Learning Good Interventions in Causal Graphs via Covering

Ayush Sawarni[1] Rahul Madhavan[1] Gaurav Sinha[2] Siddharth Barman[1]

[1]Indian Institute of Science, Bangalore
[2]Microsoft Research, Bangalore

Abstract

We study the causal bandit problem that entails identifying a near-optimal intervention from a specified set $\mathcal{A}$ of (possibly non-atomic) interventions over a given causal graph. Here, an optimal intervention in $\mathcal{A}$ is one that maximizes the expected value for a designated reward variable in the graph, and we use the standard notion of simple regret to quantify near optimality. Considering Bernoulli random variables and for causal graphs on N vertices with constant in-degree, prior work has achieved a worst case guarantee of $\widetilde{O}(N/\sqrt{T})$ for simple regret. The current work utilizes the idea of covering interventions (which are not necessarily contained within $\mathcal{A}$) and establishes a simple regret guarantee of $\widetilde{O}(\sqrt{N/T})$. Notably, and in contrast to prior work, our simple regret bound depends only on explicit parameters of the problem instance. We also go beyond prior work and achieve a simple regret guarantee for causal graphs with unobserved variables. Further, we perform experiments to show improvements over baselines in this setting.

1 INTRODUCTION

Causal Bayesian Networks (CBNs) are a prominent paradigm for modelling many real world problems Pearl [2009]. Recent applications include language modelling Sevilla [2020], medicine Koch et al. [2017], Caillet et al. [2015], Lee et al. [2018], robotics Yoshida and Nakadai [2012] and computational advertising Bottou et al. [2013].

While CBNs have been the focus of research for decades, questions related to online learning in the CBN context have been studied only recently. Prototypical questions at the interface of online learning and CBNs are captured by the causal bandits model. Causal bandits—first introduced by Lattimore et al. [2016]—merges concepts from CBNs and multi-armed bandits (MABs) to provide a framework for optimized learning over CBNs. The focus of the current work is to obtain simple regret guarantees in the causal bandit setup.

A CBN consists of a causal graph—a directed acyclic graph $\mathcal{G} = (\mathcal{V}, E)$—that provides the direction of causation amongst $N := |\mathcal{V}|$ random variables. That is, in the given graph $\mathcal{G}$, the vertices $\mathcal{V}$ correspond to variables and E corresponds to the set of (directed edges) causal relations between these variables. Here, each variable is some function of its parents. Complementarily, a variable that has no parents (known as an exogenous or independent variable) is a random variable over some distribution; see Pearl [2009] for a textbook treatment of CBNs.

We will, throughout, consider $\mathcal{V}$ to be Bernoulli random variables. In the causal bandit problem, one designates a particular vertex in the causal graph $V_N \in \mathcal{V}$ as the reward variable and seeks to optimize for the expected value of this reward variable. The optimization is over a specified set $\mathcal{A}$, which consists of interventions in the causal graph. Interventions, known as $\mathrm{do}()$ operations, fix the values of certain variables, irrespective of their parents. Specifically, in an intervention $A = \mathrm{do}(S = s)$, we fix the value of each variable $i \in S \subseteq \mathcal{V}$ to be the ith component of the given binary assignment $s \in \{0,1\}^{|S|}$. Under intervention $A = \mathrm{do}(S = s)$, the un-intervened variables (in $\mathcal{V} \backslash S$) then follow the causal relations that remain. The goal of the causal bandit learner is to perform exploratory interventions, for a given number of rounds T, and at the end of this time horizon, the learner needs to identify a near-optimal intervention from within the target set $\mathcal{A}$. That is, the overarching objective in the causal bandit problem is to identify an intervention $A \in \mathcal{A}$ under which the expected value of the reward variable, V_N, is as high as possible.

As in the classic multi-armed bandits literature Lattimore and Szepesvári [2020], Slivkins et al. [2019], the notion of simple regret is used to quantify near optimality in the causal

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:1827–1836.

bandit setup. In particular, for an algorithm that selects intervention $A_T \in \mathcal{A}$ after after T rounds, the simple regret is the difference (in expectation) between optimal reward and the reward induced by A_T.

Most prior works on causal bandits Lattimore et al. [2016], Maiti et al. [2022], Sen et al. [2017a], Lu et al. [2020], Nair et al. [2021], Sen et al. [2017b], Lu et al. [2021, 2022] address the problem with $\mathcal{A}$ restricted to atomic interventions. That is, these works hold when each intervention $A \in \mathcal{A}$ fixes some single vertex in the causal graph. Other causal bandit results Varici et al. [2022], Xiong and Chen [2023] consider settings in which all the causal relations in $\mathcal{G}$ are confined to be linear functions. In this active thread of research on causal bandits, a notable exception is the work of Yabe et al. Yabe et al. [2018], which addresses the broad setting of non-atomic interventions over general graphs and holds without assumptions on the causal relations.

Indeed, such a general form of the problem is nontrivial. In particular, the number of non-atomic interventions under consideration can be exponential in the number of variables N. Hence, a naive approach of sampling for each intervention $A \in \mathcal{A}$ can yield a simple regret proportional to $\tilde{\mathcal{O}}(\sqrt{\exp(N)/T})$. Interestingly, for the general form of the causal bandit problem, Yabe et al. Yabe et al. [2018] achieve a worst-case guarantee on simple regret of $\tilde{\mathcal{O}}(\sqrt{N^2/T})$; here, the $\tilde{\mathcal{O}}$ notation subsumes the dependence on the maximum in-degree in the causal graph and logarithmic factors. In particular, the regret guarantee of Yabe et al. [2018] depends on the optimal value of a proposed optimization problem. We also note that, even during exploration, Yabe et al. Yabe et al. [2018] consider interventions only from the target set $\mathcal{A}$ and use the solution of the proposed optimization problem to guide the exploration.

While the algorithm of Yabe et al. Yabe et al. [2018] is applicable with significant generality, it has certain key limitations. Firstly, the algorithm entails solving a non-convex optimization problem that is "time-consuming to solve" (see page 8 in Section 5 of Yabe et al. [2018]). In fact their own experiment implementations do not explicitly solve the optimization problem. Next, the regret bound is in terms of a quantity that is analytically unwieldy to estimate. In particular, their simple regret guarantee is $O\left(\sqrt{\frac{\gamma^* \log(|\mathcal{A}|T)}{T}}\right)$, where γ^* is the optimal value of a (hard to compute) non-convex optimization problem and it satisfies $\gamma^* = O(N^2)$. In addition, the regret guarantee in Yabe et al. [2018] holds for time horizon $T \gtrsim N^{16}$. Finally, their algorithm expects full observability, and does not allow for the presence of unobserved (hidden) variables in the causal graph.

The current work develops an algorithm that overcomes the above-mentioned limitations and continues to address the general form of the causal bandit problem. We use the idea of covering interventions and improve the simple regret guarantee. We also go beyond Yabe et al. [2018] and achieve a simple regret guarantee for causal graphs with unobserved variables.

1.1 OUR CONTRIBUTIONS

We present an algorithm to minimize simple regret in the causal bandit problem. Here, the learner is given a causal graph $\mathcal{G}$ on N Bernoulli random variables and a set $\mathcal{A}$ of (possibly non-atomic) interventions over $\mathcal{G}$. The learner's objective is to identify, within $\mathcal{A}$, an intervention that maximizes the expected value for a designated reward variable in $\mathcal{G}$. Furthermore, we consider a model wherein, while a near-optimal intervention is required from the target set $\mathcal{A}$, the learner is not confined to $\mathcal{A}$ during the exploration phase. In particular, we use the construct of covering interventions (see Definition 1) during exploration and show that this flexibility leads to multiple improvements over prior work. Indeed, this model is applicable in many settings wherein the learner is not confined to the target set during exploration. Consider, as stylized examples: (i) the display advertising context, wherein, during testing, one can intervene upon features, which during deployment, cannot be altered, and (ii) robotic control, in which, during simulations, hypothetical configurations can be deployed.

In fact, our result is robust enough to be used in settings where certain variables cannot be intervened upon even during exploration. One can consider such 'off-limits' variables as unobserved and then utilize our extension to graphs with unobserved parts (see Section 4). The list below summarizes our main contributions:

- For the causal bandit problem, we improve the worst-case guarantee for simple regret from $\tilde{\mathcal{O}}(\sqrt{N^2/T})$ to $\tilde{\mathcal{O}}(\sqrt{N/T})$.[1] Here, the $\tilde{\mathcal{O}}(\cdot)$ notation subsumes the dependence on the maximum in-degree d in the graph and logarithmic factors; see Theorem 1 for an explicit bound. Our algorithm can address arbitrary causal graphs. Though, as in prior works Yabe et al. [2018], Acharya et al. [2018], our result is particularly relevant for graphs in which the maximum in-degree d is sufficiently smaller than N.
- We obtain a novel simple regret algorithm for causal graphs with unobserved variables. This extension addresses the most general setting for causal Bayesian networks (see Definition 1.3.1 in Pearl [2000]) and addresses a key limitation of almost all[2] prior works on causal bandits. We detail the extension in Section 4.

[1] As mentioned previously, T denotes the time horizon (i.e., number of exploratory interventions) and N denotes the number of vertices in the causal graph.

[2] The exceptions here are the recent works of Maiti et al. Maiti et al. [2022] along with Xiong and Chen Xiong and Chen [2023]. These works are discussed at the end of the section.

- Our experiments show a marked improvement on the baselines from prior work (see Section 5), thereby substantiating the theoretical guarantees.

Our worst-case guarantee for simple regret is in terms of only the explicit parameters, such as the number of variables N and the maximum in-degree in $\mathcal{G}$; see Theorem 1. By contrast, the simple regret bound provided in Yabe et al. [2018] depends on analytically complex quantities. In addition, our guarantee holds for time horizon $T \gtrsim N^3$. This is a marked improvement over Yabe et al. [2018], which requires $T \gtrsim N^{16}$. In fact, our algorithm (Algorithm 1) is notably simple – we view this as a positive feature, which aids in implementation and adaptation of the developed method. Here, it is also relevant to note that the key technical contribution of the work is the involved regret analysis (see Section 3.1).

Covering interventions as a complementary tool for exploration. We note that covering interventions do not conform to the existing causal-bandit framework of exploring solely within the specified set of interventions $\mathcal{A}$. However, instead of viewing $\mathcal{A}$ as a confined set of 'arms,' one can work with the enriched perspective that causal bandits are an optimization problem. Indeed, the goal of the optimization problem is to identify the best intervention in $\mathcal{A}$, but—similar to many other optimization methods—exploration can happen outside the feasible region (i.e., outside $\mathcal{A}$). In this spirit, the use of covering interventions can be identified as a complementary exploration model. This model leverages the richer context of the causal bandits setting (e.g., the causal graph itself) and, as mentioned previously, is potentially applicable in various real-world contexts. Overall, covering interventions are theoretically interesting and enable notable improvements, including novel simple regret guarantees with unobserved variables.

1.2 ADDITIONAL RELATED WORK

Lattimore et al. Lattimore et al. [2016] first addressed the causal bandit, though only for parallel causal graphs and with atomic interventions. Maiti et al. Maiti et al. [2022] extended this work on atomic interventions to provide simple regret guarantees in the presence of unobserved or hidden variables. An importance sampling based approach was studied in Sen et al. [2017a] to identify atomic soft interventions that minimize simple regret. Lu et al. Lu et al. [2020] provide guarantees for cumulative regret for general causal graphs (which include hidden variables). Nair et al. Nair et al. [2021] looked at cumulative as well as simple regret in case of the budgeted setting where the observation-intervention trade-off was studied when interventions are costlier than observations. Sen et al. Sen et al. [2017b] extend the model causal bandits to include contextual causal bandits and study cumulative regret in this context. Lu et al. Lu et al. [2021] study cumulative regret in the case where the full graph structure is not known. The work Lu et al. [2022] extends the model for causal bandits to include causal Markov decision processes (C-MDPs) using a modification of the algorithm in Azar et al. [2017].

There are two recent works that focus on non-atomic interventions in the causal bandit context. The paper by Varici et al. Varici et al. [2022] studies cumulative regret for causal bandits with non-atomic interventions, albeit in the specific context of linear structural equation models. Xiong and Chen Xiong and Chen [2023] obtain sample-complexity bounds for identification of near-optimal interventions, with a particular focus on binary generalized linear models (BGLMs). The worst-case sample complexity guarantee obtained in Xiong and Chen [2023] is proportional to the size of the intervention set $\mathcal{A}$, i.e., proportional to $|\mathcal{A}|$. By contrast, the simple regret bound obtained in the current work has only a logarithmic dependence on $|\mathcal{A}|$; recall that $|\mathcal{A}|$ can be exponentially large. Xiong and Chen Xiong and Chen [2023] also address the case of unobserved (hidden) variables. However, this work assumes identifiability (the fact that all interventional distributions can be estimated through observations alone). We require no such assumption.

Apart from these works on causal bandits, we utilize the idea of covering interventions proposed by Acharya et al. Acharya et al. [2018]. They use covering interventions for distribution learning and testing problems over causal graphs. On the other hand, we use covering interventions for simple regret minimization. It is important to note that a direct use of the distribution learning algorithm (Algorithm 3) from Acharya et al. [2018] leads to a suboptimal regret bound for the causal bandit problem. Specifically, the learning algorithm of Acharya et al. Acharya et al. [2018] requires $\tilde{\mathcal{O}}(N^2\varepsilon^{-4})$ samples to learn interventional distributions up to a total variation distance of ε; see Theorem 3.4 in Acharya et al. [2018]. Hence, if used for identifying a near-optimal intervention in $\mathcal{A}$, this method would incur $\tilde{\mathcal{O}}\left(\frac{\sqrt{N}}{T^{1/4}}\right)$ simple regret.

2 NOTATION AND PRELIMINARIES

We study the causal bandit problem over causal graphs $\mathcal{G} = (\mathcal{V}, E)$. In the given (directed and acyclic) graph $\mathcal{G}$ the vertices, $\mathcal{V}$, correspond to Bernoulli random variables and E is the set of directed edges that capture causal relations between these variables.

We will use V_i or i, interchangeably, to refer to the ith node of the given causal graph $\mathcal{G}$. Since $\mathcal{G}$ is directed and acyclic, it admits a topological ordering. We will, throughout, assume that the vertices in $\mathcal{V}$ are indexed to respect a topological order, i.e., for each pair of indices $i < j$, vertex V_i appears before V_j in the topological order. Note that for

any subset of vertices $\mathcal{U} \subseteq \mathcal{V}$ the indexing of the vertices within $\mathcal{U}$ follows the topological ordering of these vertices. Furthermore, in the set $\mathcal{V}$, the last vertex with respect to the indexing (and, equivalently, the topological ordering) is the designated reward variable. That is, in a causal graph with $N := |\mathcal{V}|$ vertices, V_N is the reward variable.

Write $\mathrm{Pa}(i)$ to denote the set of parents of node V_i. Also, we define the set of parents for a subset of vertices $\mathcal{U} \subseteq \mathcal{V}$ as $\mathrm{Pa}(\mathcal{U}) := (\cup_{V \in \mathcal{U}} \mathrm{Pa}(V)) \setminus \mathcal{U}$. We use the following notations to indicate subsets of the vertices: write $[i, j] := \{V_i, V_{i+1}, V_{i+2} \dots V_j\}$ and, similarly, $(i, j] = [i + 1, j]$, $(i, j) = [i + 1, j - 1]$ and $[i, j) = [i, j - 1]$. Write the ancestor set $\mathrm{Ac}(i) := [1, i) \setminus \mathrm{Pa}(i)$, i.e., $\mathrm{Ac}(i)$ denotes the set of vertices that precede V_i in the topological ordering, excluding the parents $\mathrm{Pa}(V_i)$.

An intervention is defined as an $N = |\mathcal{V}|$ dimensional vector $A \in \{0, 1, *\}^N$ that encapsulates the values assigned to each vertex in $\mathcal{G}$; in particular, $A_i = *$ denotes that V_i is not intervened upon, while $A_i = 1$ and $A_i = 0$ denote that, in the intervention, V_i is set to 1 and 0, respectively. In addition, $\mathcal{V}(A) := \{V_i \in \mathcal{V} : A_i = *\}$ denotes the set of vertices that are not intervened under A. Also, for any subset of vertices $\mathcal{U} \subseteq \mathcal{V}$, write $\mathcal{V}_{\mathcal{U}}(A) := \mathcal{U} \cap \mathcal{V}(A)$.

Binary vectors $\mathbf{z} \in \{0, 1\}^N$ will be used to denote an assignment to the vertices (random variables) in $\mathcal{V}$. Here, $\mathbf{z}_i$ denotes the assignment to vertex V_i. For any subset of vertices $\mathcal{U} \subseteq \mathcal{V}$, we will use $\mathbf{z}_U \in \{0, 1\}^{|\mathcal{U}|}$ to denote an assignment to the vertices in $\mathcal{U}$. Let $Z(A)$ denote the set of all binary assignments that comply with an intervention A and have the reward $V_N = 1$, i.e., $Z(A) := \{\mathbf{z} \in \{0, 1\}^N : \mathbf{z}_i = A_i$, for all $i \in \mathcal{V} \setminus (\mathcal{V}(A))$, and $\mathbf{z}_N = 1\}$.

We use the following short-hand notations in our analysis to denote the conditional and interventional probability distributions:

$$\begin{aligned}
\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_U\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | U = \mathbf{z}_U\right]. \\
\mathcal{P}_{\mathbf{z}_U}\left(\mathbf{z}_i\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | \mathrm{do}\left(U = \mathbf{z}_U\right)\right] \\
&= \mathbb{P}_{\mathrm{do}(U = \mathbf{z}_U)}\left[V_i = \mathbf{z}_i\right]. \\
\mathcal{P}_{\mathbf{z}_U}\left(\mathbf{z}_i \mid \mathbf{z}_W\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | \mathrm{do}\left(U = \mathbf{z}_U\right), W = \mathbf{z}_W\right]. \\
\mathcal{P}_A\left(\mathbf{z}_i \mid \mathbf{z}_W\right) &= \mathbb{P}\left[V_i = \mathbf{z}_i | \mathrm{do}\left(A\right), W = \mathbf{z}_W\right].
\end{aligned}$$

It is important to note that intervening on all parent nodes of a vertex is the same as conditioning on them

$$\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(i)}}\left(\mathbf{z}_i\right) = \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \tag{1}$$

We use $\mu\left(A\right)$ to denote the expected reward under intervention A, i.e., $\mu\left(A\right) = \mathbb{P}\left[V_N = 1 | \mathrm{do}(A)\right]$. Specifically,

$$\mu\left(A\right) = \sum_{\mathbf{z} \in Z(A)} \prod_{i \in \mathcal{V}(A)} \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$$

We use $\widehat{\mu}(A)$ and $\widehat{\mathcal{P}}\left(\right)$ to denote the estimates for the corresponding quantities, and $\Delta\mathcal{P}\left(\right)$ to denote the error in the estimates. In particular, for an empirical estimation in which vertex V_i is sampled T_i times, with parents taking value $\mathbf{z}_{\mathrm{Pa}(i)} \in \{0, 1\}^{|\mathrm{Pa}(i)|}$, we have estimate

$$\widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) = \frac{\sum_{s=1}^{T_i} \mathbb{I}[Y_{i,s} = \mathbf{z}_i]}{T_i},$$

where $Y_{i,s}$ is the s-th sample of vertex V_i. In addition, we have

$$\begin{aligned}
\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) &= \mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) - \widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \\
\widehat{\mu}(A) &= \sum_{\mathbf{z} \in Z(A)} \prod_{i \in \mathcal{V}(A)} \widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)
\end{aligned} \tag{2}$$

Recall that, in the causal bandits problem, the objective is to find—from within a specified collection of interventions $\mathcal{A}$—an intervention with maximum possible expected reward. We will write $A^* \in \mathcal{A}$ to denote the optimal intervention and $\mu(A^*)$ for the optimal reward, i.e., $\mu(A^*) = \max_{A \in \mathcal{A}} \mu(A)$. Also, for any algorithm, let $A_T \in \mathcal{A}$ be the (randomized) output computed after T rounds; in each round the algorithm performs an intervention and observes a sample under it.[3] The simple regret of the algorithm is defined as

$$R_T = \mathbb{E}\left[\mu(A^*) - \mu(A_T)\right]. \tag{3}$$

3 FINDING NEAR-OPTIMAL INTERVENTION VIA COVERING

To find a near-optimal intervention from the given set of interventions $\mathcal{A}$ (specifically, to bound simple regret), instead of directly performing each $A \in \mathcal{A}$, we utilize interventions from a curated set of interventions $\mathcal{I}$, referred to as the covering intervention set (see Definition 1). The obtained samples are then used to estimate the interventional distribution for each $A \in \mathcal{A}$ and, hence, find a near-optimal intervention within $\mathcal{A}$. The notion of covering intervention set was formulated in Acharya et al. [2018] and is defined next.

Definition 1 (Covering Intervention Set)**.** *A collection of interventions $\mathcal{I}$ is said to be a* covering intervention set *iff, for each vertex $i \in \mathcal{V}$ and every assignment $\mathbf{z}_{\mathrm{Pa}(i)} \in \{0, 1\}^{|\mathrm{Pa}(i)|}$, there exists an intervention $I \in \mathcal{I}$ with the properties that*

- *Vertex i not intervened in I (i.e., $I_i = *$).*
- *Every vertex in $\mathrm{Pa}(i)$ is intervened (i.e., $I_p \neq *$, for all $p \in \mathrm{Pa}(i)$).*
- *I restricted to $\mathrm{Pa}(i)$ has the assignment $\mathbf{z}_{\mathrm{Pa}(i)}$ (i.e., $I_p = \mathbf{z}_{\mathrm{Pa}(i),p}$ for all $p \in \mathrm{Pa}(i)$).*

[3] Note that while the computed intervention must be contained in set $\mathcal{A}$, the interventions performed in the T rounds are not necessarily from $\mathcal{A}$.

It is shown in Acharya et al. [2018] that, for any causal graph $\mathcal{G}$ with N vertices and in-degree at most d, one can construct—using a randomized method—a covering intervention set $\mathcal{I}$ of size $O\left(d\, 2^d \log(NT)\right)$.

Specifically, for count $k = 3d\, 2^d(\log N + 2d + \log T)$, one can populate k interventions $I \in \{0, 1, *\}^N$ as follows: for each variable $i \in \mathcal{V}$, independently, set

$$I_i = \begin{cases} 0 & \text{with probability } \frac{d}{2(1+d)}, \\ 1 & \text{with probability } \frac{d}{2(1+d)}, \\ * & \text{otherwise.} \end{cases}$$

All the constructed k interventions constitute the set $\mathcal{I}$. This randomized construction is known to succeed (in providing a covering interventions set) with probability at least $(1 - 1/T)$. Formally,[4]

Lemma 1 (Acharya et al. [2018])**.** *For any moderately large $T \in \mathbb{Z}_+$, every causal graph $\mathcal{G}$—with N vertices and in-degree at most d—admits a covering intervention set $\mathcal{I}$ of size $k = 3d\, 2^d(\log N + 2d + \log T)$. Furthermore, such a set $\mathcal{I}$ can be found with probability at least $(1 - 1/T)$.*

We will write ConstructCover$(\mathcal{G})$ to denote the randomized construction of $\mathcal{I}$ mentioned above. ConstructCover$(\cdot)$ will be used as a subroutine in our simple-regret algorithm (Algorithm 1).

Theorem 1, stated below, is the main result of this section. The theorem asserts that, for causal graphs with constant in-degree and N vertices, Algorithm 1 achieves a simple regret of $\widetilde{O}\left(\sqrt{N/T}\right)$.

Given a causal graph $\mathcal{G}$ and a collection of interventions $\mathcal{A}$, Algorithm 1 first obtains a covering intervention set $\mathcal{I}$, for the graph $\mathcal{G}$, via the subroutine ConstructCover. Then, the algorithms performs, $T/|\mathcal{I}|$ times, each intervention $I \in \mathcal{I}$. Since $\mathcal{I}$ is a covering intervention set, for each vertex $\widehat{i} \in \mathcal{V}$, there exists an intervention $\widehat{I} \in \mathcal{I}$ under which all the parents $\mathrm{Pa}\left(\widehat{i}\right)$ are intervened upon, but $\widehat{i}$ itself is not. The intervention $\widehat{I}$ has already been performed $T/|\mathcal{I}|$ times by the algorithm. Using these $T/|\mathcal{I}|$ independent samples and for a specific assignment $\mathbf{z}_{\mathrm{Pa}\left(\widehat{i}\right)}$ (induced under $\widehat{I}$), we have the estimate $\widehat{\mathcal{P}}\left(\mathbf{z}_{\widehat{i}} \mid \mathbf{z}_{\mathrm{Pa}\left(\widehat{i}\right)}\right)$. Hence, for every vertex $i \in \mathcal{V}$ and every assignment $\mathbf{z}_{\mathrm{Pa}(i)}$, the algorithm has an estimate $\widehat{\mathcal{P}}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$ in hand. Using these probability estimates, the algorithm computes the reward estimates $\widehat{\mu}(A)$ for each intervention $A \in \mathcal{A}$; see equation (2). Finally, enumerating over the given set $\mathcal{A}$, the algorithm returns the intervention with the maximum estimated reward. It is relevant to note that this patently simple algorithm requires a technically involved regret analysis (detailed in Section 3.1). Indeed, the analysis is a key contribution of the current work.

[4]This lemma is a direct implication of Lemma 2 from Acharya et al. [2018], instantiated with $\delta = \frac{1}{T}$, $K = 2$.

Algorithm 1 Covering Interventions Algorithm

Input: Causal graph $\mathcal{G}$, target intervention set $\mathcal{A}$, and time horizon $T \in \mathbb{Z}_+$.

1: Set $\mathcal{I} \leftarrow$ ConstructCover$(\mathcal{G})$.
2: For each $I \in \mathcal{I}$, intervene with $\mathrm{do}(I)$ and collect $\frac{T}{|\mathcal{I}|}$ samples.
3: **for** each intervention $A \in \mathcal{A}$ **do**
4: Compute $\widehat{\mu}(A)$ using equation (2).
5: **end for**
6: **return** $\arg\max_{A \in \mathcal{A}} \widehat{\mu}(A)$.

Theorem 1. *Let $\mathcal{G}$ be any given causal graph with N vertices and in-degree at most d. Also, let $\mathcal{I}$ be a covering intervention set of $\mathcal{G}$. Then, Algorithm 1—when executed for any (moderately large) time horizon T—achieves simple regret*

$$R_T = O\left(\sqrt{\frac{N|\mathcal{I}|\log\left(|\mathcal{A}|T\right)}{T}}\right).$$

Hence, using Lemma 1, we obtain the following bound on the simple regret of Algorithm 1

$$R_T = O\left(\sqrt{\frac{N\, d2^d\, \log|\mathcal{A}|}{T}}\log T\right).$$

For graphs with additional structure (e.g. bounded out degree or trees), one can obtain covering intervention sets with size smaller than the one provided in Lemma 1 (see Lemma 2 in Acharya et al. [2018]). Since the regret guarantee of Algorithm 1 depends on the size of the covering intervention set, the simple regret bound improves for such specific graphs.

3.1 REGRET ANALYSIS

For each intervention $A \in \mathcal{A}$, the estimate $\widehat{\mu}(A)$ can be expressed as

$$\widehat{\mu}\left(A\right) = \sum_{\mathbf{z} \in Z(A)} \prod_{i \in \mathcal{V}(A)} \left(\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) + \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)\right).$$

Expanding the product, we obtain

$$\widehat{\mu}\left(A\right) = \mu(A) + \sum_{\mathbf{z} \in Z(A)} \left(\sum_{i \in \mathcal{V}(A)} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \times \prod_{j \in \mathcal{V}(A), j \neq i} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \; + \; \mathcal{L}_{\mathbf{z}} \right).$$

Here, $\mathcal{L}_{\mathbf{z}}$ represents all the product entries in the expansion that include more than one error term of the form $\Delta\mathcal{P}(\cdot \mid \cdot)$.

Specifically,

$$\mathcal{L}_{\mathbf{z}} = \sum_{k=2}^{|\mathcal{V}(A)|} \sum_{\substack{U \subseteq \mathcal{V}(A) \\ |U|=k}} \left[\left(\prod_{i \in U} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \right) \times \left(\prod_{j \in \mathcal{V}(A) \setminus U} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \right) \right] \quad (4)$$

We further write $\mathcal{H}_{\mathbf{z}}$ to represent the sum of the entries with a single error term:

$$\mathcal{H}_{\mathbf{z}} \coloneqq \sum_{i \in \mathcal{V}(A)} \Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right) \prod_{\substack{j \in \mathcal{V}(A) \\ j \neq i}} \mathcal{P}\left(\mathbf{z}_j \mid \mathbf{z}_{\mathrm{Pa}(j)}\right) \quad (5)$$

Hence,

$$\widehat{\mu}(A) - \mu(A) = \sum_{\mathbf{z} \in Z(A)} \left(\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}}\right).$$

We will establish upper bounds on the sums of $\mathcal{L}_{\mathbf{z}}$s and $\mathcal{H}_{\mathbf{z}}$s and in Lemma 3 and Lemma 4, respectively. These lemmas show that the sum of the $\mathcal{H}$ terms dominates the sum of the $\mathcal{L}$ terms. Furthermore, these bounds imply that the estimated reward $\widehat{\mu}(A)$ is sufficiently close to the true expected reward $\mu(A)$ for each $A \in \mathcal{A}$. In the interest of space, the proofs of the following three lemmas are deferred to Appendix A in the Supplementary Material.

Lemma 2. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 1, write $\mathcal{E}$ to denote the event that $|\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)| \leq \sqrt{\frac{|\mathcal{I}|(d+\log(NT))}{T}}$, for all vertices $i \in \mathcal{V}$ and all assignments $\mathbf{z}_{\mathrm{Pa}(i)} \in \{0,1\}^{|\mathrm{Pa}(i)|}$. Then, $\mathbb{P}\{\mathcal{E}\} \geq \left(1 - \frac{2}{T}\right)$.*

Lemma 3. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 1, the following event holds with probability at least $\left(1 - \frac{2}{T}\right)$:*

$$\sum_{\mathbf{z} \in Z(A)} |\mathcal{L}_{\mathbf{z}}| \leq 4(N\eta)^2 \qquad \textit{for all } A \in \mathcal{A}.$$

Here, parameter $\eta = \sqrt{\frac{|\mathcal{I}|(d+\log(NT))}{T}}$ and T is moderately large.

Lemma 4. *For estimates obtained via a covering intervention set $\mathcal{I}$, as in Algorithm 1, the following event holds with probability at least $\left(1 - \frac{2}{T}\right)$:*

$$\left| \sum_{\mathbf{z} \in Z(A)} \mathcal{H}_{\mathbf{z}} \right| \leq \sqrt{\frac{N|\mathcal{I}|\log(|\mathcal{A}|T)}{T}} \qquad \textit{for all } A \in \mathcal{A}.$$

Recall that the random variables $\mathcal{L}_{\mathbf{z}}$ and $\mathcal{H}_{\mathbf{z}}$ depend on the error terms $\Delta\mathcal{P}\left(\mathbf{z}_i \mid \mathbf{z}_{\mathrm{Pa}(i)}\right)$. Moreover, in Lemma 3 and 4, the considered sums can range over exponentially many such variables. The technically involved contribution of these lemmas is that we obtain small error bounds even in such settings of exponentially large sums.

Proof of Theorem 1. Lemma 1 implies that, with probability at least $\left(1 - \frac{1}{T}\right)$, the set $\mathcal{I}$ obtained in Line 1 of Algorithm 1 is indeed a covering intervention set. We combine this guarantee with Lemmas 3 and 4. In particular, with probability at least $\left(1 - \frac{5}{T}\right)$, we have, for all $A \in \mathcal{A}$:

$$\begin{aligned} &|\mu(A) - \widehat{\mu}(A)| \\ &= \left| \sum_{\mathbf{z} \in Z(A)} \left(\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}}\right) \right| \\ &\leq \sqrt{\frac{N|\mathcal{I}|\log(|\mathcal{A}|T)}{T}} + \frac{4N^2|\mathcal{I}|(d+\log(NT))}{T} \\ &\leq 2\sqrt{\frac{N|\mathcal{I}|\log(|\mathcal{A}|T)}{T}} \qquad (\text{for } T \gtrsim N^3) \end{aligned}$$

Let $A_T \in \mathcal{A}$ be the intervention returned by Algorithm 1 (after T rounds of interventions), i.e., $A_T = \arg\max_{A \in \mathcal{A}} \widehat{\mu}(A)$. In addition, $A^* = \arg\max_{A \in \mathcal{A}} \mu(A)$ be the optimal intervention. Hence, with probability at least $\left(1 - \frac{5}{T}\right)$, we have

$$\mu(A^*) - \mu(A_T) \leq 4\sqrt{\frac{N|\mathcal{I}|\log(|\mathcal{A}|T)}{T}} \quad (6)$$

This guarantee gives us the desired upper bound on the simple regret, R_T, of Algorithm 1:

$$\begin{aligned} R_T &= \mathbb{E}\left[\mu(A^*) - \mu(A_T)\right] \\ &\leq \left(4\sqrt{\frac{N|\mathcal{I}|\log(|\mathcal{A}|T)}{T}}\right)\left(1 - \frac{5}{T}\right) + \frac{5}{T} \\ &\leq 5\sqrt{\frac{N|\mathcal{I}|\log(|\mathcal{A}|T)}{T}}. \end{aligned}$$

Since the size of the covering intervention set satisfies $|\mathcal{I}| = 3d \cdot 2^d(\log N + 2d + \log T)$ (see Lemma 1), we also have the following explicit form of the simple regret bound

$$R_T = O\left(\sqrt{\frac{N\, d2^d\, \log|\mathcal{A}|}{T}} \log T\right).$$

The theorem stands proved. □

4 ALGORITHM FOR GRAPHS WITH UNOBSERVED VARIABLES

We now extend our algorithm to causal graphs with unobserved variables. In particular, we study Semi Markovian

Bayesian Networks (SMBNs) where we have the causal graph defined as $\mathcal{G} = (\mathcal{V}, E, E')$. Here, E is the set of directed edges, and E' is the set of bi-directed edges denoting the presence of an unobserved common parent. Any general causal graph can be projected to an equivalent SMBN [Tian and Pearl, 2002]. Hence, without loss of generality and throughout this section, we assume that the causal graph is an SMBN. It is relevant to note that in an SMBN all the vertices in $\mathcal{V}$ are observable and the unobserved variables are encapsulated by the edges E'.

Assume that the vertices $\mathcal{V}$ are topologically ordered (based on the directed edges E) and the ordering is preserved in any subset $\mathcal{U} \subset \mathcal{V}$. The SMBN graph $\mathcal{G}$ can be decomposed into a disjoint set of vertices known as *confounded components* (c-components), where each c-component is the maximal set of vertices that are connected through a bi-directed edge in E'. Let $\mathcal{C}(A)$ denote all the c-components of $\mathcal{G}$ under intervention A. We use C_i to denote the i^{th} c-component in $\mathcal{C}(A)$. We assume that any C_i maintains the topological order (induced by the directed edges E). Now, the joint distribution of the vertices for an assignment $\mathbf{z} \in Z(A)$, under intervention A, can be written as

$$\mathbb{P}\left[V = \mathbf{z} \mid \mathrm{do}(A)\right] = \prod_{C_i \in \mathcal{C}(A)} \mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}\left(\mathbf{z}_{C_i}\right).$$

Under an empirical estimation, we represent the s^{th} sample from the distribution $\mathcal{P}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}\left(\mathbf{z}_{C_i}\right)$ via the indicator random variable $Y_s(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)})$, which takes the value one when $\mathcal{V}_{C_i} = \mathbf{z}_{C_i}$, else it takes the value zero. Let $n\left(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)$ be the total number of samples in this for the pair $(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)})$. We compute the probability estimates as follows

$$\widehat{\mathcal{P}}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}\left(\mathbf{z}_{C_i}\right) = \frac{\sum_{s=1}^{T_i} Y_s\left(\mathbf{z}_{C_i}, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)}{n\left(C_i, \mathbf{z}_{\mathrm{Pa}(C_i)}\right)} \tag{7}$$

$$\widehat{\mu}(A) = \sum_{\mathbf{z} \in Z(A)} \prod_{C_i \in \mathcal{C}(A)} \widehat{\mathcal{P}}_{\mathbf{z}_{\mathrm{Pa}(C_i)}}\left(\mathbf{z}_{C_i}\right) \tag{8}$$

Next, we extend the definition of covering intervention set (Definition 1) for SMBNs:

Definition 2. *A set of intervention $\mathcal{I}$ is a covering intervention set if for all subsets S of every c-component in $\mathcal{G}$, and every assignment $\mathbf{z}_{\mathrm{Pa}(S)} \in \{0,1\}^{|\mathrm{Pa}(S)|}$ there exists and $I \in \mathcal{I}$ with the properties that*

- *No vertex in S is intervened in I.*
- *Every vertex in $\mathrm{Pa}(S)$ is intervened in I.*
- *$\mathrm{Pa}(S)$ is intervened with assignment $\mathbf{z}_{\mathrm{Pa}(S)}$.*

We construct a covering intervention set as before using the randomized method in Acharya et al. [2018]. The next lemma states that the randomized method provides a covering intervention set of size $\widetilde{O}(\log N)$ even in the case of SMBNs. This result is a direct implication of Lemma 4.2 in Acharya et al. [2018].

Lemma 5 ([Acharya et al., 2018])**.** *For any moderately large $T \in \mathbb{Z}_+$ and any causal graph $\mathcal{G}$—with in-degree at most d and c-components of size at most ℓ—there exists a covering intervention set $\mathcal{I}$ of size $k = (3d)^{\ell}\, 2^{\ell d}(\log N + 2\ell d + \log T)$. Furthermore, such a set $\mathcal{I}$ can be found with probability at least $\left(1 - \frac{1}{T}\right)$.*

The simple regret algorithm for SMBNs is exactly the same as Algorithm 1, except for the following two changes:

- The ConstructCover subroutine returns a covering intervention set of size $(3d)^{\ell} 2^{\ell d}(\log N + 2\ell d + \log T)$.
- We use equation (8) to compute the estimates $\widehat{\mu}(A)$ for each $A \in \mathcal{A}$.

The theorem below is the main result of this section.

Theorem 2. *Let $\mathcal{G}$ be any given causal graph over N vertices and with c-components of size at most ℓ. Also, let the in-degree of the vertices in $\mathcal{G}$ be at most d. Then, for any (moderately large) time horizon T and given any covering intervention set $\mathcal{I}$ of $\mathcal{G}$, Algorithm 1 achieves simple regret*

$$\mathrm{R}_T = O\left(\sqrt{\frac{N\, 2^d\, 4^{\ell}\, |\mathcal{I}| \log\left(|\mathcal{A}|T\right)}{T}}\right).$$

Hence, using Lemma 5, we obtain the following bound on the simple regret

$$\mathrm{R}_T = O\left(\sqrt{\frac{N\, (3d\, 8^d)^{\ell}\, \log|\mathcal{A}|}{T}}\log T\right).$$

A complete proof of Theorem 2 appears in Appendix B; below, we provide a sketch.

Proof Sketch of Theorem 2. We first introduce the notion of *pseudo parents* $\mathrm{Pa}'(j)$ of each vertex j in an SMBN graph $\mathcal{G}$ (see Appendix B). This notion crucially aids the regret analysis. We show that one can essentially view the factorization of an SMBN (under an intervention A) as a factorization over a fully observable graph where each vertex V_j has the set $\mathrm{Pa}'(j)$ as its parents (see Lemma 6 in Appendix B).

Now, to bound the simple regret in the SMBN context, we express, for interventions $A \in \mathcal{A}$, the difference between the estimated means, $\widehat{\mu}(A)$, and the true means, $\mu(A)$, as follows: $\widehat{\mu}(A) - \mu(A) = \sum_{\mathbf{z} \in Z(A)} (\mathcal{H}_{\mathbf{z}} + \mathcal{L}_{\mathbf{z}})$. Here, $\mathcal{H}_{\mathbf{z}}$ and $\mathcal{L}_{\mathbf{z}}$ denote the first order and higher order terms, respectively, with respect to estimate errors $\Delta\mathcal{P}\left(\cdot \mid \cdot\right)$.

Note that, in the SMBN context, the error terms $\Delta\mathcal{P}\left(\cdot \mid \cdot\right)$, as such, are obtained for individual c-components (see equation (7)). However, we are able to obtain tractable forms for

the quantities $\mathcal{H}_{\mathbf{z}}$ and $\mathcal{L}_{\mathbf{z}}$ via the above-mentioned factorization, which considers each vertex V_j in conjunction with its pseudo parents $\mathrm{Pa}'(j)$.

Building up on these involved expressions, we establish upper bounds on the sums of $\mathcal{L}_{\mathbf{z}}$s and $\mathcal{H}_{\mathbf{z}}$s. These upper bounds imply that, with high probability, the estimated mean $\widehat{\mu}(A)$ is close to the true mean $\mu(A)$, for each intervention $A \in \mathcal{A}$. Hence, our SMBN algorithm (which selects intervention $\arg\max_{A\in\mathcal{A}} \widehat{\mu}(A)$) achieves low simple regret as stated in Theorem 2.

Remark. As mentioned previously, this extension to SMBNs enables us to address settings wherein one is allowed to intervene only on a subset of the vertices. In such a case, we can reduce the graph to an SMBN by treating the vertices that can be intervened upon as observable vertices and the rest of the vertices as unobservable.

5 EXPERIMENTS

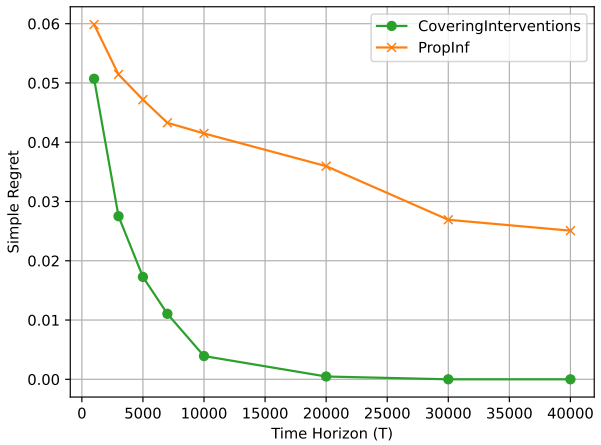

Figure 1: Plot of simple regret with rounds of exploration.

This section provides empirical evaluations of our algorithm. In the experiments, we compare our algorithm, CoveringInterventions (Algorithm 1) with PropInf, the propagating inference algorithm of Yabe et al. Yabe et al. [2018]. As in implementation of Yabe et al. [2018] (see Section 5 of the cited paper), we uniformly sample and do not explicitly solve their proposed optimization problem. The source code of our implementations is available at https://github.com/sawarniayush/learning-good-interventions-using-covering.

For the experiments, we consider a causal graph $\mathcal{G} = (\mathcal{V}, E)$ (over Bernoulli random variables) with number of nodes (variables) $N = |\mathcal{V}| = 17$ and in-degree $d = 4$. The vertex set $\mathcal{V}$ is partitioned into four subsets with cardinalities $|\mathcal{V}_1| = 7$, $|\mathcal{V}_2| = 5$, $|\mathcal{V}_3| = 4$, and $|\mathcal{V}_4| = 1$, respectively. Here, the singleton $\mathcal{V}_4$ consists of the reward variable, which is connected to all the 4 nodes in $\mathcal{V}_3$. Furthermore, the graph $\mathcal{G}$ is layered in the sense that, for each index $\ell \in \{2, 3, 4\}$ and each node $V_i \in \mathcal{V}_\ell$, the parents $\mathrm{Pa}(i) \subset \mathcal{V}_{\ell-1}$. Also, $\mathcal{V}_1$ is the set of leaf vertices – the vertices in $\mathcal{V}_1$ do not have any incoming edges.

For each non-reward variable, V_i, we set the condition probability $\mathbb{P}\{V_i = 1 \mid \mathrm{Pa}(i) = \mathbf{1}\} = 0.8$. That is, when all the parents of V_i are equal to 1,[5] then $V_i = 1$, with probability 0.8. For any other realization of the parents, the conditional probability of $V_i = 1$ is set to be 0.4, i.e., $\mathbb{P}\{V_i = 1 \mid \mathrm{Pa}(i) \neq \mathbf{1}\} = 0.4$. For the reward node V_{17} we have $\mathbb{P}\{V_{17} = 1 \mid \mathrm{Pa}(17) = \mathbf{1}\} = 0.9$ and $\mathbb{P}\{V_{17} = 1 \mid \mathrm{Pa}(17) \neq \mathbf{1}\} = 0.4$.

The set of interventions $\mathcal{A}$ is composed of all possible interventions on the leaf nodes, $\mathcal{A} = \{\mathrm{do}(\mathcal{V}_1 = s) \mid s \in \{0, 1\}^7\}$; recall that $|\mathcal{V}_1| = 7$. Note that setting each leaf node to 1 yields the optimal intervention $A^* = \mathrm{do}(\mathcal{V}_1 = \mathbf{1})$.

Simple Regret vs. Time: In our experiments, for the two algorithms, we compare the simple regret with time horizon T. In particular, for each relevant T, we execute the two algorithms 140 times and average the simple regret across these runs. We plot our results in Figure 1 and show that CoveringInterventions converges to low regret faster than PropInf.

Runtime: For this experimental setup, CoveringInterventions ran at least 8 times faster than PropInf across all the executions.[6] This runtime gap between the two implementations, highlights that CoveringInterventions scales better with the number of variables N.

6 CONCLUSION AND FUTURE WORK

Using the idea of covering interventions, this paper obtains improved simple regret guarantees for the causal bandit problem. We also generalize the guarantee to causal graphs with unobserved variables. Notably, and in contrast to prior works, our regret guarantees only depend on the explicit problem parameters. Our experiments empirically highlight that our algorithm provides improvements over baselines. Establishing lower bounds in the general causal bandit setup is an important direction of future work. It is also interesting to develop computationally efficient (simple regret) algorithms for settings in which the target set $\mathcal{A}$ is large and implicitly specified.

[5] Recall that intervening on all parent nodes of a vertex is the same as conditioning on them.

[6] The computation of the β parameters is a time consuming step in PropInf.

Acknowledgements

Siddharth Barman's reserach is supported by a SERB Core research grant (CRG/2021/006165). Ayush Sawarni gratefully acknowledges a travel grant by Microsoft Research, India.

References

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. *Advances in Neural Information Processing Systems*, 31, 2018.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Pascal Caillet, Sarah Klemm, Michel Ducher, Alexandre Aussem, and Anne-Marie Schott. Hip fracture in the elderly: a re-analysis of the epidos study with causal bayesian networks. *PLoS One*, 10(3):e0120125, 2015.

Daniel Koch, Robert S Eisinger, and Alexander Gebharter. A causal bayesian network model of disease progression mechanisms in chronic myeloid leukemia. *Journal of theoretical biology*, 433:94–105, 2017.

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

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Sujee Lee, Sijie Wang, Philip A Bain, Christine Baker, Tammy Kundinger, Craig Sommers, and Jingshan Li. Reducing copd readmissions: A causal bayesian network model. *IEEE Robotics and Automation Letters*, 3(4): 4046–4053, 2018.

Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. *Advances in Neural Information Processing Systems*, 34:24817–24828, 2021.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Efficient reinforcement learning with prior causal knowledge. In *Conference on Causal Learning and Reasoning*, pages 526–541. PMLR, 2022.

Aurghya Maiti, Vineet Nair, and Gaurav Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *Uncertainty in Artificial Intelligence*, pages 1328–1338. PMLR, 2022.

Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR, 2021.

Judea Pearl. Causality: Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

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

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

Rajat Sen, Karthikeyan Shanmugam, Murat Kocaoglu, Alex Dimakis, and Sanjay Shakkottai. Contextual bandits with latent confounders: An nmf approach. In *Artificial Intelligence and Statistics*, pages 518–527. PMLR, 2017b.

Jaime Sevilla. Explaining data using causal bayesian networks. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, pages 34–38, 2020.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12 (1-2):1–286, 2019.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 519–527, 2002.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *arXiv preprint arXiv:2208.12764*, 2022.

Nuoya Xiong and Wei Chen. Combinatorial pure exploration of causal bandits. In *International Conference on Learning Representations*, 2023.

Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. In *International Conference on Machine Learning*, pages 5512–5520. PMLR, 2018.

Takami Yoshida and Kazuhiro Nakadai. Active audio-visual integration for voice activity detection based on a causal bayesian network. In *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pages 370–375. IEEE, 2012.