'}$, breaking the equality, which contradicts the assumption.

(Only if) Let $\mathbf{X} \not\subseteq an(Y)_{G_{\overline{\mathbf{X}}}}$. Let $\mathbf{Z} = \mathbf{X} \setminus an\left(Y\right)_{G_{\overline{\mathbf{X}}}}$, which is a nonempty set and $\mathbf{X}' = \mathbf{X} \setminus \mathbf{Z}$. By Rule 3 of *do*-calculus, $\mu_{\mathbf{x}',\mathbf{z}} = \mu_{\mathbf{x}'}$, which violates the definition of MIS. □

We present an algorithm (Alg. 3) for enumerating all the MISs given a causal diagram G and a reward variable Y. The algorithm builds a set of MISs by adding a variable to a previously obtained MIS so that the resulting set is a MIS.

Algorithm 3 Minimal Intervention Set Enumeration

function $\mathsf{MISs}(G, Y)$
  Input: G a causal diagram; Y an outcome variable
  $H = G\left[An\left(Y\right)_G\right]$
  return $\mathsf{subMISs}\left(H,\ Y,\ \emptyset,\ \mathsf{reversed}\left(\mathsf{topological\text{-}sort}\left(H\right)\right) \cap \{\mathbf{V}\backslash\{Y\}\}\right)$

function $\mathsf{subMISs}(G,\ Y,\ \mathbf{X},\ \mathbf{W})$
  $\mathbb{X} = \{\mathbf{X}\}$
  for $W_i \in \mathbf{W}$ **do**
    $H = G_{\overline{W_i}}[An(Y)_{G_{\overline{W_i}}}]$
    $\mathbb{X} = \mathbb{X} \cup \mathsf{subMISs}(H,\ Y,\ \mathbf{X} \cup \{W_i\},\ \mathbf{W}^{i+1:} \cap \mathbf{V}(H))$
  return $\mathbb{X}$

We prove Prop. 2 below with the following observation — given two different MISs $\mathbf{X}$ and $\mathbf{Z}$, if $\mu_{\mathbf{x}^*} \leq \mu_{\mathbf{z}^*}$ for every SCM conforming to a given causal diagram G, then there exists a SCM $\mu_{\mathbf{x}^*} < \mu_{\mathbf{z}^*}$.

Proposition 2. *Given information $[\![G, Y]\!]$, if Y is not confounded with $an(Y)_G$ via unobserved confounders, then $pa(Y)_G$ is the only POMIS.*

Proof. Let $\mathbf{X}$ be a MIS. Let $\mathbf{X}' = \mathbf{X} \setminus pa(Y)_G$ and $\mathbf{Z} = pa(Y)_G \setminus \mathbf{X}$.

$$
\begin{aligned}
\mu_{\mathbf{x}} &= \sum_{\mathbf{z}} \mathbb{E}[Y \mid do(\mathbf{x}), \mathbf{z}] P(\mathbf{z}|do(\mathbf{x})) \\
&= \sum_{\mathbf{z}} \mathbb{E}[Y \mid do(\mathbf{x}, \mathbf{z})] P(\mathbf{z}|do(\mathbf{x})) && \because \text{Rule 2} \\
&= \sum_{\mathbf{z}} \mathbb{E}[Y \mid do(\mathbf{x}[pa(Y)_G], \mathbf{z})] P(\mathbf{z}|do(\mathbf{x})) \\
&\leq \sum_{\mathbf{z}} \mu_{\mathbf{pa}_Y^*} P(\mathbf{z}|do(\mathbf{x})) \\
&= \mu_{\mathbf{pa}_Y^*}
\end{aligned}
$$

$\square$

We describe two models for Fig. 3b showing $\mu_\emptyset > \mu_{x^*}$ and $\mu_\emptyset < \mu_{x^*}$, respectively. Let $\oplus$ represent the exclusive-or function.

- $\mu_\emptyset > \mu_{x^*}$: Let the domains of U, X, and Z be $\{0, 1\}$ and let $0 < P\left(U = 1\right) = \alpha < 1$. $\mathbf{F}$ consists of $f_Z\left(u\right) = 1 - u$, $f_X\left(z\right) = z$, and $f_Y\left(z, u\right) = z \oplus u$. Then, $\mu_{x^*} = \mu_{z^*} = \max\left(\alpha, 1 - \alpha\right)$, which is smaller than $\mu_\emptyset = 1$:
$$\begin{aligned}\mu_{z^*} &= \sum_{u,x,y} y \cdot P\left(y|x, u\right) P\left(x|z^*\right) P\left(u\right)\\ &= \sum_{u,x} P\left(y = 1|x, u\right) P\left(x|z^*\right) P\left(u\right)\\ &= \sum_{u} P\left(y = 1|X = z^*, u\right) P\left(u\right)\\ &= \alpha P\left(y = 1|X = z^*, 1\right) + \left(1 - \alpha\right) P\left(y = 1|X = z^*, 0\right)\\ &= \alpha\delta_{z^*,0} + \left(1 - \alpha\right)\delta_{z^*,1}\\ &= \max\left(\alpha, 1 - \alpha\right)\end{aligned}$$
Since $\mu_\emptyset = 1$, observation is strictly better than intervening on either Z or X.
- $\mu_\emptyset < \mu_{x^*}$: Changing $f_Y\left(x, u\right)$ to $x + u$, we observe $\mu_{x^*} = \mu_{z^*} = 1 + \alpha > \mu_\emptyset = 1$.

Deterministic relations can be modified to a probabilistic one by introducing binary unobserved variables U_X, U_Y, and U_Z and modifying functions for X, Y, and Z to exclusive-or with U_X, U_Y, and U_Z, respectively. By setting probability of them being 1 small enough, one can keep the orders $\mu_\emptyset > \mu_{x^*}$ or $\mu_\emptyset < \mu_{x^*}$.

We devise two models where $\mu_{z^*} > \mu_{x^*}$ and $\mu_{z^*} < \mu_{x^*}$, respectively, for Fig. 3c. Let U_X, U_Y, and U_Z be variable-specific exogenous variables affecting X, Y, and Z, respectively. Let U be the unobserved confounder between X and Y.

- $\mu_{z^*} > \mu_{x^*}$: Let $P(U = 1) = 0.5$ with $\forall_{U_V \in \{U_X, U_Y, U_Z\}} P(U_V = 1) = \epsilon \approx 0$. Let $f_Y(x, u, u_Y) = x \oplus (1 - u) \oplus u_Y$, $f_X(z, u, u_X) = z \oplus u_X \oplus u$, and $f_Z(u_Z) = u_Z$. Then, $\mu_x = (1 - \epsilon) \cdot P(U_X = 1 - x) + \epsilon \cdot P(U_X = x) \approx P(U_X = 1 - x) = 0.5$ while $\mu_{z^*} = 1 - 2\epsilon + \epsilon^2 \approx 1$.
- $\mu_{z^*} < \mu_{x^*}$: Let probabilities of exogenous variables being 1 be 0.5. Let $f_Y(x, u, u_Y) = x + u + u_Y$, $f_X(z, u, u_X) = z \oplus u_X \oplus u$, and $f_Z(u_Z) = u_Z$. Then, $\mu_x = x + 0.5 + 0.5 = x + 1$ and $\mu_z = P(x = 1|do(z)) + 0.5 + 0.5$. Therefore, $\mu_{z^*} = 1.5 < 2 = \mu_{x^*}$.

Appendix B Graphical characterization of POMIS

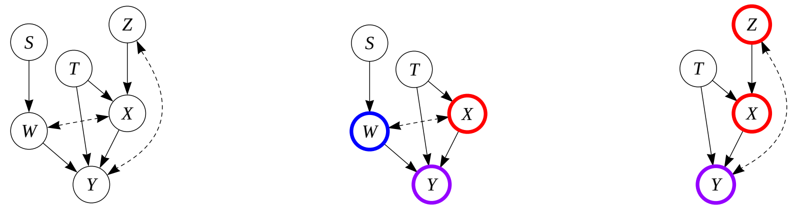
We present an algorithm (Alg. 4) retrieving a MUCT given $[\![G, Y]\!]$.

Algorithm 4 Minimal Unobserved Confounders' Territory

function $\mathsf{MUCT}(G, Y)$
  $H = G\left[An\left(Y\right)_G\right]$
  $\mathbf{Q} = \{Y\}$; $\mathbf{T} = \{Y\}$
  while $\mathbf{Q} \neq \emptyset$ **do**
    remove an element Q_1 from $\mathbf{Q}$
    $\mathbf{W} = \mathsf{CC}\left(Q_1\right)_H$; $\mathbf{T} = \mathbf{T} \cup \mathbf{W}$; $\mathbf{Q} = \left(\mathbf{Q} \cup de\left(\mathbf{W}\right)_H\right) \setminus \mathbf{T}$
  return $\mathbf{T}$

The following proposition and corollary will be used partly to prove propositions and theorems in the main text.

Proposition 10 (Subsumption)**.** *Let* $\mathbf{T}$ *and* $\mathbf{X}$ *be the MUCT and IB for* G *with respect to* Y*, respectively. Then, for any* $\mathbf{Z} \subseteq \mathbf{V} \setminus \{Y\}$*,* $\mu_{\mathbf{s}^*} \geq \mu_{\mathbf{z}^*}$ *where* $\mathbf{S} = (\mathbf{T} \cap \mathbf{Z}) \cup \mathbf{X}$*.*



(a) original G (b) for U_1 (c) for U_2

Figure 6: A causal diagram and colored graphs for unobserved confounders.

Proof. (Case $\mathbf{Z} \supseteq \mathbf{X}$) By definition of IB and Rule 3 of *do*-calculus, $\mu_{\mathbf{z}^*} = \mathbb{E}\left[Y|do(\mathbf{z}^*\left[\mathbf{T}\cup\mathbf{X}\right])\right]$. Since $\mathbf{Z}\cap(\mathbf{T}\cup\mathbf{X}) = \mathbf{S}$, $\mu_{\mathbf{z}^*} = \mu_{\mathbf{s}^*}$.

(Otherwise) Let $\mathbf{X}' = \mathbf{X} \setminus \mathbf{Z}$. Then,

$$\mu_{\mathbf{z}^*} = \sum_{\mathbf{x}'} \mathbb{E}\left[Y|do(\mathbf{z}^*), \mathbf{x}'\right] P\left(\mathbf{x}'|do(\mathbf{z}^*)\right)$$

The first term becomes

$$\begin{aligned}
\mathbb{E}\left[Y|do(\mathbf{z}^*), \mathbf{x}'\right] &= \mathbb{E}\left[Y|do(\mathbf{z}^*), do(\mathbf{x}')\right] && \because \text{Rule 2 } (Y \perp\!\!\!\perp \mathbf{X}' \mid \mathbf{Z})_{G_{\overline{\mathbf{Z}}\underline{\mathbf{X}'}}} \\
&= \mathbb{E}\left[Y|do(\mathbf{z}^*\left[\mathbf{T}\cup\mathbf{X}\right]), do(\mathbf{x}')\right] && \because \text{Rule 3} \\
&\leq \mu_{\mathbf{s}^*} && \because (\mathbf{Z}\cap(\mathbf{T}\cup\mathbf{X}))\cup\mathbf{X}' = \mathbf{S}
\end{aligned}$$

Finally, $\mu_{\mathbf{z}^*} \leq \mu_{\mathbf{s}^*}$ because $\sum_{\mathbf{x}'} \mu_{\mathbf{s}^*} P\left(\mathbf{x}'|do(\mathbf{z}^*)\right) = \mu_{\mathbf{s}^*}$. $\square$

Corollary 11. *Given $[\![G, Y]\!]$, no POMIS intersects with an $(\mathbf{X})_G \setminus \mathbf{X}$ where $\mathbf{X} = \mathsf{IB}(G, Y)$.*

The proposition says that rewards of arms related to intervening on $\mathbf{Z}$ cannot be better than intervening on $\mathbf{Z}$ and the border $\mathbf{X}$ together. Further, since intervening outside of the territory and the border is ineffective, one can intervene only $\mathbf{Z}$ that are inside the territory altogether with the border.

Proposition 4. $\mathsf{IB}(G, Y)$ *is a POMIS given $[\![G, Y]\!]$.*

Proof. Let $\mathbf{X}$ be $\mathsf{IB}(G, Y)$. In this proof, every unobserved variables $\mathbf{U}$ is a binary variable with its domain being $\{0, 1\}$. An easy case is when $\mathbf{T} = \{Y\}$ where $\mathbf{X}$ is the parents of Y in G. We construct a SCM such that

1. Each endogenous variable V associates with an unobserved variable U_V;
2. $f_Y = 1 - (\bigvee \mathbf{u}^Y \oplus (\bigvee \mathbf{x}))$ with $P(\mathbf{u}^Y = 0) \approx 1$;
3. $f_V = (\bigoplus \mathbf{u}^V) \oplus (\bigoplus \mathbf{pa}_V)$ for $V \in \mathbf{V} \setminus \{Y\}$ with $P(u_j) = 0.5$ for every $U_j \in \mathbf{U} \setminus \mathbf{U}^Y$.

Then, $\mathbb{E}[Y|do(\mathbf{X} = 0)] = P(\mathbf{u}^Y = 0) \approx 1$ while all others yield expectations less than or equal to 0.5, otherwise.

Now, we consider a general case where $\mathbf{T} \supset \{Y\}$, that is, there exists at least one unobserved confounder between Y and its ancestors. As a first step, we prove the existence of a SCM M, conforming to $H = G\left[\mathbf{T}\cup\mathbf{X}\right]$, which yields the maximum outcome *only* through $do(\mathbf{X} = 0)$. To do so, we construct a SCM for each unobserved confounder in $H\left[\mathbf{T}\right]$. Let $\mathbf{U}' = \{U_j\}_{j=1}^m$ be unobserved confounders in $H[\mathbf{T}]$. Then, those m individual SCMs $\{M_i\}_{i=1}^m$ will be integrated into a single SCM M so that any intervention other than $\mathbf{x} = 0$ negatively affects the outcome of Y.

We proceed to describe M_i for $U_i \in \mathbf{U}'$. Let $B^{(i)}$ and $R^{(i)}$ be two children of U_i. Let

$$H_i = H\left[De\left(\left\{B^{(i)}, R^{(i)}\right\}\right)_H \cup \left(\mathbf{X} \cap pa\left(De\left(\left\{B^{(i)}, R^{(i)}\right\}\right)_H\right)\right)\right]$$

with all bidirected edges removed except U_i. Functions for variables in $H[\mathbf{T}]$ will be described below. We label (i.e., color code) vertices in $De\left(B^{(i)}\right)_H \setminus De\left(R^{(i)}\right)_H$ as blue and $De\left(R^{(i)}\right)_H \setminus De\left(B^{(i)}\right)_H$

		M_1			M_2			M				
U_1	U_2	$w^{(1)}$	$x^{(1)}$	$y^{(1)}$	$z^{(2)}$	$x^{(1)}$	$y^{(1)}$	w	z	x	y'	y
0	0	1	0	2	0	0	2	00 01	00 00	00 00	10 10	1
	1				1	1	1		01 00	01 00	01 10	
1	0	0	1	1	0	0	2	00 00	00 00	00 01	10 01	
	1				1	1	1		01 00	01 01	01 01	

Table 1: Values with $s = t = 0$ where values for M are shown as binary. y' represents $4y^{(2)} + y^{(1)}$ before the value is binarized.

as red, and $De\left(B^{(i)}\right)_H \cap De\left(R^{(i)}\right)_H$ as purple. Each of $B^{(i)}$ and $R^{(i)}$ perceives that U_i is a parent colored as blue with value u_i and red with value $1-u_i$, respectively. Those blue, red, purple variables are assigned 3 if any of their parents in $\mathbf{X}$ is not 0. Otherwise, their values are determined as follows. For every blue and red vertex, its corresponding function returns the common value of its parents of the same color and returns 3 if colored parents' values are not homogeneous. For every purple vertex, its corresponding function returns 2 if every blue, red, and purple parent is 0, 1, and 2, respectively, and returns 1 if 1, 0, and 1, respectively. For other cases, the function returns 3. One can view the value 3 as a parity propagated to Y.

Now, we integrate m SCMs into one. In M_i, two bits are sufficient to represent every variable. Then, we build a unified SCM where each variable in $\mathbf{T}$ is represented with $2m$ bits where a SCM for U_i will take $2i-1^{\text{th}}$ and $2i^{\text{th}}$ bit (n.b. the right most digit representing $2^0 = 1$ corresponds to the first bit). We then binarize Y by setting 1 if $2i-1^{\text{th}}$ and $2i^{\text{th}}$ bits are 01 or 10 for every $1 \leq i \leq m$ and 0 otherwise. Let $P\left(u_i = 1\right) = 0.5$ for $U_i \in \mathbf{U}'$. This unified SCM M provides a core mechanism to output $Y = 1$ if $do(\mathbf{X} = 0)$ and $Y = 0$ if $do(\mathbf{X} \neq 0)$. If any of variable in $\mathbf{T}$ is intervened, then at least one sub-SCM among m sub-SCMs will be disrupted yielding an expectation smaller than or equal to 0.5.

We now extend the core SCM for $H[\mathbf{T} \cup \mathbf{X}]$ to a SCM for G. However, we can ignore joint probability distributions for any exogenous variables only affecting endogenous variables the outside of H. In addition, functions for endogenous variables lying outside H is irrelevant to the reward distribution. We define a function for $V \in An(Y)_G \setminus \mathbf{T}$ and joint distributions for unobserved variables except $\mathbf{U}'$. Let $f_V = \bigoplus \mathbf{u}^V \oplus \bigoplus \mathbf{pa}_V$ and $P(u_i = 0) = 0.5$ for $U_i \in \mathbf{U}$ whose child(ren) disjoint to $\mathbf{T}$ and $P(u_j = 0) \approx 1$ for $U_j \in \mathbf{U}$ whose child(ren) intersects with $\mathbf{T}$. This ensures that the core mechanism will only be randomly disturbed with a small probability close to 0 preserving the best arm being $do(\mathbf{X} = 0)$. □

We provide an example in Fig. 6 illustrating how sub-SCMs are constructed. Further, values of variables for M_1, M_2 and a unified M are shown in Table 1 with $s = t = 0$.

Proposition 5. *Given* $[\![G, Y]\!]$, $\mathsf{IB}(G_{\overline{\mathbf{W}}}, Y)$ *is a POMIS, for any* $\mathbf{W} \subseteq \mathbf{V} \setminus \{Y\}$.

Proof. Let $\mathbf{T} = \mathsf{MUCT}(G_{\overline{\mathbf{W}}}, Y)$, $\mathbf{X} = \mathsf{IB}(G_{\overline{\mathbf{W}}}, Y)$, and $\mathbf{T}_0 = \mathsf{MUCT}(G, Y)$. We adopt the strategy used in Prop. 4 where a SCM is constructed so that an optimal arm is $do(\mathsf{IB}(G, Y) = 0)$. We first similarly build a SCM for $G\left[\mathbf{T} \cup \mathbf{X}\right]$ while ignoring, for now, *dangling* unobserved confounders between $\mathbf{T}$ and $\mathbf{T}_0 \setminus \mathbf{T}$. Let $\mathbf{U}'$ be such unobserved confounders. Then, we modify the SCM so that a dangling unobserved confounder $U_i \in \mathbf{U}'$ flips (i.e., $0 \leftrightarrow 1$) the value of its endogenous child in $\mathbf{T}$ when $u_i = 1$. Let $P\left(\mathbf{u}' \neq 0\right)$ be close to 0 so that $\mathbb{E}\left[Y | do(\mathbf{X} = 0)\right]$ is close to 1. However, intervening on $\mathbf{X} \neq 0$ or on $\mathbf{Z}$ where $\mathbf{Z} \cap \mathbf{T} \neq \emptyset$ will make corresponding expectations around 0.5 or below. □

Theorem 6. *Given* $[\![G, Y]\!]$, $\mathbf{X} \subseteq \mathbf{V} \setminus \{Y\}$ *is a POMIS if and only if* $\mathsf{IB}(G_{\overline{\mathbf{X}}}, Y) = \mathbf{X}$.

Proof. (If part) A special case of Prop. 5 where $\mathbf{W} = \mathbf{X}$.

(Only if part) Let $\mathbf{W} \subseteq \mathbf{V} \setminus \{Y\}$. Let $\mathbf{T}$ and $\mathbf{X}$ be $\mathsf{MUCT}(G_{\overline{\mathbf{W}}}, Y)$ and $\mathsf{IB}(G_{\overline{\mathbf{W}}}, Y)$ and $\mathbf{T}_0$ and $\mathbf{X}_0$ be $\mathsf{MUCT}(G, Y)$ and $\mathsf{IB}(G, Y)$, respectively. We will prove that $\mathbf{W}$ is not a POMIS when $\mathbf{W} \neq \mathbf{X}$. We can limit $\mathbf{W} \subseteq \mathbf{X}_0 \cup \mathbf{T}_0 \setminus \{Y\}$ due to Prop. 10. Let $\mathbf{X}' = \mathbf{X} \setminus \mathbf{W}$. First, we can observe that

$\mathbf{W} \subseteq An(\mathbf{X})_G$ since otherwise $\mathbf{W} \cap \mathbf{T} \neq \emptyset$, which contradicts that $\mathbf{W}$ is neither a descendant of some variable nor confounded in $G_{\overline{\mathbf{W}}}$. Then,

$$\begin{aligned}
\mu_{\mathbf{w}^*} &= \sum_{\mathbf{x}'} \mathbb{E}\left[Y|do(\mathbf{w}^*), \mathbf{x}'\right] P\left(\mathbf{x}'|do(\mathbf{w}^*)\right) && \because \text{Basic algebra} \\
&= \sum_{\mathbf{x}'} \mathbb{E}\left[Y|do(\mathbf{w}^*), do(\mathbf{x}')\right] P\left(\mathbf{x}'|do(\mathbf{w}^*)\right) && \because \text{Rule 2 } (Y \perp\!\!\!\perp \mathbf{X}' \mid \mathbf{W})_{G_{\overline{\mathbf{W}}\underline{\mathbf{X}'}}} \\
&= \sum_{\mathbf{x}'} \mathbb{E}\left[Y|do\left(\mathbf{w}^*\left[\mathbf{X}\right]\right), do(\mathbf{x}')\right] P\left(\mathbf{x}'|do(\mathbf{w}^*)\right) && \because \text{Rule 3 } (Y \perp\!\!\!\perp (\mathbf{W} \setminus \mathbf{X}) \mid \mathbf{X})_{G_{\overline{\mathbf{X}, \mathbf{W}\setminus\mathbf{X}}}} \\
&\leq \sum_{\mathbf{x}'} \mu_{\mathbf{x}^*} P\left(\mathbf{x}'|do(\mathbf{w}^*)\right) \\
&= \mu_{\mathbf{x}^*}
\end{aligned}$$

Therefore, $\mathbf{W}$ is not a POMIS. □

Appendix C Algorithmic characterization of POMIS

Proposition 7. *Let $\mathbf{T}$ and $\mathbf{X}$ be the $\mathsf{MUCT}(G_{\overline{\mathbf{W}}}, Y)$ and $\mathsf{IB}(G_{\overline{\mathbf{W}}}, Y)$, respectively, relative to G and Y. Then, for any $\mathbf{Z} \subseteq \mathbf{V} \setminus \mathbf{T}$, $\mathsf{MUCT}(G_{\overline{\mathbf{X}\cup\mathbf{Z}}}, Y) = \mathbf{T}$ and $\mathsf{IB}(G_{\overline{\mathbf{X}\cup\mathbf{Z}}}, Y) = \mathbf{X}$.*

Proof. Given that $\mathbf{T}$ is a MUCT for G being intervened on $\mathbf{X}$, additional intervention outside $\mathbf{T}$ does not affect $\mathbf{T}$ being the MUCT and $\mathbf{X}$ being the IB. □

Proposition 8. *Let $H = G_{\overline{\mathbf{X}}}[\mathbf{T} \cup \mathbf{X}]$ where $\mathbf{T}$ and $\mathbf{X}$ are MUCT and IB given $[\![G_{\overline{\mathbf{W}}}, Y]\!]$, respectively. Then, for any $\mathbf{W}' \subseteq \mathbf{T} \setminus \{Y\}$, $H_{\overline{\mathbf{W}'}}$ and $G_{\overline{\mathbf{W}\cup\mathbf{W}'}}$ yield the same MUCT and IB with respect to Y.*

Proof. $G_{\overline{\mathbf{W}\cup\mathbf{W}'}}$ and $H_{\overline{\mathbf{W}'}}$ share the same edges among $\mathbf{T} \cup \mathbf{X}$ except the fact that $\mathbf{X}$ has no parent in $H_{\overline{\mathbf{W}'}}$, which is irrelevant to identifying MUCT $\mathbf{T}$ and IB $\mathbf{X}$ in both diagrams. □

Proposition 12. *If $\mathbf{W}$ is a POMIS, for any $\mathbf{W}' \subset \mathbf{W}$, $\mathbf{W} \setminus \mathbf{W}' \subset \mathsf{MUCT}(G_{\overline{\mathbf{W}'}}, Y) \cup \mathsf{IB}(G_{\overline{\mathbf{W}'}}, Y)$.*

Proof. Otherwise, $\mathbf{W}$ is not a MIS since intervening on $\mathsf{IB}(G_{\overline{\mathbf{W}'}}, Y)$ is preferred to intervening on $\mathbf{W}$. □

We illustrate how the algorithm works with a causal graph G in Fig. 4a. The graph and its manipulated graphs are shown in Fig. 4. Given G and Y, POMISs obtains a MUCT-and-IB induced subgraph with the IB intervened, $G_{\overline{\{S,T\}}}[\{W, X, Y, Z\} \cup \{S, T\}]$ (Line 2), which is the same as G in this example. While recording the first POMIS, the IB for $G_{\overline{\emptyset}}$, POMISs calls $\mathsf{subPOMISs}$ (Line 3) with $(Y, X, Z, T, W, S) \cap \{W, X, Z\} = (X, Z, W)$ for the parameter $\boldsymbol{\pi}$ assuming (Y, X, Z, T, W, S) is acquired among many reversed topological orders of H (Line 3). This corresponds to requesting $\mathsf{subPOMISs}$ to compute the IBs for the passed causal graph with each variable in (X, Z, W) intervened (Lines 6–10). With $G_{\overline{X}}$, its corresponding MUCT and IB are $\{Y\}$ and $\{T, W, X\}$ (Fig. 4b). The IB will be recorded as a POMIS (Line 9) but no subsequent call for $\mathsf{subPOMISs}$ will be made since $(X, Z, W)^{i+1:} \cap \{Y\} = \emptyset$. Given $G_{\overline{Z}}$, the same set of MUCT and IB is obtained as $G_{\overline{X}}$ (Fig. 4c). It is unnecessary to record the IB $\{T, W, X\}$ since it contains X, which precedes Z in the given order (Line 7, 8). The MUCT and IB given $G_{\overline{W}}$ are $\{X, Y, Z\}$ and $\{T, W\}$ (Fig. 4d), respectively. It will record the IB $\{T, W\}$, which disjoints to $\{X, Z\}$. With $\{X, Y, Z\}$ as MUCT, the algorithm checks whether subsequent calls are necessary. However, since both X and Z precede W in the order, no recursive call is made. After all, the complete set of POMISs $\{\{S, T\}, \{T, W\}, \{T, W, X\}\}$ given $[\![G, Y]\!]$ will be returned only examining 4 IBs compared to 2^5=32. If an order such that W precedes Z is considered, then a recursive call for $\mathsf{subPOMISs}$ will be made for $G_{\overline{\{W,Z\}}}$ after examining $G_{\overline{W}}$. It is unclear at this moment how different ordering affects the number of recursive calls.

Theorem 9 (Soundness and Completeness)**.** *Given information $[\![G, Y]\!]$, the algorithm POMISs (Alg. 1) returns all, and only POMISs.*

Proof. Let $\mathbb{P}_{G,Y}$ be all POMISs for $[\![G, Y]\!]$. Let π be an arbitrary sequence of $\mathbf{V}$ such that $\mathbf{V} = \{\pi(1), \pi(2), \ldots, \pi(n)\}$ where $n = |\mathbf{V}|$. Let $A \prec \mathbf{B}$ if $\forall_{B\in\mathbf{B}}\pi^{-1}(A) < \pi^{-1}(B)$ and $A \preceq \mathbf{B}$ be similarly defined. For readability, let $\mathbf{T}^{(Q)} = \mathsf{MUCT}(G_{\overline{\{Q\}}}, Y)$ and $\mathbf{X}^{(Q)} = \mathsf{IB}(G_{\overline{\{Q\}}}, Y)$. We first show

$$\begin{aligned}
\mathbb{P}_{G,Y} &= \{\mathsf{IB}\left(G_{\overline{\mathbf{Z}}}, Y\right)\}_{\mathbf{Z}\subseteq\mathbf{T}\setminus\{Y\}} \\
&= \{\mathsf{IB}\left(G, Y\right)\} \cup \left\{\mathsf{IB}(G_{\overline{\mathbf{Z}}}, Y)\right\}_{\emptyset\neq\mathbf{Z}\subseteq\mathbf{T}\setminus\{Y\}} \\
&= \{\mathsf{IB}\left(G, Y\right)\} \cup \bigcup_{Q\in\mathbf{T}\setminus\{Y\}} \left\{\mathsf{IB}(G_{\overline{\{Q\},\mathbf{Z}}}, Y)\right\}_{\mathbf{Z}\subseteq\mathbf{T}\setminus\{Q,Y\}:Q\prec\mathbf{Z}} \\
&= \{\mathsf{IB}\left(G, Y\right)\} \cup \bigcup_{Q\in\mathbf{T}\setminus\{Y\}:Q\preceq\mathbf{X}^{(Q)}} \left\{\mathsf{IB}(G_{\overline{\mathbf{X}^{(Q)}}}[\mathbf{T}^{(Q)} \cup \mathbf{X}^{(Q)}]_{\overline{\mathbf{Z}}}, Y)\right\}_{\mathbf{Z}\subseteq\mathbf{T}^{(Q)}\setminus\{Y\}:Q\prec\mathbf{Z}}
\end{aligned}$$

The third line partitions $2^{\mathbf{T}\setminus\{Y\}}_{>0}$, the power set of $\mathbf{T} \setminus \{Y\}$ excluding an empty set, so that sets of variables in each partition include the same variable, e.g., Q, with no variable smaller than Q with respect to π. The fourth line follows from Prop. 7 (change for $\mathbf{Z}$) and Prop. 8 (change of arguments for IB). An additional constraint for Q being $Q \preceq \mathbf{X}^{(Q)}$ avoids redundant computations since two different variables, e.g., Q' and Q'', can both yield the same MUCT and IB, $\mathbf{T}^{(Q')} = \mathbf{T}^{(Q'')}$ and $\mathbf{X}^{(Q')} = \mathbf{X}^{(Q'')}$. Since

$$\left\{\mathsf{IB}(G_{\overline{\mathbf{X}^{(Q)}}}[\mathbf{T}^{(Q)} \cup \mathbf{X}^{(Q)}]_{\overline{\mathbf{Z}}}, Y)\right\}_{\mathbf{Z}\subseteq\mathbf{T}^{(Q)}\setminus\{Y\}} = \mathbb{P}_{G_{\overline{\mathbf{X}^{(Q)}}}[\mathbf{T}^{(Q)}\cup\mathbf{X}^{(Q)}],Y},$$

we can rewrite $\mathbb{P}_{G,Y}$ as (abusing notations)

$$\mathbb{P}_{G,Y} = \{\mathsf{IB}\left(G, Y\right)\} \cup \bigcup_{Q\in\mathbf{T}\setminus\{Y\}:Q\preceq\mathbf{X}^{(Q)}} \mathbb{P}^{Q\prec\mathbf{Z}}_{G_{\overline{\mathbf{X}^{(Q)}}}[\mathbf{T}^{(Q)}\cup\mathbf{X}^{(Q)}],Y}$$

where the superscript $Q \prec \mathbf{Z}$ serves as an additional constraint for efficiency.

The algorithm implements the above equality where parameter $\boldsymbol{\pi}$ in $\mathsf{subPOMISs}$ carries the constraint $Q \prec \mathbf{Z}$ and parameter $\mathbf{O}$ conveys the constraint $Q \preceq \mathbf{X}^{(Q)}$. Let $\mathbf{W}$ be an arbitrary POMIS in $\mathbb{P}_{G,Y}$. We can index its element (i.e., variable) $\mathbf{W} = \{W_i\}_{i=1}^{|\mathbf{W}|}$ with respect to π so as to $\pi^{-1}(W_i) < \pi^{-1}(W_j)$ if $i < j$. Then, there exists a sequence of recursive calls of $\mathsf{subPOMISs}$ where variables in $\mathbf{W}$ are sequentially determined to be intervened skipping the already determined to do so (i.e., $\mathbf{W} \cap \mathsf{IB}(G', Y)$ where G' is the first argument of $\mathsf{subPOMISs}$). Therefore, the algorithm completely enumerates all POMISs effectively avoiding redundant computations. □

Appendix D Experiments

Task 1: $P(U_{X_1} = 1) = 0.54$, $P(U_{X_2} = 1) = 0.67$, $P(U_Y = 1) = 0.58$, $P(U_{Z_1} = 1) = 0.54$, and $P(U_{Z_2} = 1) = 0.44$, and functions:

$$\begin{aligned}
f_{Z_1}(u_{Z_1}) &= u_{Z_1} \\
f_{Z_2}(u_{Z_2}) &= u_{Z_2} \\
f_{X_1}(z_1, z_2, u_{X_1}) &= z_1 \oplus z_2 \oplus u_{X_1} \\
f_{X_2}(z_1, z_2, u_{X_2}) &= 1 \oplus z_1 \oplus z_2 \oplus u_{X_2} \\
f_Y(x_1, x_2, u_Y) &= (x_1 \wedge x_2) \vee u_Y
\end{aligned}$$

Task 2: $P(U_X = 1) = 0.11$, $P(U_Y = 1) = 0.15$, $P(U_Z = 1) = 0.6$, and $P(U_{XY} = 1) = 0.51$ and functions:

$$\begin{aligned}
f_Z(u_Z) &= u_Z \\
f_X(z, u_X, u_XY) &= u_X \oplus u_{XY} \oplus z \\
f_Y(x, u_Y, u_XY) &= 1 \oplus u_Y \oplus u_{XY} \oplus x
\end{aligned}$$

Task 3: $P(U_S = 1) = 0.45$, $P(U_T = 1) = 0.81$, $P(U_W = 1) = 0.07$, $P(U_X = 1) = 0.06$, $P(U_Y = 1) = 0.06$, $P(U_Z = 1) = 0.05$, $P(U_{WX} = 1) = 0.51$, and $P(U_{YZ} = 1) = 0.54$, and functions:

$$\begin{aligned}
f_S(u_S) &= u_S \\
f_T(u_T) &= u_T \\
f_W(s, u_W, u_{WX}) &= u_W \oplus u_{WX} \oplus s \\
f_Z(u_Z, u_{YZ}) &= u_Z \oplus u_{YZ} \\
f_X(t, z, u_X, u_{WX}) &= 1 \oplus t \oplus z \oplus u_X \oplus u_{WX} \\
f_Y(t, w, x, u_Y, u_{YZ}) &= t \oplus w \oplus x \oplus u_Y \oplus u_{YZ}
\end{aligned}$$

Acknowledgments

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

References

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

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems 28*, pages 1342–1350. 2015.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404 – 1422, 2012.

Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *International Conference on Machine Learning*, pages 521–529, 2014.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *Proceedings of Conference On Learning Theory (COLT)*, pages 355–366, 2008.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1156–1164, 2017.

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

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213, 2012.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems 30*, pages 7021–7031, 2017.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems 29*, pages 1181–1189. 2016.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? Technical Report R-36, Purdue AI Lab, Department of Computer Science, Purdue University, 2018.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Proceedings of The 27th Conference on Learning Theory*, pages 975–999, 2014.

Pedro A. Ortega and Daniel A. Braun. Generalized Thompson sampling for sequential decision-making and causal inference. *Complex Adaptive Systems Modeling*, 2(2), 2014.

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

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. Second ed., 2009.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3057–3066, 2017.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. A Bradford Book, 2001.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Csaba Szepesvári. *Algorithms for reinforcement learning*. Morgan and Claypool, 2010.

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

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002.

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