Proceedings of Machine Learning Research vol 236:423–461, 2024 3rd Conference on Causal Learning and Reasoning

Confounded Budgeted Causal Bandits

Fateme Jamshidi FATEME.JAMSHIDI@EPFL.CH
EPFL, Switzerland

Jalal Etesami J.ETESAMI@TUM.DE
TUM, Germany

Negar Kiyavash NEGAR.KIYAVASH@EPFL.CH
EPFL, Switzerland

Editors: Francesco Locatello and Vanessa Didelez

Abstract

We study the problem of learning "good" interventions in a stochastic environment modeled by its underlying causal graph. Good interventions refer to interventions that maximize rewards. Specifically, we consider the setting of a pre-specified budget constraint, where interventions can have non-uniform costs. We show that this problem can be formulated as maximizing the expected reward for a stochastic multi-armed bandit with side information. We propose an algorithm to minimize the *cumulative regret* in general causal graphs. This algorithm trades off observations and interventions based on their costs to achieve the optimal reward. This algorithm generalizes the state-of-the-art methods by allowing non-uniform costs and hidden confounders in the causal graph. Furthermore, we develop an algorithm to minimize the *simple regret* in the budgeted setting with non-uniform costs and also general causal graphs. We provide theoretical guarantees, including both upper and lower bounds, as well as empirical evaluations of our algorithms. Our empirical results showcase that our algorithms outperform the state of the art.

Keywords: Causal inference, Multi-armed bandits.

1. Introduction

Multi-armed bandits (MAB) problem has been widely studied in sequential decision-making literature (Lai et al., 1985; Even-Dar et al., 2006). In this problem, a learner sequentially selects an arm to pull and receives a stochastic reward. The learner tries different arms with the goal of maximizing the expected reward. A commonly used assumption in the literature is that the arms are statistically independent. In other words, the distribution of one arm's reward contains no information about the reward of the other arms. Under this assumption, a variety of approaches have been developed in the literature to solve the MAB problem, such as Thompson sampling (Thompson, 1933) and variants of Upper Confidence Bound (UCB) (Auer et al., 2002; Cappé et al., 2013). Recently, a variant of the problem where dependencies among different arms are allowed has been studied. In such a setting, prevalent in real-world problems, pulling an arm reveals additional information about other arms. Examples of applications can be found in various settings, such as linear optimization (Dani et al., 2008), combinatorial bandits (Cesa-Bianchi and Lugosi, 2012), and Lipschitz bandits (Magureanu et al., 2014).

An effective and succinct representation of interdependencies among a set of variables (e.g., arms) can be captured by its corresponding causal graph (Pearl, 1995). In the field of causal discovery, a significant array of algorithms has been devised with the goal of identifying the underlying causal graph (Spirtes et al., 2000; Margaritis and Thrun, 1999; Chickering, 2002; Mokhtarian

et al., 2022). Moreover, the study of sample complexity, crucial for understanding the efficiency of these algorithms, has received considerable attention (Kalisch and Bühlman, 2007; Jamshidi et al., 2023b; Acharya et al., 2023). Such graphs have been successfully used in a wide range of applications from agriculture (Splawa-Neyman et al., 1990) and genetics (Meinshausen et al., 2016) to marketing (Kim et al., 2008) to model the causal relationships. In this work, we study the MAB problem in a stochastic environment in which the dependencies among different arms are modeled by the underlying causal graph. We assume that this causal graph is available to the learner as side information [1]. This formulation is known as causal MAB, and it has recently gained increasing attention in literature (Bareinboim et al., 2015; Lattimore et al., 2016; Lee and Bareinboim, 2018, 2019; Lu et al., 2020; Nair et al., 2021; Maiti et al., 2022).

In certain variants of the MAB problem, pulling an arm is associated with a cost (Kocaoglu et al., 2017; Lindgren et al., 2018; Nair et al., 2021). In this setting, the challenge of the learner with a limited budget is to use the budget for exploring different arms effectively in order to maximize the reward. As an example, consider a treatment-effect problem in which the goal of a practitioner is to measure the effectiveness of different treatments and ultimately find the most effective one. In this example, the effectiveness of the treatments (e.g., the percentage of recovered patients) denotes the reward. On the other hand, different treatments may have different costs. Suppose that there are two treatments available: A) a medicament and B) a surgery. As pulling arm B is more expensive than arm A in this problem, the practitioner's challenge is to use her given budget effectively to try both treatments and maximize the reward.

We study causal MAbs with *non-uniform* costs for pulling arms. As we discuss in Sections 3 and 4, having non-uniform costs leads to different learning algorithms and theoretical guarantees. Additionally, we relax the existing structural assumptions on the underlying causal graph, as such structural assumptions may not be valid in many real-world problems. These assumptions were put into place to simplify accounting for the information pulling an arm reveals about other arms. For instance, a standard result in causal inference literature implies that when the causal graph does not have any unblocked backdoor path (see Appendix A for definitions) between the intervened variables and the reward variable, the effect of any intervention (pulling any arm) is equal to the conditional expectation of the reward given that arm (Pearl, 2009). Lastly, previous work has mainly considered the case where the causal graph is fully observable. We relax this assumption by allowing for so-called unobserved confounders, i.e., variables we cannot observe.

Contribution: Our main contributions are as follows.

- We generalize the setting studied in the state of the art in causal bandit literature by allowing non-uniform costs and hidden confounders in the causal graph. Non-uniform costs introduce additional complexity to the MAB problem in terms of the trade-off between exploration and exploitation. It becomes crucial for the learner to select an arm for exploration not only based on its reward but also its associated cost. To address this complexity, we propose algorithms that incorporate cost-dependent exploration criteria both in the setting of simple and cumulative regret. General causal graphs with hidden confounders add yet another challenge: how to avoid spurious correlations in the data as a result of the confounders and harness true causal relationships to learn about other arms besides the one being played. To overcome this challenge, we propose

1. It is pertinent to note that using side information is observed in other domains, such as causal effect identification (Tikka et al., 2019; Akbari et al., 2023) and causal imitation learning (Jamshidi et al., 2023a).

estimators in Section 3.1 for the expected reward of arms that leverage both observational and interventional data.

- We propose two algorithms (Algorithm 1 in Section 3 and Algorithm 2 in Section 4) to minimize the cumulative and simple regrets, respectively[2] and upper bound their expected regrets. We prove that by leveraging causal information, Algorithm 1 achieves better cumulative regret than the optimal classic MAB algorithm. In Algorithm 2, we propose a new threshold that accounts for the cost of pulling arms to identify infrequent arms more effectively. As a result, Algorithm 2 outperforms prior work (Nair et al., 2021) even in their own settings (when the costs are uniform and the causal graph has no backdoor). Moreover, we present lower bounds on both simple and cumulative regrets of any algorithm and discuss their relations with the presented upper bounds in Section 5.

- We evaluate our proposed algorithms in Section 6. Our simulation results show that our algorithms perform well for general causal graphs and non-uniform costs and outperform the state of the art even in the settings they were specifically designed for.

1.1. Related Work

Authors in Tran-Thanh et al. (2012) propose F-KUBE, an algorithm for a budgeted MAB problem without utilizing the underlying causal graph. Lattimore et al. (2016) study the problem of minimizing the simple regret in a special causal graph called parallel graphs[3] after T steps where the cost of pulling all arms is one. They propose an algorithm with average regret of $\mathcal{O}\big(\sqrt{\frac{a}{T}\log\frac{NT}{a}}\big)$, where a defined in Remark 9 depends on the underlying causal model and N is the number of intervenable variables.

The authors in Nair et al. (2021) study a causal MAB problem in which the learner has a limited budget B, all interventions have the same cost $c \geq 1$, and the cost of observation is one. They consider the problem of minimizing the simple regret in special causal graphs called no-backdoor graphs[4]. They show that their proposed algorithm's expected regret is upper bounded by $\mathcal{O}\big(\sqrt{\frac{ca}{B}\log\frac{NB}{ca}}\big)$. We also study this particular setting in Section 4 as a special case of our setting but allow for non-uniform costs and derive a tighter bound for the expected regret (Remark 9).

Nair et al. (2021) studies non-budgeted a causal MAB problem with general causal graphs when the objective is the cumulative regret. The proposed algorithm in Nair et al. (2021) requires access to the distribution of parents of the reward variable for each intervention. This restrictive assumption is also required in Lu et al. (2020). Maiti et al. (2022) studies a causal MAB problem when all costs are assumed to be one for both simple and cumulative regret objectives. In the case of simple regret, the proposed algorithm for causal graphs with possibly hidden confounders attains an expected simple regret upper bounded by $\mathcal{O}\big(\sqrt{\frac{b}{T}\log\frac{NT}{b}}\big)$, where b depends on the causal model. In the case of cumulative regret, the proposed algorithm only works for causal graphs with no hidden variables.

2. Our theoretical results both generalize the results in Nair et al. (2021) and Maiti et al. (2022) (to allow for non-uniform costs and general causal graphs) and correct the oversights and errors in the proofs of these papers which affect the validity of the bounds claimed therein (see Section H for details).
3. It is composed of variable set $\mathbf{V} = \{X_1, \ldots, X_N, Y\}$ and edges from each X_i to Y.
4. The graphs in which all backdoor paths from each intervenable variable to the reward variable are blocked. Please refer to Appendix A for details.

We generalize both aforementioned results to non-uniform cost settings and derive tighter theoretical bounds on the regret.

2. Preliminaries

Throughout this paper, random variables and their realizations are denoted by capital and lowercase letters, respectively. We use bold capital and lowercase letters to denote sets of variables and their realizations, respectively.

Causal structure: Let $\mathcal{G} = (\mathbf{V}, \mathbf{E}^d, \mathbf{E}^b)$ denote an acyclic-directed mixed graph (ADMG) with the set of observed variables $\mathbf{V}$, the set of *directed* edges $\mathbf{E}^d \subseteq \mathbf{V} \times \mathbf{V}$ and the set of *bidirected* edges $\mathbf{E}^b \subseteq \binom{\mathbf{V}}{2}$. The existence of a bidirected edge between nodes V_1 and V_2 represents a hidden confounder that influences both V_1 and V_2.

Given two arbitrary variables $V_1, V_2 \in \mathbf{V}$, when $(V_1, V_2) \in \mathbf{E}^d$, V_1 is a parent of V_2 and V_2 is a child of V_1. The set of parents of V_2 is denoted by $\mathbf{Pa}(V_2)$.

Given two subsets of variables $\mathbf{R}$ and $\mathbf{S}$ and their realizations $\mathbf{r}$ and $\mathbf{s}$, respectively, let $P_{\mathbf{s}}(\mathbf{r}) := P(\mathbf{R} = \mathbf{r}|do(\mathbf{S} = \mathbf{s}))$ denote the post-interventional distribution of $\mathbf{R}$ after intervening on $\mathbf{S}$.

Definition 1 (C-component (Tian and Pearl, 2002b)) *Two observed variables V_1 and V_2 are said to be in a c-component of an ADMG $\mathcal{G}$, if and only if they are connected by a bi-directed path.*

As an example, in Figure 1, $\{X_1, X_2, X_3, X_5\}$ and $\{X_4\}$ are two c-components of $\mathcal{G}$.

Definition 2 (Identifiability (Tian and Pearl, 2002b)) *Given an ADMG $\mathcal{G} = (\mathbf{V}, \mathbf{E}^d, \mathbf{E}^b)$, and two disjoint subsets $\mathbf{R}, \mathbf{S} \subseteq \mathbf{V}$, $P_{\mathbf{s}}(\mathbf{r})$ is said to be identifiable in $\mathcal{G}$ if $P_{\mathbf{s}}(\mathbf{r})$ is uniquely computable from $P(\mathbf{V})$.*

Causal multi-arm bandits: Let $\mathbf{X} = \{X_1, \ldots, X_N\} \subseteq \mathbf{V}$ and $Y \in \mathbf{V}$ denote the set of *intervenable* variables (variables that the learner is allowed to intervene on) and the reward variable, respectively. For ease of presentation, we assume that all variables are binary. All our results can be extended to sets of finite-domain variables.

In the causal MAB setting, at each round, a learner can explore either by intervening in the system or merely observing it. If the learner decides to intervene, they will select an intervenable variable, e.g., $X_i \in \mathbf{X}$, set its value, e.g., $do(X_i = x)$, and observe the remaining variables, i.e., $\mathbf{V} \setminus \{X_i\}$. This choice of action (arm) is denoted by $a_{i,x}$. On the other hand, when the decision is to observe, i.e., $do()$, she merely observes all observed variables. This action is denoted by a_0.

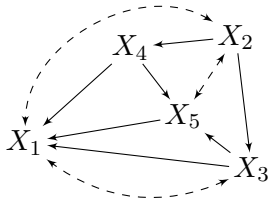

Figure 1: An ADMAG $\mathcal{G}$ over $\mathbf{V} = \{X_1, ..., X_5\}$. Bidirected edges are represented by dashed edges.

We denote the set of possible actions by $\mathcal{A} := \{a_{i,x} | i \in [N], x \in \{0,1\}\} \cup \{a_0\}$. We assume that the cost of pulling an arm $a \in \mathcal{A}$ is $c_a \in \mathbb{R}_+$ and denote the set of costs by $\mathcal{C} := \{c_a | a \in \mathcal{A}\}$. Moreover, without loss of generality, we assume that the cost of a_0 is one, i.e., $c_0 = 1$. At time t, the arm pulled by the learner is denoted by a^t, the reward received is denoted by y^t, and the observed values of the variables in an arbitrary subset $\mathbf{S} \subseteq \mathbf{V}$ are denoted by $\mathbf{s}^t$.

Problem Setting: We study a causal MAB problem, in which a learner with a budget $B \geq 0$ aims to minimize either its *simple regret* (Section 4) or *cumulative regret* (Section 3). It is assumed that the learner knows the underlying causal graph. This problem is known as the budgeted causal MAB (Nair et al., 2021). Regret is a commonly used measure to evaluate the performance of learners in a bandit setting, and it captures the foregone utility from the actual action choice against the optimum action (Cesa-Bianchi and Lugosi, 2006).

In order to formally introduce the regret, we first define the average reward of action $a \in \mathcal{A}$ as follows: $\mu_a := \mathbb{E}[Y|a]$. For example, $\mu_{a_{i,x}}$ denotes $\mathbb{E}[Y|do(X_i = x)]$.

Simple regret: Let a^* denote the arm that maximizes the expected reward with budget B. The simple regret of a learner using budget B is defined by

$$R_s(B) := \mu_{a^*} - \mu_{\tilde{a}_B}, \tag{1}$$

where $\tilde{a}_B$ denotes the arm selected by the learner after exhausting budget B. When the learner's objective is the minimize simple regret, it suffices to find the best arm at the final step (i.e., after spending their budget) without having to worry about the intermediate actions that they chose.

In many real-world applications, it is important that the learner does not pull sub-optimal arms too often during her exploration. In this case, the objective function should reflect the intermediate regrets the learner accumulates.

Cumulative regret: Let T_B^ℓ denote the time step that a learner ℓ consumes its budget B, i.e., at time step $T_B^\ell + 1$, it does not have enough budget to perform even the lowest cost action. In this case, the expected reward accumulated by the learner ℓ will be $\mathcal{R}^\ell(B) := \sum_{t \leq T_B^\ell} \mu_{a^t}$, where μ_{a^t} is the rewards of action taken at time t. Furthermore, let $\mathcal{R}^*(B)$ denote the expected reward accumulated by the optimum learner with budget B. Then, the cumulative regret of the learner ℓ using budget B is given by

$$R_c(B) := \mathcal{R}^*(B) - \mathcal{R}^\ell(B). \tag{2}$$

As we consider a single learner in this work, in the rest of the paper, we drop the superscript ℓ. A learner minimizing cumulative regret must trade off exploration vs. exploitation.

Remark 3 *We can define a non-budgeted causal MAB problem in which there is no cost associated with pulling an arm, but the learner has limited time T to either identify the best arm or minimize cumulative regret during T steps. This problem is a special case of the budgeted MAB problem. Assume all arms have the same cost $c_a = c > 0$, then a budgeted causal MAB with budget B is equivalent to a non-budgeted causal bandit with the time limit $T = B/c$ and the simple and cumulative regrets are given by $R_s(B/c)$ and $R_c(B/c)$, respectively.*

3. Cumulative Regret in General Graphs

In this section, we study the budgeted causal MAB problem in general causal graphs with hidden confounders when the learner's objective is to minimize cumulative regret. We propose Algorithm 1, developed based on Upper Confidence Bound (UCB) algorithm (Auer et al., 2002), which generalizes the state-of-the-art in causal MAB in two ways: it allows for non-uniform costs among the arms and as well as the existence of hidden confounders in the causal graph.

Non-uniform costs change the optimal exploitation policy as, depending on the costs, pulling the arm with the highest reward repeatedly, in general, does not maximize the learner's accumulated reward within the budget. Indeed our empirical studies in Section 6.1 show that Algorithm 1 outperforms existing causal MAB algorithms designed for uniform costs.

Algorithm 1 works in general graphs and relaxes existing structural assumptions on the underlying causal graph in the literature. Recently, Maiti et al. (2022) studied the non-budgeted causal MAB problem with graphs that have no hidden confounders. This is a limiting assumption in many real-world applications such as medical science, epidemiology, and sociology when it is impossible to ensure that all common confounders are measured in a study (Leek and Storey, 2007; Imai et al., 2010; Colombo et al., 2012). Our proposed algorithm merely requires the identifiability assumption for all intervenable variables in $\mathcal{G}$. A sufficient graphical condition for identifiability of $P_{x_i}(y)$ is that there does not exist a path of bi-directed edges from X_i to its children (Tian and Pearl, 2002b) which significantly relaxes the existing structural assumptions in the state of the art.

Algorithm 1 takes as inputs the causal graph $\mathcal{G}$, the budget B, and the cost set $\mathcal{C}$. In the beginning, it pulls each arm once (line 1). Assuming that the intervenable variables are binary, i.e., $X_i \in \{0, 1\}$, this requires $2N + 1$ number of steps and costs $\sum_{i,x} c_{i,x} + 1$.

Algorithm 1 Budgeted Cumulative Regret in General Graphs

Input: $\mathcal{G}$, B, $\mathcal{C}$

1: Pull each arm once and set $t = 2N + 1$;
2: Set $B^t = B - \sum_{i,x} c_{i,x} - 1$ and $\beta = 1$;
3: **while** $B^t \geq 1$ **do**
4: **if** $N_0^{t-1} < \beta^2 \log t$ or $B^t < \min_{i,x} c_{i,x}$ **then**
5: pull $a^t = a_0$
6: **else**
7: pull $a^t = \arg\max_{a \in \mathcal{A}} \bar{\mu}_a^{t-1}$
8: **for** $a \in \mathcal{A}$ **do**
9: Update $N_a^t = N_a^{t-1} + \mathbb{1}\{a^t = a\}$.
10: Update $\hat{\mu}_a^t$ and $\bar{\mu}_a^t$ using Equations (5), (8), and (9).
11: Let $\tilde{a} = \arg\max_{a \in \mathcal{A}} (\hat{\mu}_a^t / c_a)$.
12: **if** $\hat{\mu}_0^t < (\hat{\mu}_{\tilde{a}}^t / c_{\tilde{a}})$ **then**
13: Update $\beta = \min\{\frac{2\sqrt{2}}{(\hat{\mu}_{\tilde{a}}^t / c_{\tilde{a}}) - \hat{\mu}_0^t}, \sqrt{\log t}\}$
14: set $t = t + 1$.
15: Update $B^t = B^{t-1} - c_{a^{t-1}}$.

Let N_a^t denote the number of times that arm $a \in \mathcal{A}$ is pulled at the end of t rounds. We denote the estimated average reward by pulling arm a and its estimated UCB at the end of round t by $\hat{\mu}_a^t$ and $\bar{\mu}_a^t$, respectively. The procedure for these estimators will be discussed in Section 3.1. As long

as the remaining budget at round t, B^t, is larger than one, Algorithm 1 continues to explore and exploit by checking at round t whether arm a_0 is pulled at least $\beta^2 \log t$ times (this threshold might change in line 17). If so, Algorithm 1 pulls an arm with the highest $\bar{\mu}_a^t$ in line 8; otherwise, it pulls arm a_0. Afterward, in lines 10-14, it updates $\hat{\mu}_a^t$ and $\bar{\mu}_a^t$ using the newly acquired observational or interventional data, which is discussed in Section 3.1. In the end, the threshold β and the remaining budget are updated in lines 16-19.

3.1. Estimation and update steps

Herein, we explain how to estimate and update $\hat{\mu}_a^t$ and $\bar{\mu}_a^t$ in Algorithm 1. Recall that $\mu_{i,x} = \mathbb{E}[Y|do(X_i = x)]$. Therefore, to estimate $\mu_{i,x}$, it suffices to estimate $P\big(Y = 1|do(X_i = x)\big)$.

Let $\mathbf{C}_i$ and $\mathbf{W}_i$ denote the c-component containing X_i and $\mathbf{V} \setminus \{X_i\}$, respectively. Given two subsets $\mathbf{S}$ and $\mathbf{R}$ of observed variables such that $\mathbf{S} \subseteq \mathbf{R}$ and a subset of realizations $\mathbf{r}$ for $\mathbf{R}$, we use $(\mathbf{r})_{\mathbf{S}}$ to denote the restriction of $\mathbf{r}$ to the variables in $\mathbf{S}$. Given two subsets of variables $\mathbf{S}_1$ and $\mathbf{S}_2$, and realizations $\mathbf{s}_1$ for $\mathbf{S}_1$ and $\mathbf{s}_2$ for $\mathbf{S}_2$, we denote the assignments to $\mathbf{S}_1 \cup \mathbf{S}_2$ by $\mathbf{s}_1 \circ \mathbf{s}_2$.

Under the identifiability assumption for intervenable variables, Bhattacharyya et al. (2020) shows that $P_x(\mathbf{w}_i) := P(\mathbf{W}_i = \mathbf{w}_i|do(X_i = x))$ can be factorized as follows,

$$P_x(\mathbf{w}_i) = \sum_{x' \in \{0,1\}} \prod_{V_j \in \mathbf{C}_i} P\big((x' \circ \mathbf{w}_i)_{V_j} | (x' \circ \mathbf{w}_i)_{\mathbf{Z}_j}\big) \prod_{V_j \notin \mathbf{C}_i} P\big((\mathbf{w}_i)_{V_j} | (x \circ \mathbf{w}_i)_{\mathbf{Z}_j}\big), \tag{3}$$

where $\mathbf{Z}_j = (\bigcup_{V_k \in \mathbf{C}_j} \mathbf{Pa}(V_k) \cup \mathbf{C}_j) \setminus V_j$ and $\mathbf{C}_j$ is c-component of V_j.

Using (3), the expected reward of pulling $a_{i,x}$ would be

$$\begin{aligned}\mathbb{E}[Y|do(X_i = x)] &= \sum_{\mathbf{w}_i' : y = 1} P\big(\mathbf{W}_i = \mathbf{w}_i'|do(X_i = x)\big) \\ &= \sum_{\mathbf{w}_i' : y = 1} \sum_{x' \in \{0,1\}} \prod_{V_j \in \mathbf{C}_i} P\big((x' \circ \mathbf{w}_i')_{V_j} | (x' \circ \mathbf{w}_i')_{\mathbf{Z}_j}\big) \prod_{V_j \notin \mathbf{C}_i} P\big((\mathbf{w}_i')_{V_j} | (x \circ \mathbf{w}_i')_{\mathbf{Z}_j}\big),\end{aligned} \tag{4}$$

where the first summation is over all realization of $\mathbf{W}_i = \mathbf{V} \setminus \{X_i\}$ in which $Y = 1$. This is because the terms with $Y = 0$ have no contribution to the expectation.

Define $\mathbf{O}^t := \{t' \leq t |\ a^{t'} = a_0\}$, and $\mathbf{I}_{i,x}^t := \{t' \leq t |\ a^{t'} = a_{i,x}\}$. $\mathbf{O}^t$ and $\mathbf{I}_{i,x}^t$ denote the set of time steps at which arms a_0 and $a_{i,x}$ are pulled by the end of time t, respectively. Hence, an empirical estimation of average reward of a_0 is given by

$$\hat{\mu}_0^t := \frac{1}{N_0^t} \sum_{t' \in \mathbf{O}^t} \mathbb{1}\{a^{t'} = a_0, y^{t'} = 1\}. \tag{5}$$

To estimate $\mu_{i,x}$ from observational data, it suffices to estimate each term in (4). To do so, we partition $\mathbf{O}^t$ into $|\mathbf{V}|$ number of subsets randomly and denote the j-th partition by $\mathbf{O}_j^t$. We will use the data in $\mathbf{O}_j^t$ to estimate $P(V_j|\mathbf{Z}_j)$. Given a realization x' of X_i and a realization $\mathbf{w}_i'$ of $\mathbf{W}_i$, let $\mathbf{O}_j^t(x', \mathbf{w}_i') := \big\{t' \in \mathbf{O}_j^t | \mathbf{z}_j^{t'} = (x' \circ \mathbf{w}_i')_{\mathbf{Z}_j}\big\}$. Recall that $\mathbf{w}_i'$ is an arbitrary realization of $\mathbf{W}_i$ in which $Y = 1$. To proceed, we require the following definitions,

$$S_{j,i}^t := \min_{\mathbf{w}_i'} \min_{x'} |\mathbf{O}_j^t(x', \mathbf{w}_i')|, \text{ if } V_j \in \mathbf{C}_i, \quad \tilde{S}_{j,i,x}^t := \min_{\mathbf{w}_i'} |\mathbf{O}_j^t(x, \mathbf{w}_i')|, \text{ if } V_j \notin \mathbf{C}_i,$$

where $|\mathbf{O}|$ denotes the size of set $\mathbf{O}$. We also define the minimum number of data points in the partition sets as

$$S_{i,x}^t := \min\big\{\min_{j : V_j \in \mathbf{C}_i} S_{j,i}^t, \min_{j : V_j \notin \mathbf{C}_i} \tilde{S}_{j,i,x}^t\big\}.$$

In the next step, we partition each $\mathbf{O}_j^t(x', \mathbf{w}_i')$ into $S_{i,x}^t$ number of subsets randomly and denote the s-th subset by $\mathbf{O}_j^{t,s}(x', \mathbf{w}_i')$. Let

$$\hat{P}_j^{t,s}(x', \mathbf{w}_i') := \frac{\sum_{t' \in \mathbf{O}_j^{t,s}(x', \mathbf{w}_i')} \mathbb{1}\{v_j^{t'} = (\mathbf{w}_i' \circ x')_{V_j}\}}{|\mathbf{O}_j^{t,s}(x', \mathbf{w}_i')|}, \quad V_j \in \mathbf{C}_i, \tag{6}$$

$$\hat{P}_j^{t,s}(x, \mathbf{w}_i') := \frac{\sum_{t' \in \mathbf{O}_j^{t,s}(x, \mathbf{w}_i')} \mathbb{1}\{v_j^{t'} = (\mathbf{w}_i')_{V_j}\}}{|\mathbf{O}_j^{t,s}(x, \mathbf{w}_i')|}, \quad V_j \notin \mathbf{C}_i. \tag{7}$$

Finally, the expected reward of pulling $a_{i,x}$ is estimated as follows,

$$\hat{\mu}_{i,x}^t := \frac{\sum_{t' \in \mathbf{I}_{i,x}^t} \mathbb{1}\{y^{t'} = 1\} + \sum_{s \in [S_{i,x}^t]} Y_{i,x}^s}{N_{i,x}^t + S_{i,x}^t}, \tag{8}$$

where $[S_{i,x}^t] = \{1, ..., S_{i,x}^t\}$ and $Y_{i,x}^s := \sum_{\mathbf{w}_i' : y=1} \sum_{x' \in \{0,1\}} \prod_{V_j \in \mathbf{C}_i} \hat{P}_j^{t,s}(x', \mathbf{w}_i') \prod_{V_j \notin \mathbf{C}_i} \hat{P}_j^{t,s}(x, \mathbf{w}_i')$.

Lemma 4 *$\hat{\mu}_{i,x}^t$ in (8) and $\hat{\mu}_0^t$ in (5) are unbiased estimators of $\mu_{i,x}$ and μ_0.*

Analogous to UCB algorithm in bandits literature (Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006), Algorithm 1 computes UCB estimate of μ_a at round t using the following equations,

$$\bar{\mu}_{i,x}^t := \hat{\mu}_{i,x}^t + \sqrt{\frac{2 \ln t}{N_{i,x}^t + S_{i,x}^t}}, \quad \bar{\mu}_0^t := \hat{\mu}_0^t + \sqrt{\frac{2 \ln t}{N_0^t}}. \tag{9}$$

Let $a^* := \arg\max_{a \in \mathcal{A}} \frac{\mu_a}{c_a}$ and for $a \in \mathcal{A}$, let $\delta_a := \frac{\mu_{a^*}}{c_{a^*}} - \frac{\mu_a}{c_a}$. Recall that $p_{i,x} = P(X_i = x)$.

Theorem 5 *The expected cumulative regret of Algorithm 1 is bounded by*

$$\delta_0 \Big(\frac{8 \ln B}{\delta_0^2} + 1 + \frac{\pi^2}{3} \Big) + \sum_{\delta_{i,x} > 0} \delta_{i,x} \Big(\frac{8 \ln B}{\delta_{i,x}^2} + 2 - \frac{8 p_{i,x}}{18 \delta_0^2 |\mathbf{V}|} \ln b_{i,x} \cdot \tau_{i,x,b} + \frac{\pi^2}{3} \Big),$$

where $b_{i,x} := \frac{8}{\delta_{i,x}^2} \ln(\frac{B}{\max_a c_a}) + 1$, $\tau_{i,x,b} := \max\big\{0, 1 - |\mathbf{V}| \cdot \mathcal{W}_i \cdot b_{i,x}^{-p_{i,x}^2/(2|\mathbf{V}|)}\big\}$, and $\mathcal{W}_i$ denotes the alphabet size of variables in $\mathbf{V} \setminus \{X_i, Y\}$.

The proof of Theorem 5 is provided in Appendix B. This theorem ensures that the maximum number of pulling a sub-optimal arm a is bounded by a factor of δ_a.

4. Simple Regret in General Graphs

In this section, we study the budgeted causal MAB problem with general graph $\mathcal{G}$ for a learner whose objective is simple regret. The novelty of our results is that they generalize the state-of-the-art by allowing non-uniform costs for arms. As discussed in the previous section, having non-uniform costs may change the trade-off between exploration vs. exploitation and hence requires a different treatment than non-budgeted causal MAB. Our experiments in Section 6 showcase that, indeed, our algorithm outperforms the state of the art, which is designed for uniform costs.

Algorithm 2 Budgeted Simple Regret in General Graphs

Input: $\mathcal{G}, B, \mathcal{C}$

1: **for** $t \in \{1, 2, \ldots, B/2\}$ **do**
2: Pull arm a_0 and observe $\mathbf{v}^t$
3: $\hat{\mu}_0 = 2(\sum_{t=1}^{B/2} y^t)/B$
4: **for** $a_{i,x} \in \mathcal{A}$ **do**
5: Estimate $\hat{\mu}_{i,x}$ using Alg. 3 in Appendix C
6: Estimate $\hat{q}_{i,x}$ using Equation (11)
7: Compute $n(\hat{\mathbf{q}})$ using Equation (10)
8: Construct $\mathcal{A}' := \{a_{i,x} \in \mathcal{A} | \hat{q}_{i,x}^{k_i} \leq \frac{1}{n(\hat{\mathbf{q}})}\}$
9: **if** $|\mathcal{A}'| = 0$ **then**
10: Pull arm a_0 for the remaining $\frac{B}{2}$ rounds
11: Re-estimate $\hat{\mu}_0 = (\sum_{t=1}^{B/2} y^t)/B$
12: **for** $a_{i,x} \in \mathcal{A}$ **do**
13: Re-estimate $\hat{\mu}_{i,x}$ using Alg. 3
14: **else**
15: Compute $n = \frac{B}{2\sum_{i,x} c_{i,x} \mathbb{1}\{a_{i,x} \in \mathcal{A}'\}}$
16: Pull each arm $a_{i,x} \in \mathcal{A}'$ for n rounds
17: **for** $a_{i,x} \in \mathcal{A}'$ **do**
18: $\hat{\mu}_{i,x} = \frac{1}{n} \sum_{t=\frac{B}{2}+1}^{\frac{B}{2}+n|\mathcal{A}'|} y^t \mathbb{1}\{a^t = a_{i,x}\}$
19: **return** $\hat{a}^* \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a$.

Under the identifiability assumption for all intervenable variables in $\mathcal{G}$, we present Algorithm 2 to minimize the simple regret for a budget B. This algorithm generalizes the one in Maiti et al. (2022) to a budgeted causal MAB setting when the arms have non-uniform costs. It uses its given budget B to estimate the average reward of each arm and then outputs an arm with the maximum estimated average reward. More specifically, Algorithm 2 takes the causal graph $\mathcal{G}$, the budget B, and the cost set $\mathcal{C}$ as inputs.

It pulls arm a_0, i.e., collects observational data until it has exhausted half of its budget. This leads to an initial estimate of the expected reward of each arm $a \in \mathcal{A}$ (lines 4-8). Note that estimating the expected rewards is possible due to the identifiability assumption of intervenable variables and is done by Algorithm 3 presented in Appendix C.

Algorithm 3 is proposed by Bhattacharyya et al. (2020) to estimate $\mathbb{E}[Y|do(X)]$ from observational data when the causal effect $P_x(y)$ is identifiable in $\mathcal{G}$.

When an arm is observed frequently during the first part of the algorithm, the initial estimate of its expected reward becomes accurate. Algorithm 2 spends the other half of its budget to explore the so-called infrequent arms (lines 9-23). An arm $a_{i,x} \in \mathcal{A}$ is considered to be infrequent if $\hat{q}_{i,x} \leq \left(\frac{1}{n(\hat{\mathbf{q}})}\right)^{1/k_i}$, where

$$n(\hat{\mathbf{q}}) := \min\Big\{\tau | \sum_{i,x} c_{i,x} \mathbb{1}\{\hat{q}_{i,x} < \Big(\frac{1}{\tau}\Big)^{\frac{1}{k_i}}\} \leq \tau\Big\}, \tag{10}$$

$$\hat{q}_{i,x} := \frac{2}{B} \min_{\mathbf{z}} \Big\{ \sum_{t=1}^{B/2} \mathbb{1}\{x_i^t = x, \widetilde{\mathbf{Pa}}^t(x_i) = \mathbf{z}\}\Big\}, \tag{11}$$

where $\widetilde{\mathbf{Pa}}(X_i) := (\bigcup_{V_j \in \mathbf{C}_i} \mathbf{Pa}(V_j) \cup \mathbf{C}_i) \setminus X_i$, and $\mathbf{C}_i$ denotes the c-component in $\mathcal{G}$ containing X_i. The size of C_i is denoted by k_i.

Let $\mathcal{A}'$ denote the set of all infrequently observed arms. If $\mathcal{A}' = \emptyset$, Algorithm 2 spends the remaining budget for observation, i.e., pulls a_0. Otherwise, it uses the remaining budget to pull the infrequent arms and update their corresponding estimations. Finally, it outputs an arm with the maximum estimated average reward.

Remark 6 *Consider the special case of no-backdoor graphs (causal graphs with no unblocked backdoor paths from intervenable variables to the reward variable Y). This graphical constraint ensures that for all $X \in \mathbf{X}$, $\mathbb{E}[Y|do(X = x)] = \mathbb{E}[Y|X = x]$. This is due to the second rule of do-calculus (Pearl, 1995). For causal MABs with no-backdoor graphs, $\mu_{i,x}$ can be estimated using observation as follows $\sum_{t=1}^{B/2} y^t \mathbb{1}\{x_i^t = x\} / \sum_{t=1}^{B/2} \mathbb{1}\{x_i^t = x\}$. When the interventions have non-uniform costs, redefining $\hat{q}_{i,x} = \frac{2}{B} \sum_{t=1}^{B/2} \mathbb{1}\{x_i^t = x\}$ yields drastically lower regrets. This special case and our improvements are discussed in Appendix E.*

Theorem 7 *The expected simple regret of Algorithm 2 is bounded by $\mathcal{O}\Big(\sqrt{\frac{n(\mathbf{q})}{B} \log \frac{NB}{n(\mathbf{q})}}\Big)$.*

The proof of this theorem is provided in Appendix D.

Remark 8 *Maiti et al. (2022) proposes an algorithm for non-budgeted causal MAB with general causal graphs, which is a special case of our setting in all costs are one. By setting $c_{i,x} = 1$ for all i and x in Theorem 7, we can recover their expected simple regret bound.*

Remark 9 *Nair et al. (2021) studies the causal MAB problem with no-backdoor graphs and an additional constraint on the costs that is $c_{i,x} = c > 1$ for all i and x and $c_0 = 1$. Note that this setting does not satisfy the non-budgeted assumption in Maiti et al. (2022). Moreover, their algorithm uses a different exploration set than $\mathcal{A}'$ that seems to result in both worse performance and theoretical bound. Specifically, the threshold for determining the infrequent arms in Nair et al. (2021) is given by $m'(\mathbf{q}) := \min\{\tau | \sum_{i,x} \mathbb{1}\{p_{i,x} < \frac{1}{\tau}\} \leq \tau\}$. As we show in Appendix E, in this setting, $n(\mathbf{q}) \leq cm'(\mathbf{q})$ for all $c > 1$ and $\mathbf{q}$. Nair et al. (2021) shows that the expected simple regret of their algorithm is bounded by $\mathcal{O}\Big(\sqrt{\frac{cm'(\mathbf{q})}{B} \log \frac{NB}{cm'(\mathbf{q})}}\Big)$. Given that $n(\mathbf{q}) \leq cm'(\mathbf{q})$ for all $c > 1$, even in the special setting of Nair et al. (2021), our algorithm achieves better expected simple regret. This is also shown empirically in our experiment in Section 6.2.*

5. Lower Bounds

Simple regret: As mentioned earlier, Maiti et al. (2022) studies a special setting of causal MAB problem with uniform costs ($c_{i,x} = 1$ for all i, x) in general causal graphs when the objective function of the learner is simple regret. In particular, they showed that there is a large class of causal graphs called tree graphs, such that for any graph $\mathcal{T}$ in this class with N intervenable nodes and a positive integer $M \leq N$, there exists a joint distribution $P(\cdot)$ compatible[5] with graph $\mathcal{T}$, such that $n(\mathbf{q}) = M$ and the expected simple regret of any causal MAB algorithm is $\Omega\big(\sqrt{n(\mathbf{q})/B}\big)$. Comparing this result with the bound introduced in Theorem 7, we observe that for certain categories of graphs with uniform costs, the expected simple regret yielded by Algorithm 2 differs from the minimum value at most by a factor of $\sqrt{\log\left(NB/n(\mathbf{q})\right)}$.

5. Compatibility also known as Markov property (Pearl, 2009) means that the $P(\cdot)$ factorizes according to the graph $\mathcal{T}$.

Cumulative regret: We prove $\min_{A_B} \max_{\mathcal{C},\mathcal{G}_N,P'} R_c(A_B, \mathcal{G}_N, P, \mathcal{C}) \geq \Omega(\sqrt{\lfloor B/c \rfloor KN})$ in appendix F, where $R_c(A_B, \mathcal{G}_N, P, \mathcal{C})$ denotes the cumulative regret of an adaptive algorithm A_B with total budget B on a causal graph $\mathcal{G}_N$ with N nodes each of which has domain $\{1, ..., K\}$. The reward distribution is P and the set of costs is given by $\mathcal{C} = \{1 \leq c_a \leq c | a \in \mathcal{A}\}$. This shows that for any algorithm, there exists a causal bandits problem characterized by $(\mathcal{G}_N, P, \mathcal{C})$ such that it suffers at least $\Omega(\sqrt{\lfloor B/c \rfloor KN})$ of cumulative regret.

6. Experiments

Herein, we present our empirical evaluations of our algorithm in comparison with state of the art. Throughout, each point in figures is obtained as an average of 100 trials[6].

6.1. Cumulative Regret in General Graphs

In this section, we compared the performance of Algorithm 1 with algorithms *CRM* and *F-KUBE* by Maiti et al. (2022) and Tran-Thanh et al. (2012), respectively. *CRM* is a causal MAB algorithm designed for general causal graphs where all of the variables are observable (no hidden confounders exist). *F-KUBE* is a budgeted MAB algorithm with non-uniform costs that does not use the knowledge of the causal graph. We used a graph with 6 intervenable variables, $N = 6$, and modeled each V_i with at least a parent in $\mathcal{G}$ to be the XOR of its parents with probability 0.8 or the XNOR of its parents, otherwise. Moreover, for each variable V_i without any parents, we modeled $V_i \sim Bernoulli(0.5 + 0.5\epsilon)$, where $\epsilon \sim Uniform(0, 1)$.

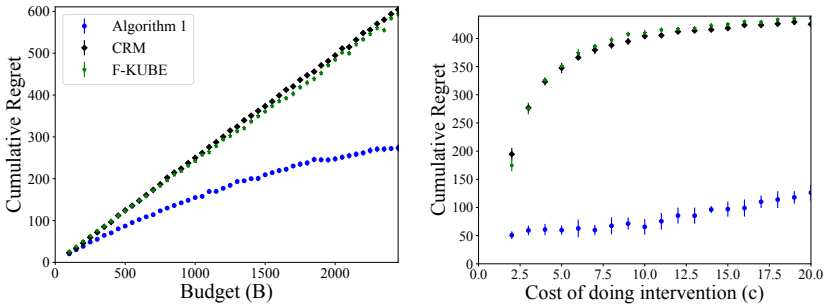

Figure 2: Cumulative regret on a general graph with $N = 6$.

The cumulative regret vs. budget plot in Figure 2 depicts the performance of algorithms by assuming that the cost of pulling $a_{i,x}$ for $i \in [N], x \in \{0, 1\}$ is selected randomly from $\{2, 3\}$. As the budget increases, the cumulative regret of all of the algorithms increases. However, the growth rate of the cumulative regrets of *F-KUBE* and *CRM* are higher than our algorithm. Moreover, since the cumulative regrets of *F-KUBE* and *CRM* do not converge to a constant for $B \leq 2500$, they fail to identify the optimal arm within this budget range while the regret of our algorithm remains a constant for large budgets, which indicates that it could identify the best arm in the experiment.

The cumulative regret vs. intervention cost in Figure 2 illustrates the performance of the algorithms when the budget was fixed to 1000 and $c_{i,x} = c$ for all $i \in [N], x \in \{0, 1\}$ such that

6. Python implementations are provided in the supplementary.

$c \in \{2, 3, \ldots, 20\}$ (uniform cost for all interventional arms). As shown in this figure, the cumulative regret of Algorithm 1 grows slower than the others. Note that since *CRM* considers only the causal graphs without hidden variables, for fairness, we compared these algorithms for the graph without hidden variables. The underlying graph for this experiment and additional experiments on graphs with hidden variables are provided in Appendix G.1.

6.2. Simple Regret in No-backdoor Graphs

In order to be able to compare our algorithm for simple regret with several related works, we studied the causal MAB for the special case of no-backdoor graphs in this section. We compared the performance of Algorithm 2 with two causal bandit algorithms γ-NB (Nair et al., 2021) and *PB* (Lattimore et al., 2016). *PB* is a non-budgeted algorithm that is designed to minimize the simple regret when the graph has no backdoor. γ-NB is a budgeted version of *PB* that allows uniform costs on arms, i.e., $c_{i,x} = c > 1$ for all i and x.

We used the same setting as in Nair et al. (2021) and Lattimore et al. (2016) in which the underlying graph has 50 intervenable variables and all of these variables are parents of the reward variable Y. This particular structure is called a parallel graph. We modeled $X_i \sim Bernoulli(p_i)$ with $p_1 = p_2 = 0.02$ for $i \in \{1, 2\}$ and $p_i = 0.5$ for $i \in \{3, \ldots, 50\}$. Moreover, we modeled the reward variable as $Y \sim Bernoulli(\frac{1}{2} + \epsilon)$ if $X_1 = 1$, and otherwise, $Y \sim Bernoulli(\frac{1}{2} - \epsilon')$, where $\epsilon = 0.3$ and $\epsilon' = \frac{p_1 \epsilon}{1 - p_1}$.

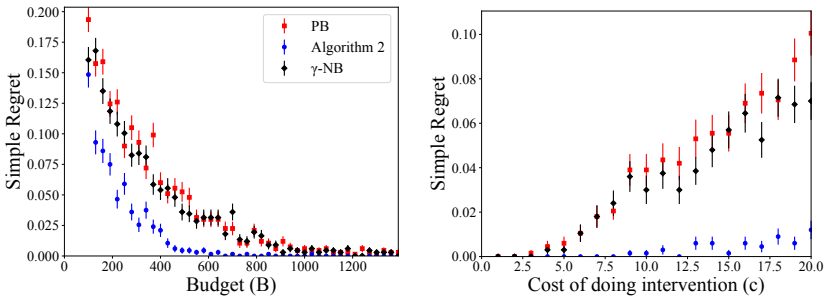

Figure 3: Simple regret on the parallel graph with $N = 50$.

The simple regret vs. budget plot in Figure 3 was obtained by selecting the cost of pulling each arm $a_{i,x}$ for $i \in [N], x \in \{0, 1\}$ randomly from $\{2, 3, 4, 5\}$. The simple regrets of all of the algorithms converge to zero as the budget increases, but Algorithm 2 demonstrates faster convergence. In the simple regret vs. the cost of intervention plot in Figure 3, we considered the setting in which the budget was fixed to 1500 and $c_{i,x} = c$ for $i \in [N], x \in \{0, 1\}$ such that $c \in \{1, 2, \ldots, 20\}$. The simple regret is increasing in terms of intervention costs, as expected. Since Algorithm 2 uses a different exploration set compared to the others, it drastically outperforms them even in a setting favorable to them. Additional experiments are presented in Appendix G.2 including an experiment using *Successive Rejects* algorithm in Audibert et al. (2010) which is a baseline MAB algorithm[7].

7. Successive Rejects is not included in the experiments of the main text as it fails to perform well for large N.

6.3. Simple Regret in General Graphs

We compared the performance of Algorithm 2 with two algorithms, SRM (Maiti et al., 2022) and *Successive Rejects* for general graphs in addition to the special structures of the previous section. SRM is a causal MAB algorithm for minimizing simple regret in the non-budgeted setting where the underlying graph is general. Here, we used a causal graph that violates the no-backdoor criterion. The graph has $N = 7$ intervenable variables, i.e., it has 15 arms (14 interventional and one observational arm). We used the same procedure to construct the model as Section 6.1.

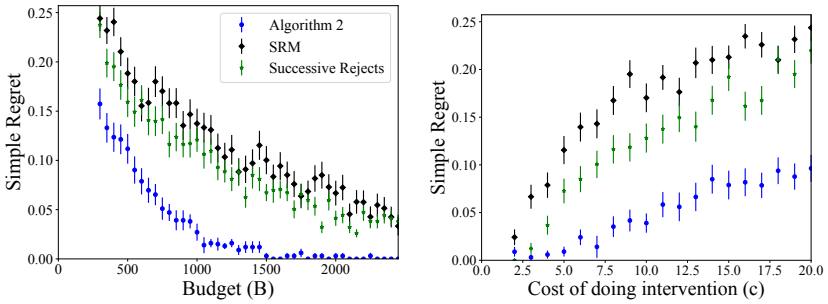

Figure 4: Simple regret on a general graph with $N = 7$.

The simple regret vs. budget plot in Figure 4 illustrates the performance of algorithms when the cost of each interventional arm was selected randomly from $\{5, 6, 7\}$. Algorithm 2 converges to 0 faster than the others as B grows. The simple regret vs. the cost of intervention plot in Figure 4 compares the performance of the algorithms when the budget is fixed to 800, and the cost of all interventional arms is equal to c, where $c \in \{2, 3, \ldots, 20\}$. Additional experiments and the underlying graph of this experiment are provided in Appendix G.3.

7. Conclusion

We studied the budgeted causal MAB problem with non-uniform costs for different arms in general causal graphs in which all intervenable variables have identifiable causal effects. We considered two different learners; one with simple regret as its objective and the other with cumulative regret objective. For each learner, we proposed an algorithm and provided theoretical guarantees. Furthermore, through empirical studies in different scenarios, we evaluated the performances of our proposed algorithms and showed that they outperform the state-of-the-art.

References

Jayadev Acharya, Sourbh Bhadane, Arnab Bhattacharyya, Saravanan Kandasamy, and Ziteng Sun. Sample complexity of distinguishing cause from effect. In *International Conference on Artificial Intelligence and Statistics*, pages 10487–10504. PMLR, 2023.

Sina Akbari, Fateme Jamshidi, Ehsan Mokhtarian, Matthew James Vowels, Jalal Etesami, and Negar Kiyavash. Causal effect identification in uncertain causal networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010.

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

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

Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, Ashwin Maran, and Vinodchandran N Variyam. Learning and sampling of atomic interventions from observations. In *International Conference on Machine Learning*, pages 842–853. PMLR, 2020.

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*, pages 1516–1541, 2013.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

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

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. *COLT*, pages 355–366, 2008.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

Fateme Jamshidi, Sina Akbari, and Negar Kiyavash. Causal imitability under context-specific independence relations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Fateme Jamshidi, Luca Ganassali, and Negar Kiyavash. On sample complexity of conditional independence testing with von mises estimator with application to causal discovery. *arXiv preprint arXiv:2310.13553*, 2023b.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

Dan J Kim, Donald L Ferrin, and H Raghav Rao. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision support systems*, 44(2):544–564, 2008.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pages 1875–1884. PMLR, 2017.

Tze Leung Lai, Herbert Robbins, et al. 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. *Advances in Neural Information Processing Systems*, 29, 2016.

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

Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems*, 31, 2018.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4164–4172, 2019.

Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

Erik Lindgren, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Experimental design for cost-aware learning of causal graphs. *Advances in Neural Information Processing Systems*, 31, 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.

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

Aurghya Maiti, Vineet Nair, and Gaurav Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, 12, 1999.

Nicolai Meinshausen, Alain Hauser, Joris M Mooij, Jonas Peters, Philip Versteeg, and Peter Bühlmann. Methods for causal inference from gene perturbation experiments and validation. *Proceedings of the National Academy of Sciences*, 113(27):7361–7368, 2016.

Ehsan Mokhtarian, Sina Akbari, Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Learning bayesian networks in the presence of structural side information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7814–7822, 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. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

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

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

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. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 519–527, San Francisco, CA, USA, 2002a. Morgan Kaufmann Publishers Inc. ISBN 1558608974.

Jin Tian and Judea Pearl. *A general identification condition for causal effects*. eScholarship, University of California, 2002b.

Santtu Tikka, Antti Hyttinen, and Juha Karvanen. Identifying causal effects via context-specific independence relations. *Advances in neural information processing systems*, 32, 2019.

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas Jennings. Knapsack based optimal policies for budget–limited multi–armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 1134–1140, 2012.

Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '88, page 69–78, NLD, 1990. North-Holland Publishing Co. ISBN 0444886508.

Appendix A. Technical preliminaries

Definition 10 (Directed Path) *Let $V_1, V_2, \ldots, V_m$ be a set of distinct vertices in an ADMG $\mathcal{G}$. There is a directed path from V_1 to V_m if $V_i \in \mathbf{Pa}(V_{i+1})$ for every $1 \leq i \leq m-1$.*

Definition 11 (Descendant) *Let X_i and X_j be two vertices in an ADMG $\mathcal{G}$. X_j is called a descendant of X_i if there exists a directed path from X_i to X_j.*

Definition 12 (Blocked) *Given a causal graph $\mathcal{G}$ and two vertices $X_1, X_n \in \mathbf{V}$, a path between X_1 and X_n is called blocked by a set of vertices $\mathbf{W}$ (with neither X_1 nor X_n in $\mathbf{W}$) whenever there is a vertex X_k, such that one of the followings holds:*

(1) $X_k \in \mathbf{W}$ and $X_{k-1} \to X_k \to X_{k+1}$ or $X_{k-1} \leftarrow X_k \leftarrow X_{k+1}$ or $X_{k-1} \leftarrow X_k \to X_{k+1}$,

(2) $X_{k-1} \to X_k \leftarrow X_{k+1}$ and neither X_k nor any of its descendants is in $\mathbf{W}$.

Lemma 13 (Chernoff inequalities) *Let X be a random variable. Then, for every $s \geq 0$, the followings hold:*

(1) $P(X \geq \mathbb{E}[X] + s) \leq \min_{\lambda \geq 0} \mathbb{E}\big[\exp\big(\lambda(X - \mathbb{E}[X])\big)\big] \exp(-\lambda s)$

(2) $P(X \leq \mathbb{E}[X] - s) \leq \min_{\lambda \geq 0} \mathbb{E}\big[\exp\big(\lambda(\mathbb{E}[X] - X)\big)\big] \exp(-\lambda s)$

Lemma 14 (Hoeffding inequalities) *Let X be a random variable such that $X \in [a, b]$. Therefore, for every $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\Big[\exp\big(\lambda(X - \mathbb{E}[X])\big)\Big] \leq \exp\Big(\frac{\lambda^2 (b-a)^2}{8}\Big).$$

Lemma 15 (Chernoff-Hoeffeding inequality) *Assume $X^1, \ldots, X^T$ are independent random variables such that $0 \leq X^t \leq 1$ for $t = 1, \ldots, T$. Then, for every $\epsilon > 0$, the following inequalities hold:*

(1) $P\Big(\sum_{t\in[T]} X^t - \mathbb{E}\big[\sum_{t\in[T]} X^t\big] \geq \epsilon\Big) \leq \exp(\frac{-2\epsilon^2}{T})$,

(2) $P\Big(\sum_{t\in[T]} X^t - \mathbb{E}\big[\sum_{t\in[T]} X^t\big] \leq -\epsilon\Big) \leq \exp(\frac{-2\epsilon^2}{T})$.

Appendix B. Proofs of Section 3

In order to prove Theorem 5, we require several technical lemmas which we present below.

In the following lemma, we show that $\hat{\mu}^t_{i,x}$ is an unbiased estimator of $\mu_{i,x}$.

Lemma 16 *$\hat{\mu}^t_{i,x}$ in (8) and $\hat{\mu}^t_0$ in (5) are unbiased estimators of $\mu_{i,x}$ and μ_0.*

Proof Recall Equation (8):

$$\hat{\mu}^t_{i,x} := \frac{\sum_{t' \in \mathbf{I}^t_{i,j}} \mathbb{1}\{Y^{t'} = 1\} + \sum_{s \in [S^t_{i,x}]} Y^s_{i,x}}{N^t_{i,x} + S^t_{i,x}}.$$

Note that $Y^s_{i,x}$ is an unbiased estimator of $\mu_{i,x}$ because we partition the time steps that arm a_0 was pulled into $|\mathbf{V}|$ different number of subsets. Taking expectations from both sides of the above equation yields

$$\begin{aligned}
\mathbb{E}[\hat{\mu}^t_{i,x}] &= \mathbb{E}\left[\frac{\sum_{t'\in\mathbf{I}^t_{i,j}} \mathbb{1}\{Y^{t'}=1\} + \sum_{s\in[S^t_{i,x}]} Y^s_{i,x}}{N^t_{i,x}+S^t_{i,x}}\right] \\
&= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} \mathbb{E}\left[\frac{\sum_{t'\in\mathbf{I}^t_{i,j}} \mathbb{1}\{Y^{t'}=1\} + \sum_{s\in[S^t_{i,x}]} Y^s_{i,x}}{N^t_{i,x}+S^t_{i,x}}\Big| N^t_{i,x}=a, S^t_{i,x}=b\right] P(N^t_{i,x}=a, S^t_{i,x}=b) \\
&= \sum_{a=1}^{\infty}\sum_{b=0}^{\infty} \mathbb{E}\Big[\frac{a\mu_{i,x}+b\mu_{i,x}}{a+b}|N^t_{i,x}=a, S^t_{i,x}=b\Big] P(N^t_{i,x}=a, S^t_{i,x}=b) \\
&= \mu_{i,x}\sum_{a=1}^{\infty}\sum_{b=0}^{\infty} P(N^t_{i,x}=a, S^t_{i,x}=b) = \mu_{i,x}.
\end{aligned}$$

Similarly, one can show that $\hat{\mu}^t_0$ is an unbiased estimator of μ_0. ■

Next, we show a concentration result for $\hat{\mu}^t_{i,x}$ in (8).

Lemma 17 *For $\hat{\mu}^t_{i,x}$ given in Equation (8), we have $P\big(|\hat{\mu}^t_{i,x}-\mu_{i,x}|\geq\epsilon\big)\leq 2\exp\big(-2(N^t_{i,x}+S^t_{i,x})\epsilon^2\big)$.*

Proof

$$\begin{aligned}
&P(\hat{\mu}^t_{i,x}-\mu_{i,x}\geq\epsilon) = P\left(\frac{\sum_{j\in\mathbf{I}^t_{i,x}} \mathbb{1}\{Y^{t'}=1\} + \sum_{s\in[S^t_{i,x}]} Y^s_{i,x}}{N^t_{i,x}+S^t_{i,x}} \geq \mu_{i,x}+\epsilon\right) \\
&= P\left(\sum_{t'\in\mathbf{I}^t_{i,x}} \mathbb{1}\{Y^{t'}=1\} + \sum_{s\in[S^t_{i,x}]} Y^s_{i,x} \geq (N^t_{i,x}+S^t_{i,x})\mu_{i,x} + (N^t_{i,x}+S^t_{i,x})\epsilon\right) \\
&\overset{(a)}{\leq} \min_{\lambda\geq 0}\mathbb{E}\left[\exp\Big(\lambda\big(\sum_{t'\in\mathbf{I}^t_{i,x}}(\mathbb{1}\{Y^{t'}=1\}-\mu_{i,x}) + \sum_{s\in[S^t_{i,x}]}(Y^s_{i,x}-\mu_{i,x})\big)\Big)\right]\exp\big(-\lambda(N^t_{i,x}+S^t_{i,x})\epsilon\big) \\
&= \min_{\lambda\geq 0}\mathbb{E}\left[\prod_{t'\in\mathbf{I}^t_{i,x}}\exp\big(\lambda(\mathbb{1}\{Y^{t'}=1\}-\mu_{i,x})\big)\prod_{s\in[S^t_{i,x}]}\exp\big(\lambda(Y^s_{i,x}-\mu_{i,x})\big)\right]\exp\big(-\lambda(N^t_{i,x}+S^t_{i,x})\epsilon\big) \\
&\overset{(b)}{=} \min_{\lambda\geq 0}\prod_{t'\in\mathbf{I}^t_{i,x}}\mathbb{E}\Big[\exp\big(\lambda(\mathbb{1}\{Y^{t'}=1\}-\mu_{i,x})\big)\Big]\prod_{s\in[S^t_{i,x}]}\mathbb{E}\Big[\exp\big(\lambda(Y^s_{i,x}-\mu_{i,x})\big)\Big]\exp\big(-\lambda(N^t_{i,x}+S^t_{i,x})\epsilon\big) \\
&\overset{(c)}{\leq} \min_{\lambda\geq 0}\exp\Big(\frac{N^t_{i,x}\lambda^2}{8}+\frac{S^t_{i,x}\lambda^2}{8}-\lambda(N^t_{i,x}+S^t_{i,x})\epsilon\Big) \\
&\overset{(d)}{\leq} \exp\big(-2(N^t_{i,x}+S^t_{i,x})\epsilon^2\big).
\end{aligned} \tag{12}$$

The inequality in (a) holds using Lemma 13. The equality in (b) is true since the terms in the product are independent. The inequality in (c) follows from Lemma 14 where $Y^{t'}, Y^s_{i,x} \in \{0,1\}$. Finally, the inequality in (d) follows after substituting the optimal $\lambda := 4\epsilon$. In a similar way, it can be shown

$$P\big(\hat{\mu}^t_{i,x} - \mu_{i,x} \leq -\epsilon\big) \leq \exp\big(-2(N^t_{i,x} + S^t_{i,x})\epsilon^2\big).$$

Therefore, we get

$$P\big(|\hat{\mu}^t_{i,x} - \mu_{i,x}| \geq \epsilon\big) \leq 2\exp\big(-2(N^t_{i,x} + S^t_{i,x})\epsilon^2\big).$$

■

In the following lemma, we introduce a bound for the expectation of $S^T_{i,x}$. Recall that $S^T_{i,x} := \min\Big\{\min_{j:V_j\in\mathbf{C}_i} S^T_{j,i}, \min_{j:V_j\notin\mathbf{C}_i} \tilde{S}^T_{j,i,x}\Big\}$ where

$$\begin{cases} S^T_{j,i} := \min_{\mathbf{w}'_i} \min_{x'} |\mathbf{O}^T_j(x', \mathbf{w}'_i)| & \text{if } V_j \in \mathbf{C}_i, \\ \tilde{S}^T_{j,i,x} := \min_{\mathbf{w}'_i} |\mathbf{O}^T_j(x, \mathbf{w}'_i)| & \text{if } V_j \notin \mathbf{C}_i. \end{cases}$$

Lemma 18 *Let $\mathcal{W}_i$ be the size of the domain set of $\mathbf{V}\setminus\{X_i, Y\}$ and $p_{i,x} := \min_j \min_{\mathbf{w}'_i} p_{j,i,x}(\mathbf{w}'_i)$. Moreover, define $\tau_{i,x,T} := \max(0, 1 - |\mathbf{V}|\mathcal{W}_i T^{-\frac{p^2_{i,x}}{2|\mathbf{V}|}})$. Then,*

$$\mathbb{E}[S^T_{i,x}] \geq \frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N^T_0]\tau_{i,x,T} - 1.$$

Proof

We define

$$\hat{p}^T_{j,i,x}(\mathbf{w}'_i) := \begin{cases} \frac{\min_{x'} |\mathbf{O}^T_j(x', \mathbf{w}'_i)|}{|\mathbf{O}^T_j|} & \text{if } V_j \in \mathbf{C}_i, \\ \frac{|\mathbf{O}^T_j(x, \mathbf{w}'_i)|}{|\mathbf{O}^T_j|} & \text{otherwise.} \end{cases}$$

where $|O^T_j| = \Big\lfloor \frac{N^T_0}{|\mathbf{V}|} \Big\rfloor$. Moreover, let $\hat{p}^T_{i,x} := \min_j \min_{\mathbf{w}'_i} \hat{p}^T_{j,i,x}(\mathbf{w}'_i)$. Using the above definition, we have

$$P\Big(\hat{p}^T_{j,i,x}(\mathbf{w}'_i) \leq \frac{p_{i,x}}{2}\Big) \overset{(a)}{\leq} P\Big(\hat{p}^T_{j,i,x}(\mathbf{w}'_i) \leq p_{j,i,x}(\mathbf{w}'_i) - \frac{p_{i,x}}{2}\Big) \overset{(b)}{\leq} \exp\Big(-2\frac{p^2_{i,x}}{4}\frac{\ln T}{|\mathbf{V}|}\Big) = T^{-\frac{p^2_{i,x}}{2|\mathbf{V}|}}. \tag{13}$$

Inequality (a) holds because $\frac{p_{i,x}}{2} \leq p_{j,i,x}(\mathbf{w}'_i) - \frac{p_{i,x}}{2}$ and (b) follows from Lemma 15. Therefore, we have

$$P\Big(\min_j \min_{\mathbf{w}'_i} \hat{p}^T_{j,i,x}(\mathbf{w}'_i) \leq \frac{p_{i,x}}{2}\Big) \leq \sum_j \sum_{\mathbf{w}'_i} P\Big(\hat{p}^T_{j,i,x}(\mathbf{w}'_i) \leq \frac{p_{i,x}}{2}\Big) \overset{(a)}{\leq} |\mathbf{V}|\mathcal{W}_i T^{-\frac{p^2_{i,x}}{2|\mathbf{V}|}}. \tag{14}$$

where $\mathcal{W}_i$ is the alphabet size of $\mathbf{V} \setminus \{X_i, Y\}$ and the inequality (a) follows from Equation (13). Finally, from the definition of $S_{i,x}^T$, we get

$$\begin{aligned}
\mathbb{E}[S_{i,x}^T] &\geq \mathbb{E}\left[\min_j \min_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i') \left\lfloor \frac{N_0^T}{|\mathbf{V}|} \right\rfloor\right] \\
&\geq \frac{1}{|\mathbf{V}|}\mathbb{E}\Big[\min_j \min_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i')(N_0^T - |\mathbf{V}|)\Big] \\
&= \frac{1}{|\mathbf{V}|}\mathbb{E}\Big[\min_j \min_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i')N_0^T - \max_j \max_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i')|\mathbf{V}|\Big] \\
&\geq \frac{1}{|\mathbf{V}|}\mathbb{E}\Big[\min_j \min_{\mathbf{w}'} \hat{p}_{j,i,x}^T(\mathbf{w}_i')N_0^T\Big] - 1 \\
&= \frac{1}{|\mathbf{V}|}\sum_{n=1}^{\infty} n\mathbb{E}\Big[\min_j \min_{\mathbf{w}'} \hat{p}_{j,i,x}^T(\mathbf{w}_i')|N_0^T = n\Big]P(N_0^T = n) - 1 \\
&\geq \frac{1}{|\mathbf{V}|}\sum_{n=1}^{\infty} n\frac{p_{i,x}}{2}P\Big(\min_j \min_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i') > \frac{p_{i,x}}{2}|N_0^T = n\Big)P(N_0^T = n) - 1 \\
&\geq \frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N_0^T]\max(0, 1 - |\mathbf{V}|\mathcal{W}_i T^{-\frac{p_{i,x}^2}{2|\mathbf{V}|}}) - 1 \\
&= \frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N_0^T]\tau_{i,x,T} - 1.
\end{aligned}$$

■

Lemma 19 *Suppose that $a_{i,x}$ is not the optimal arm, i.e., $a^* \neq a_{i,x}$. In this case, we have*

$$\mathbb{E}[N_{i,x}^T] \leq \frac{8\ln T}{\delta_{i,x}^2} + 2 - \frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N_0^l]\tau_{i,x,l} + \frac{\pi^2}{3},$$

where $l := \frac{8\ln T}{\delta_{i,x}^2} + 1$. Moreover, if $a^* \neq a_0$, then,

$$\mathbb{E}[N_0^T] \leq \frac{8\ln T}{\delta_0^2} + 1 + \frac{\pi^2}{3}.$$

Proof Define $E_{i,x}^t$ to be effective number of pulling arm $a_{i,x}$ at the end of time t and let $E_{i,x}^t := N_{i,x}^t + S_{i,x}^t$. Using this definition, we rewrite $N_{i,x}^T$ as follows

$$N_{i,x}^T = \sum_{t\in[l]} \mathbb{1}\{a^t = a_{i,x}, E_{i,x}^t \leq l\} + \sum_{t\in[l+1,T]} \mathbb{1}\{a^t = a_{i,x}, E_{i,x}^t > l\}. \tag{15}$$

Let $m := \max\{t|E_{i,x}^t \leq l\}$, then, the first part of Equation (15) will be equal to $N_{i,x}^m$, i.e.,

$$N_{i,x}^T = N_{i,x}^m + \sum_{t\in[l+1,T]} \mathbb{1}\{a^t = a_{i,x}, E_{i,x}^t > l\}.$$

Since $N^m_{i,x} = E^m_{i,x} - S^m_{i,x}$, we get $N^m_{i,x} = \sum_{t\in[m]} \mathbb{1}\{a^t = a_{i,x}\} = l - S^m_{i,x}$. This allows us to rewrite Equation (15) as

$$N^T_{i,x} = l - S^m_{i,x} + \sum_{t\in[l+1,T]} \mathbb{1}\{a^t = a_{i,x}, E^t_{i,x} > l\},$$

and since $m \geq l$, we have $S^m_{i,x} \geq S^l_{i,x}$. Therefore,

$$N^T_{i,x} \leq l - S^l_{i,x} + \sum_{t\in[l+1,T]} \mathbb{1}\{a^t = a_{i,x}, E^t_{i,x} > l\}. \tag{16}$$

By taking expectation on both sides of Equation (16) we have

$$\mathbb{E}[N^T_{i,x}] \leq l - \mathbb{E}[S^l_{i,x}] + \sum_{t\in[l+1,T]} P(a^t = a_{i,x}, E^t_{i,x} > l).$$

Using Lemma 18, we rewrite the above inequality as

$$\mathbb{E}[N^T_{i,x}] \leq l + 1 - \frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N^l_0]\tau_{i,x,l} + \sum_{t\in[l+1,T]} P(a^t = a_{i,x}, E^t_{i,x} > l). \tag{17}$$

Next, we bound $\sum_{t\in[l+1,T]} P(a^t = a_{i,x}, E^t_{i,x} > l)$,

$$\sum_{t\in[l+1,T]} P(a^t = a_{i,x}, E^t_{i,x} > l) \leq \sum_{t\in[l+1,T]} P(\bar{\mu}^{t-1}_{i,x} \geq \bar{\mu}^{t-1}_{a^*}, E^t_{i,x} > l) = \sum_{t\in[l,T-1]} P(\bar{\mu}^t_{i,x} \geq \bar{\mu}^t_{a^*}, E^t_{i,x} \geq l).$$

For clarity, we use $\hat{\mu}^t_a(E^t_a)$ instead of $\hat{\mu}^t_a$. By substituting the definitions of the UCB, the right hand side of the equation becomes

$$\begin{aligned}
&\sum_{t\in[l,T-1]} P\Big(\frac{\hat{\mu}^t_{i,x}(E^t_{i,x})}{c_{i,x}} + \sqrt{\frac{2\ln t}{c^2_{i,x}E^t_{i,x}}} \geq \frac{\hat{\mu}^t_{a^*}(E^t_{a^*})}{c_{a^*}} + \sqrt{\frac{2\ln t}{c^2_{a^*}E^t_{a^*}}}, E^t_{i,x} \geq l\Big) \\
&\overset{(a)}{\leq} \sum_{t\in[l,T-1]} P\Big(\max_{t_1\in[l+1,t]} \frac{\hat{\mu}^t_{i,x}(t_1)}{c_{i,x}} + \sqrt{\frac{2\ln t}{c^2_{i,x}t_1}} \geq \min_{t_2\in[l+1,t]} \frac{\hat{\mu}^t_{a^*}(t_2)}{c_{a^*}} + \sqrt{\frac{2\ln t}{c^2_{a^*}t_2}}\Big) \\
&\leq \sum_{t\in[T]} \sum_{t_1\in[l,t-1]} \sum_{t_2\in[l,t-1]} P\Big(\frac{\hat{\mu}^t_{i,x}(t_1)}{c_{i,x}} + \sqrt{\frac{2\ln t}{c^2_{i,x}t_1}} \geq \frac{\hat{\mu}^t_{a^*}(t_2)}{c_{a^*}} + \sqrt{\frac{2\ln t}{c^2_{a^*}t_2}}\Big).
\end{aligned}$$

Inequality (a) holds because

$$\max_{t_1\in[l+1,t]} \hat{\mu}^t_{i,x}(t_1) + \sqrt{\frac{2\ln t}{t_1}} \geq \hat{\mu}^t_{i,x}(E^t_{i,x}) + \sqrt{\frac{2\ln t}{E^t_{i,x}}},$$

and

$$\min_{t_2\in[l+1,t]} \hat{\mu}^t_{a^*}(t_2) + \sqrt{\frac{2\ln t}{t_2}} \leq \hat{\mu}^t_{a^*}(E^t_{a^*}) + \sqrt{\frac{2\ln t}{E^t_{a^*}}}.$$

It can be shown that if none of the following hold, then $\frac{\hat{\mu}^t_{i,x}(t_1)}{c_{i,x}} + \sqrt{\frac{2\ln t}{c^2_{i,x}t_1}} \geq \frac{\hat{\mu}^t_{a^*}(t_2)}{c_{a^*}} + \sqrt{\frac{2\ln t}{c^2_{a^*}t_2}}$ does not hold as well,

$$\hat{\mu}^t_{i,x}(t_1) - \mu_{i,x} \geq \sqrt{\frac{2\ln t}{t_1}}, \tag{18}$$

$$\hat{\mu}^t_{a^*}(t_2) - \mu_{a^*} \leq -\sqrt{\frac{2\ln t}{t_2}}, \tag{19}$$

$$\frac{\mu_{a^*}}{c_{a^*}} - \frac{\mu_{i,x}}{c_{i,x}} \leq 2\sqrt{\frac{2\ln t}{c^2_{i,x}t_1}}. \tag{20}$$

Now, we bound the probability of events in Equations (18) and (19),

$$P\big(\hat{\mu}^t_{i,x}(t_1) - \mu_{i,x} \geq \sqrt{\frac{2\ln t}{t_1}}\big) \leq \exp\big(-2\frac{2\ln t}{t_1}t_1\big) = t^{-4},$$

$$P\big(\hat{\mu}^t_{a^*}(t_2) - \mu_{a^*} \leq -\sqrt{\frac{2\ln t}{t_2}}\big) \leq \exp\big(-2\frac{2\ln t}{t_2}t_2\big) = t^{-4},$$

where we used Lemma 15 to obtain the both above inequalities. Furthermore, by assuming that $l := \frac{8\ln T}{\delta^2_{i,x}} + 1$, the event in Equation (20) is false,

$$\sum_{t\in[l+1,T]} P(a^t = a_{i,x}, E^t_{i,x} > l) \leq \sum_{t\in[T]} \sum_{t_1\in[l,t-1]} \sum_{t_2\in[l,t-1]} 2t^{-4} \leq \frac{\pi^2}{3}. \tag{21}$$

Therefore, if $a^* \neq a_{i,x}$, using Equations (17) and (21), we obtain the following bound for $N^T_{i,x}$:

$$\mathbb{E}[N^T_{i,x}] \leq \frac{8\ln T}{\delta^2_{i,x}} + 2 - \frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N^l_0]\tau_{i,x,l} + \frac{\pi^2}{3}.$$

For the second part of the proof, suppose that $a^* \neq a_0$. In this case, we decompose N^T_0 in two parts,

$$N^T_0 = \sum_{t\in[T]} \mathbb{1}\{a^t = a_0\} = \sum_{t\in[l]} \mathbb{1}\{a^t = a_0, N^T_0 \leq l\} + \sum_{t\in[l+1,T]} \mathbb{1}\{a^t = a_0, N^T_0 > l\}.$$

By taking expectation from both sides of the above inequality, we get

$$\begin{aligned}\mathbb{E}[N^T_0] &= \sum_{t\in[l]} P(a^t = a_0, N^T_0 \leq l) + \sum_{t\in[l+1,T]} P(a^t = a_0, N^T_0 > l) \\ &\leq l + \sum_{t\in[l+1,T]} P(a^t = a_0, N^t_0 > l).\end{aligned}$$

Following the same procedure used to bound $\sum_{t\in[l+1,T]} P(a^t = a_{i,x}, E^t_{i,x} > l)$ and for $l = \frac{8\ln T}{\delta^2_0} + 1$, we obtain

$$\sum_{t\in[l+1,T]} P(a^t = a_0, N^t_0 > l) \leq \frac{\pi^2}{3}.$$

This implies the following bound for $\mathbb{E}[N_0^T]$,

$$\mathbb{E}[N_0^T] \leq \frac{8\ln T}{\delta_0^2} + 1 + \frac{\pi^2}{3}.$$

■

Lemma 20 *If $a^* = a_0$, we have the following bound for $\mathbb{E}[N_0^T]$,*

$$\mathbb{E}[N_0^T] \geq T - 2N(2 + \frac{\pi^2}{3}) - \sum_{i,x} \frac{8\ln T}{\delta_{i,x}^2}.$$

Proof From definition of N_a^T, we know $N_0^T = T - \sum_{i,x} N_{i,x}^T$. If $a^* \neq a_{i,x}$ we have the following by Lemma 19:

$$\mathbb{E}[N_{i,x}^T] \leq \frac{8\ln T}{\delta_{i,x}^2} + 2 + \frac{\pi^2}{3}.$$

Then,

$$\mathbb{E}[N_0^T] \geq T - 2N(2 + \frac{\pi^2}{3}) - \sum_{i,x} \frac{8\ln T}{\delta_{i,x}^2}.$$

■

Lemma 21 *Suppose that $a^* \neq a_0$ and let $\delta_0 := \frac{\mu_{a^*}}{c_{a^*}} - \mu_0$. We also define*

$$\hat{p}_{j,i,x}^T(\mathbf{w}_i') := \begin{cases} \frac{\min_{x'} |\mathbf{O}_j^T(x', \mathbf{w}_i')|}{|\mathbf{O}_j^T|} & \text{if } V_j \in \mathbf{C}_i, \\ \frac{|\mathbf{O}_j^T(x, \mathbf{w}_i')|}{|\mathbf{O}_j^T|} & \text{otherwise.} \end{cases}$$

Moreover, let $\hat{p}_{i,x}^T := \min_j \min_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i')$, $p_{i,x} := \min_j \min_{\mathbf{w}_i'} p_{j,i,x}(\mathbf{w}_i')$, and $p := \min_{i,x} p_{i,x}$. Then,

$$P\Big(|\hat{\mu}_0^T - \mu_0| \geq \frac{\delta_0}{4}\Big) \leq 2T^{-\frac{\delta_0^2}{8}},$$

and

$$P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}\Big) \leq |\mathbf{V}|\mathcal{W}_i T^{-\frac{p^2}{2|\mathbf{V}|}} + 2T^{-\frac{p\delta_0^2 c_{i,x}^2}{16|\mathbf{V}|}},$$

where $\mathcal{W}_i$ is the size of domain from which $\mathbf{V} \setminus \{X_i, Y\}$ takes values.

Proof By Algorithm 1, the number of times that arm a_0 is pulled at the end of T rounds, denoted by N_0^T, is at least $\beta^2 \ln T$ and since $\beta \geq 1$, we have $N_0^T \geq \ln T$. Using Lemma 15, we

$$P\Big(|\hat{\mu}_0^T - \mu_0| \geq \frac{\delta_0}{4}\Big) \leq 2\exp(-2\frac{\delta_0^2}{16}\ln T) = 2T^{-\frac{\delta_0^2}{8}}.$$

Next, we prove the second inequality as follows.

$$P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}\Big) = P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}, \hat{p}_{i,x}^T \leq \frac{p}{2}\Big) + P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}, \hat{p}_{i,x}^T > \frac{p}{2}\Big)$$
$$\leq P(\hat{p}_{i,x}^T \leq \frac{p}{2}) + P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}, \hat{p}_{i,x}^T > \frac{p}{2}\Big) \tag{22}$$

Now, we bound the first part in Equation (22). To do so, first, we get

$$P\Big(\hat{p}_{j,i,x}^T(\mathbf{w}_i') \leq \frac{p}{2}\Big) \overset{(a)}{\leq} P\Big(\hat{p}_{j,i,x}^T(\mathbf{w}_i') \leq p_{j,i,x}(\mathbf{w}_i') - \frac{p}{2}\Big) \overset{(b)}{\leq} \exp(-2\frac{p^2}{4}.\frac{\ln T}{|\mathbf{V}|}) = T^{-\frac{p^2}{2|\mathbf{V}|}}. \tag{23}$$

Note that in Equation (23), the inequality in (a) holds because $\frac{p}{2} \leq p_{j,i,x}(\mathbf{w}') - \frac{p}{2}$ and the inequality in (b) follows from Lemma 15. Therefore,

$$P(\hat{p}_{i,x}^T \leq \frac{p}{2}) = P\big(\min_j \min_{\mathbf{w}_i'} \hat{p}_{j,i,x}^T(\mathbf{w}_i') \leq \frac{p}{2}\big) \leq \sum_j \sum_{\mathbf{w}_i'} P\big(\hat{p}_{j,i,x}^T(\mathbf{w}_i') \leq \frac{p}{2}\big) \overset{(a)}{\leq} |\mathbf{V}|\mathcal{W}_i T^{-\frac{p^2}{2|\mathbf{V}|}}, \tag{24}$$

where the inequality in (a) follows from Equation (23).

Next, we bound the second part of Equation (22). From Algorithm 1, we have $\beta \geq 1$, and therefore, $N_0^T \geq \ln T$. Now, if $\hat{p}_{i,x}^T > \frac{p}{2}$, then $S_{i,x}^T > \frac{p}{2}\frac{N_0^T}{|\mathbf{V}|} \geq \frac{p}{2|\mathbf{V}|}\ln T$. Therefore, using Lemma 17 and 15 we have the following bound for the second part of Equation (22):

$$P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}, \hat{p}_{i,x}^T > \frac{p}{2}\Big) \leq 2\exp\big(-2.\frac{\delta_0^2 c_{i,x}^2}{16}\frac{p}{2|\mathbf{V}|}\ln T\big) = 2T^{-\frac{p\delta_0^2 c_{i,x}^2}{16|\mathbf{V}|}}. \tag{25}$$

Finally, using Equations (24) and (25), we rewrite Equation (22) as follows,

$$P\Big(|\frac{\hat{\mu}_{i,x}^T}{c_{i,x}} - \frac{\mu_{i,x}}{c_{i,x}}| \geq \frac{\delta_0}{4}\Big) \leq |\mathbf{V}|\mathcal{W}_i T^{-\frac{p^2}{2|\mathbf{V}|}} + 2T^{-\frac{p\delta_0^2 c_{i,x}^2}{16|\mathbf{V}|}}.$$

■

Lemma 22 *Assume that $a^* \neq a_0$ and let $\delta_0 = \frac{\mu_{a^*}}{c_{a^*}} - \mu_0$. If $T \geq \max\big\{e^{\frac{32}{\delta_0}}, \arg\min_t\{t^{\frac{p\delta_0^2}{16|\mathbf{V}|}} \geq \frac{8N(3+|\mathbf{V}|\mathcal{W})}{3}\}\big\}$, where $\mathcal{W} := \max_i \mathcal{W}_i$, then $\mathbb{E}[\beta^2] \geq \frac{8}{9\delta_0}$.*

Proof For each arm $a \in \mathcal{A}$, let e_a be the event that $|\frac{\hat{\mu}_a^T}{c_a} - \frac{\mu_a}{c_a}| \leq \frac{\delta_0}{4}$ and define $e := \cap_{a\in\mathcal{A}} e_a$. Furthermore, let $\bar{e}_a$ and $\bar{e}$ denote the compliment of the events e_a and e, respectively. Lemma 21 implies the following inequalities for a_0 and every $a_{i,x} \in \mathcal{A}$,

$$P(\bar{e}_0) \leq 2T^{-\frac{\delta_0^2}{8}},$$

$$P(\bar{e}_{i,x}) \leq |\mathbf{V}|\mathcal{W}_i T^{-\frac{p^2}{2|\mathbf{V}|}} + 2T^{-\frac{p\delta_0^2 c_{i,x}^2}{16|\mathbf{V}|}}.$$

Therefore, using the above equations and the union bound, we get

$$\begin{aligned} P(\bar{e}) &\overset{(a)}{\leq} 2T^{-\frac{\delta_0^2}{8}} + 2N\big(|\mathbf{V}|\mathcal{W}T^{-\frac{p^2}{2|\mathbf{V}|}} + 2T^{-\frac{p\delta_0^2}{16|\mathbf{V}|}}\big) \\ &\overset{(b)}{\leq} 2NT^{-\frac{p\delta_0^2}{16|\mathbf{V}|}} + 2N\big(|\mathbf{V}|\mathcal{W}T^{-\frac{p\delta_0^2}{16|\mathbf{V}|}} + 2T^{-\frac{p\delta_0^2}{16|\mathbf{V}|}}\big) \\ &= 2N(3+|\mathbf{V}|\mathcal{W})T^{-\frac{p\delta_0^2}{16|\mathbf{V}|}}, \end{aligned}$$

where the inequality in (a) holds since $c_{i,x} \geq 1$ for every $i \in [N], x \in \{0,1\}$ and (b) holds since $p \leq 1, \delta_0 \leq 1$.

Let $\hat{\mu}_{\tilde{a}}^T := \max_{a\in\mathcal{A}} \frac{\hat{\mu}_a^T}{c_a}$. By the definition of $\tilde{a}$ and δ_0, we have $\frac{\mu_{\tilde{a}}}{c_{\tilde{a}}} - \mu_0 \leq \delta_0$. If event e is true, then

$$\begin{aligned} -\frac{\delta_0}{2} \leq \frac{\hat{\mu}_{\tilde{a}}^T}{c_{\tilde{a}}} - \hat{\mu}_0^T + (\mu_0 - \frac{\mu_{\tilde{a}}}{c_{\tilde{a}}}) \leq \frac{\delta_0}{2} \\ \implies \frac{\hat{\mu}_{\tilde{a}}^T}{c_{\tilde{a}}} - \hat{\mu}_0^T \leq \frac{3\delta_0}{2}. \end{aligned} \tag{26}$$

From Algorithm 1, at time T, we have $\beta = \min\{\frac{2\sqrt{2}}{\hat{\mu}_{\tilde{a}}^T/c_{\tilde{a}} - \hat{\mu}_0^T}, \sqrt{\log T}\}$. By assuming $\log T \geq \frac{32}{\delta_0^2}$ and using Equation (26), we have the following bound at the end of round T

$$\beta^2 \geq \frac{32}{9\delta_0^2}.$$

This implies that if event e is true, then, $\beta^2 \geq \frac{32}{9\delta_0^2}$. Now, we bound $\mathbb{E}[\beta^2]$:

$$\mathbb{E}[\beta^2] \geq \frac{32}{9\delta_0^2}P(e) \geq \frac{32}{9\delta_0^2}\Big(1 - 2N(3+|\mathbf{V}|\mathcal{W})T^{-\frac{p\delta_0^2}{16|\mathbf{V}|}}\Big).$$

Since $T^{\frac{p\delta_0^2}{16|\mathbf{V}|}} \geq \frac{8N(3+|\mathbf{V}|\mathcal{W})}{3}$, then,

$$\mathbb{E}[\beta^2] \geq \frac{8}{9\delta_0^2}.$$

■

We are now ready to prove Theorem 5.

Theorem 5 *The expected cumulative regret of Algorithm 1 is bounded by*

$$\delta_0\Big(\frac{8\ln B}{\delta_0^2} + 1 + \frac{\pi^2}{3}\Big) + \sum_{\delta_{i,x}>0} \delta_{i,x}\Big(\frac{8\ln B}{\delta_{i,x}^2} + 2 - \frac{8p_{i,x}}{18\delta_0^2|\mathbf{V}|}\ln b_{i,x}\cdot\tau_{i,x,b} + \frac{\pi^2}{3}\Big),$$

where $b_{i,x} := \frac{8}{\delta_{i,x}^2}\ln(\frac{B}{\max_a c_a}) + 1$, $\tau_{i,x,b} := \max\big\{0, 1 - |\mathbf{V}|\cdot\mathcal{W}_i\cdot b_{i,x}^{-p_{i,x}^2/(2|\mathbf{V}|)}\big\}$, *and* $\mathcal{W}_i$ *denotes the alphabet size of variables in* $\mathbf{V}\setminus\{X_i, Y\}$.

Algorithm 3 Compute $\hat{\mu}_{i,x}$ using observational samples

Input: ADMG $\mathcal{G}$, observational samples, indices $i \in [N]$ and $x \in \{0,1\}$

1: Reduce ADMG $\mathcal{G}$ to $\mathcal{H}_i$ using Algorithm 4;
2: Compute $\hat{D}_{i,x}$ by Algorithm 5 (use $\mathcal{H}_i$ as input);
3: Use Equation (28) to compute $\hat{\mu}_{i,x}$;
4: **Return:** $\hat{\mu}_{i,x}$.

Proof Let T_B denote the number of rounds that Algorithm 1 plays the arms before the budget B is exhausted. Therefore, we know $\frac{B}{\max_a c_a} \leq T_B \leq B$. Using Lemma 19, we get the following for T_B satisfying Lemma 22:

$$\begin{aligned}
\mathbb{E}[R_c(B)] &\leq \delta_0\Big(\frac{8\ln T_B}{\delta_0^2}+1+\frac{\pi^2}{3}\Big)+\sum_{\delta_{i,x}>0}\delta_{i,x}\Big(\frac{8\ln T_B}{\delta_{i,x}^2}+2-\frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[N_0^l]\tau_{i,x,l}+\frac{\pi^2}{3}\Big)\\
&\overset{(a)}{\leq} \delta_0\Big(\frac{8\ln T_B}{\delta_0^2}+1+\frac{\pi^2}{3}\Big)+\sum_{\delta_{i,x}>0}\delta_{i,x}\Big(\frac{8\ln T_B}{\delta_{i,x}^2}+2-\frac{p_{i,x}}{2|\mathbf{V}|}\mathbb{E}[\beta^2]\cdot\ln l\cdot\tau_{i,x,l}+\frac{\pi^2}{3}\Big)\\
&\overset{(b)}{\leq} \delta_0\Big(\frac{8\ln T_B}{\delta_0^2}+1+\frac{\pi^2}{3}\Big)+\sum_{\delta_{i,x}>0}\delta_{i,x}\Big(\frac{8\ln T_B}{\delta_{i,x}^2}+2-\frac{8p_{i,x}}{18\delta_0^2|\mathbf{V}|}\ln l\cdot\tau_{i,x,l}+\frac{\pi^2}{3}\Big)\\
&\leq \delta_0\Big(\frac{8\ln B}{\delta_0^2}+1+\frac{\pi^2}{3}\Big)+\sum_{\delta_{i,x}>0}\delta_{i,x}\Big(\frac{8\ln B}{\delta_{i,x}^2}+2-\frac{8p_{i,x}}{18\delta_0^2|\mathbf{V}|}\ln b\cdot\tau_{i,x,b}+\frac{\pi^2}{3}\Big),
\end{aligned} \tag{27}$$

where $l = \frac{8\ln T_B}{\delta_{i,x}^2}+1$, $b = \frac{8}{\delta_{i,x}^2}\ln(\frac{B}{\max_a c_a})+1$, $\tau_{i,x,l} = \max\{0, 1-|\mathbf{V}|\mathcal{W}_i l^{-\frac{p_{i,x}^2}{2|\mathbf{V}|}}\}$, $\tau_{i,x,b} = \max\{0, 1-|\mathbf{V}|\mathcal{W}_i b^{-\frac{p_{i,x}^2}{2|\mathbf{V}|}}\}$. Furthermore, the inequality in (a) follows from the fact $\mathbb{E}[N_0^l] \geq \mathbb{E}[\beta^2]\ln l$ and the inequality in (b) follows from Lemma 22. The last inequality holds since $\frac{B}{\max_a c_a} \leq T_B \leq B$. ■

Appendix C. Estimating the expected reward from observational distribution

In this section, we use a procedure (proposed by Bhattacharyya et al. (2020); Maiti et al. (2022)) to compute $\hat{\mu}_{i,x}$ for each $a_{i,x}$ using the observational data obtained by pulling arm a_0. Algorithm 3 summarizes the steps of this procedure.

Algorithm 3 takes the underlying causal graph $\mathcal{G}$, observational data that were collected by pulling arm a_0 for the first $\frac{B}{2}$ rounds and indices i, x as inputs. In line 1, it reduces $\mathcal{G}$ to $\mathcal{H}_i$ using Algorithm 4. Then, it computes the Bayes-net $\hat{D}_{i,x}$ in $\mathcal{H}_i$ using Algorithm 5 proposed by Bhattacharyya et al. (2020) in line 2. Finally, using $\hat{D}_{i,x}$, it computes $\hat{\mu}_{i,x}$ by substituting the distribution $P_{D_{i,x}}(Y=1, \mathbf{V}'=\mathbf{v}')$ in the following equation with its empirical estimation,

$$\mu_{i,x} = P_{\mathcal{G}}(Y=1|do(X_i)=x) = P_{\mathcal{H}_i}(Y=1|do(X_i)=x) = \sum_{\mathbf{v}'} P_{D_{i,x}}(Y=1, \mathbf{V}'=\mathbf{v}'), \tag{28}$$

where $\mathbf{V}'$ is the set of variables in $\mathcal{H}_i$ except $\{X_i, Y\}$ and $\mathbf{v}'$ is an arbitrary realization of $\mathbf{V}'$.

Algorithm 4 Reducing $\mathcal{G}$ to $\mathcal{H}_i$

Input: Causal garph $\mathcal{G}$ and index $i \in [N]$.

1: Let $\mathbf{W}_i = X_i \cup \widetilde{\mathbf{Pa}}(X_i) \cup Y$

2: Let $\mathcal{G}_i$ be the graph obtained from $\mathcal{G}$ by considering $\mathbf{V} \setminus \mathbf{W}_i$ as hidden variables.

3: Using **Projection Algorithm** (Tian and Pearl, 2002a; Verma and Pearl, 1990) do the following steps:

(a) For each observable variable $V_j \in \mathbf{V}$ in $\mathcal{G}_i$, add an observable variable V_j in $\mathcal{H}_i$.

(b) For each pair of observable variables $V_j, V_k \in \mathbf{V}$ in $\mathcal{G}_i$, add a directed edge from V_j to V_k in $\mathcal{H}_i$ if one of the followings hold:

1) There exists a directed edge from V_j to V_k in $\mathcal{G}_i$, or

2) There exists a directed path from V_j to V_k in $\mathcal{G}_i$ such that it contains only unobservable variables.

(c) For each pair of observable variables $V_j, V_k \in \mathbf{V}$ in $\mathcal{G}_i$, add a bidirected edge between V_j and V_k in $\mathcal{H}_i$ if there exists an unobservable variable U in $\mathcal{G}_i$ such that there exist two paths from U to V_j and from U to V_k in $\mathcal{G}'_i$ such that both paths contain only unobservable variables.

4: **Return**: $\mathcal{H}_i$

Appendix D. Proofs of Section 4

Theorem 7 *The expected simple regret of Algorithm 2 is bounded by* $\mathcal{O}\Big(\sqrt{\frac{n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}}\Big)$.

Proof Recall

$$q_{i,x}(\mathbf{z}) = P(X_i = x, \widetilde{\mathbf{Pa}}(X_i) = \mathbf{z}),$$
$$q_{i,x} = \min_{\mathbf{z}} q_{i,x}(\mathbf{z}),$$
$$q = \min\{q_{i,x} | q_{i,x} > 0 : i \in [N], x \in \{0,1\}\}.$$

When $q_{i,x} = 0$ for all i, x, we define $q = \frac{1}{N+1}$.

For each $X_i \in \mathbf{X}$, let k_i be the size of the c-component containing X_i, and $k = \min_i k_i$. Moreover, let $\mathcal{Z}_i$ be the size of the domain from which $\widetilde{\mathbf{Pa}}(X_i)$ takes its values and let $\mathcal{Z} := \max_i \mathcal{Z}_i$. Next, we prove a lemma that is useful in the proof of Theorem 7.

Lemma 23 *For every $i \in [N]$, define $f_{i,x}(\mathbf{z})$ to be one, if $|\hat{q}_{i,x}(\mathbf{z}) - q_{i,x}(\mathbf{z})| \geq \frac{1}{4}(1 - 2^{-1/k})q$ at the end of $\frac{B}{2}$ rounds. Let $f = 1$, if there exists $i \in [N], x \in \{0,1\}$, and $\mathbf{z}$ in the domain of $\widetilde{\mathbf{Pa}}(X_i)$ and $f_{i,x}(\mathbf{z}) = 1$. Then, the following statements hold*

(a) $P(f = 1) \leq 4N\mathcal{Z}\exp\big(-\frac{1}{16}(1 - 2^{-1/k})^2 q^2 B\big)$.

(b) If $f = 0$, therefore, $n(\hat{\mathbf{q}}) \leq 2n(\mathbf{q})$ holds at the end of $\frac{B}{2}$ rounds.

Algorithm 5 Computing $\hat{D}_{i,x}$

Input: Observational samples, i, x, $\mathcal{H}_i$ (with set of vertices $\mathbf{V}_i$) and parameter t.

for every variable $V_j \in \mathbf{C}_i$ **do**
  for every realization $V_j = v$ and $\mathbf{Z}_j = \mathbf{z}$, where $\mathbf{Z}_j$ is the set of effective parents of V_j in $\mathcal{H}_i$ **do**
    $N_{\mathbf{z}} \leftarrow$ the number of samples that $\mathbf{Z}_j = \mathbf{z}$,
    $N_{\mathbf{z},v}$ the number of samples that $\mathbf{Z}_j = \mathbf{z}$ and $V_j = v$,
    $\hat{D}_{i,x}(V_j = v|\mathbf{Z}_j = \mathbf{z}) \leftarrow \frac{N_{\mathbf{z},v}+1}{N_{\mathbf{z}+2}}$,
for every variable $V_j \in \mathbf{V}_i \setminus \mathbf{C}_i$ **do**
  for every $V_j = v$ and $\mathbf{Z}_j \setminus X_i = \mathbf{z}$, **do**
    if $X_i \in \mathbf{Z}_j$ **then**
      $N_{\mathbf{z}} \leftarrow$ the number of samples that $\mathbf{Z}_j \setminus X_i = \mathbf{z}$ and $X_i = x$,
      $N_{\mathbf{z},v} \leftarrow$ the number of samples that $\mathbf{Z}_j \setminus X_i = \mathbf{z}$, $X_i = x$ and $V_j = v$,
      if $N_{\mathbf{z}} \geq t$ **then**
        $\hat{D}_{i,x}(V_j = v|\mathbf{Z}_j = \mathbf{z}) \leftarrow \frac{N_{\mathbf{z},v}+1}{N_{\mathbf{z}+2}}$,
      else
        $\hat{D}_{i,x}(V_j = v|\mathbf{Z}_j \setminus \{X_i\} = \mathbf{z}, X_i = x) = \frac{1}{2}$,
    else
      $N_{\mathbf{z}} \leftarrow$ the number of samples that $\mathbf{Z}_j = \mathbf{z}$,
      $N_{\mathbf{z},v} \leftarrow$ the number of samples that $\mathbf{Z}_j = \mathbf{z}$ and $V_j = v$,
      if $N_{\mathbf{z}} \geq t$ **then**
        $\hat{D}_{i,x}(V_j|\mathbf{Z}_j = \mathbf{z}) \leftarrow \frac{N_{\mathbf{z},v}+1}{N_{\mathbf{z}+2}}$,
      else
        $\hat{D}_{i,x}(V_j|\mathbf{Z}_j = \mathbf{z}) \leftarrow \frac{1}{2}$,
Return$\hat{D}_{i,x}$.

Proof (a) Using Lemma 15, we get

$$P(f_{i,x}(\mathbf{z}) = 1) \leq 2\exp\big(-\frac{1}{16}(1-2^{-1/k})^2 q^2 B\big).$$

Now, define $f_{i,x}$ to be one, if there exists $\mathbf{z}$ in domain of $\widetilde{\mathbf{Pa}}(X_i)$ and $f_{i,x}(\mathbf{z}) = 1$. Using union bound, we get

$$P(f_{i,x} = 1) \leq 2\mathcal{Z}_i \exp\big(-\frac{1}{16}(1-2^{-1/k})^2 q^2 B\big).$$

Next, let $f_i = 1$ if there exists $x \in \{0,1\}$ and $f_{i,x} = 1$. Then,

$$P(f_i = 1) \leq 4\mathcal{Z}_i \exp\big(-\frac{1}{16}(1-2^{-1/k})^2 q^2 B\big)$$

Applying the union bound on the above equation, we get

$$P(f = 1) \leq 4N\mathcal{Z} \exp\big(-\frac{1}{16}(1-2^{-1/k})^2 q^2 B\big).$$

(b) First, we sort all $q_{i,x}^{k_i}$ for $i \in [N]$ and $x \in \{0,1\}$ in an ascending order. Without loss of generality, assume the sorted sequence is $q_1^{k_1} \leq q_2^{k_2} \leq \cdots \leq q_{2N}^{k_{2N}}$. Define $g_1 := \max\{i | q_i^{k_i} < \frac{1}{n(\mathbf{q})}\}$. The definition of $n(\mathbf{q})$ implies that $g_1 \leq n(\mathbf{q})$ and $q_i^{k_i} \geq \frac{1}{n(\mathbf{q})}$ for every $i > g_1$. Therefore, by assuming $|\hat{q}_{i,x}(\mathbf{z}) - q_{i,x}(\mathbf{z})| \leq \frac{1}{4}(1-2^{-1/k})q$, we get the following for every $i > g_1$:

$$(\hat{q}_i)^{k_i} \geq \big(q_i - \frac{1}{4}(1-2^{-1/k})q\big)^{k_i} \overset{\text{(a)}}{\geq} \big(q_i - \frac{1}{4}(1-2^{-1/k})q_i\big)^{k_i} \geq \frac{2^{-\frac{k_i}{k}}}{n(\mathbf{q})} \geq \frac{1}{2n(\mathbf{q})},$$

where the inequality in (a) holds since $q_i \geq q$. Hence,

$$\sum_{i\in[N]} c_i \mathbb{1}\{\hat{q}_i^{k_i} < \frac{1}{2n(\mathbf{q})}\} < \sum_{i\in[g_1]} c_i = \sum_{i\in[N]} c_i \mathbb{1}\{\hat{q}_i^{k_i} < \frac{1}{n(\mathbf{q})}\} \leq n(\mathbf{q}) \leq 2n(\mathbf{q}).$$

The above equation implies that the following inequality holds for $\tau = 2n(\mathbf{q})$,

$$\sum_{i\in[N]} c_i \mathbb{1}\{\hat{q}_i^{k_i} < \frac{1}{\tau}\} \leq \tau.$$

Then, by the definition of $n(\mathbf{q})$, we get

$$n(\hat{\mathbf{q}}) \leq 2n(\mathbf{q}). \tag{29}$$

■

Lemma 24 *Let $a_{i,x} \in \mathcal{A}$ be an arbitrary arm and $\epsilon > 0$, then $P(|\hat{\mu}_{i,x} - \mu_{i,x}| > \epsilon) \leq \exp(-\epsilon^2 \frac{q_{i,x}^{k_i} B}{M})$ at the end of $\frac{B}{2}$ rounds, where $M \geq 1$ is a constant number which is independent of the distribution but dependent on the underlying graph.*

Proof Theorem 2.5 in (Bhattacharyya et al., 2020) implies that $\hat{\mu}_{i,x}$ can be estimated with probability $1-\delta_i$ such that $|\hat{\mu}_{i,x} - \mu_{i,x}| \leq \epsilon$ using $\mathcal{O}(\frac{2^{u_i}}{q_{i,x}^{k_i}\epsilon^2} \log 2^{u_i} \log \frac{1}{\delta_i})$ number of samples, where $u_i = 2(1+k_i(d+1))^2$. Therefore, using $B = K\frac{2^{2u_i}}{q_{i,x}^{k_i}\epsilon^2} \log \frac{1}{\delta_i}$ samples, where K is a constant, we achieve

$$P\big(|\hat{\mu}_{i,x} - \mu_{i,x}| \leq \epsilon\big) \geq 1 - \delta_i.$$

Next, we re-write δ_i in terms of ϵ and B and get

$$P(|\hat{\mu}_{i,x} - \mu_{i,x}| > \epsilon) \leq \exp\Big(\epsilon^2 \frac{B q_{i,x}^{k_i}}{K 2^{2u_i}}\Big) \leq \Big(\epsilon^2 \frac{B q_{i,x}^{k_i}}{M}\Big),$$

where $M = \max\{1, K2^{2u_i}\}$. ■

We are now ready to prove Theorem 7 using the aforementioned Lemmas. Let $M' := 2^{k-1}M$ and

$$B_1 := \min_b \Big\{\sqrt{\frac{4M'n(\mathbf{q})}{b} \log \frac{Nb}{n(\mathbf{q})}} \geq 6\frac{n(\mathbf{q})}{b}\Big\},$$

$$B_2 := \min_b \Big\{\sqrt{\frac{36M'n(\mathbf{q})}{b} \log \frac{Nb}{n(\mathbf{q})}} \geq 4N\mathcal{Z} \exp\big(-\frac{1}{16}(1-2^{-1/k})^2 q^2 b\big)\Big\},$$

and assume $B \geq \max\{B_1, B_2\}$.

For every $a_{i,x} \in \mathcal{A}'$, Algorithm 2 pulls each arm $\frac{B}{2\sum_{i,x} c_{i,x}\mathbb{1}\{a_{i,x}\in\mathcal{A}'\}}$ additionally to recompute $\hat{\mu}_{i,x}$. Therefore, by the definition of $n(\mathbf{q})$ and Lemma 23, we get the following if $f = 0$ (Please see Lemma 23 for the definition of the event f),

$$\frac{B}{2\sum_{i,x} c_{i,x}\mathbb{1}\{a_{i,x}\in\mathcal{A}'\}} \geq \frac{B}{2n(\hat{\mathbf{q}})} \geq \frac{B}{4n(\mathbf{q})}. \tag{30}$$

Then, by Lemma 15, we have the following equation for every $a_{i,x} \in \mathcal{A}'$:

$$P\Big(|\hat{\mu}_{i,x} - \mu_{i,x}| \geq \epsilon | f = 0\Big) \leq 2\exp\big(-\epsilon^2 \frac{B}{2n(\mathbf{q})}\big) \leq 2\exp\big(-\epsilon^2 \frac{B}{4M'n(\mathbf{q})}\big). \tag{31}$$

For each $a_{i,x} \notin \mathcal{A}'$, we know that $\hat{q}_{i,x}^{k_i} \geq \frac{1}{n(\hat{\mathbf{q}})}$. However, depending on $q_{i,x}^{k_i}$, the proof technique varies. Below, we present the proof under two different cases:

Case 1. If $a_{i,x} \notin \mathcal{A}'$ and $q_{i,x}^{k_i} < \frac{1}{n(\mathbf{q})}$. Conditioning on $f = 0$, we have :

$$\begin{aligned} q_{i,x}^{k_i} \geq (\hat{q}_{i,x} - \frac{1}{4}(1-2^{-1/k})q)^{k_i} &\geq \Big((\frac{1}{n(\hat{\mathbf{q}})})^{1/k_i} - \frac{1}{4}(\frac{1}{n(\mathbf{q})})^{1/k_i}\Big)^{k_i} \\ &\overset{(a)}{\geq} \Big((\frac{1}{2n(\mathbf{q})})^{1/k_i} - \frac{1}{4}(\frac{1}{n(\mathbf{q})})^{1/k_i}\Big)^{k_i} \\ &\geq \frac{1}{2^{k+1}n(\mathbf{q})}, \end{aligned}$$

where the inequality in (a) follows from Lemma 23. Using the above bound for $q_{i,x}^{k_i}$ and Lemma 24 yield

$$P(|\hat{\mu}_{i,x} - \mu_{i,x}| \geq \epsilon | f = 0) \leq \exp(-\epsilon^2 \frac{B}{2^{k+1}Mn(\mathbf{q})}) = \exp(-\epsilon^2 \frac{B}{4M'n(\mathbf{q})}). \tag{32}$$

Case 2. If $a_{i,x} \notin \mathcal{A}'$ and $q_{i,x}^{k_i} \geq \frac{1}{n(\mathbf{q})}$. From Lemma 24, we have

$$P\big(|\hat{\mu}_{i,x} - \mu_{i,x}| \geq \epsilon | f = 0\big) \leq \exp(-\epsilon^2 \frac{q_{i,x}^{k_i} B}{M}) \leq \exp(-\epsilon^2 \frac{B}{4M'n(\mathbf{q})}). \tag{33}$$

For $a = a_0$, Lemma 15 gives us

$$P(|\hat{\mu}_0 - \mu_0| \geq \epsilon) \leq 2\exp(-\epsilon^2 B) \leq 2\exp(-\epsilon^2 \frac{B}{4M'n(\mathbf{q})}). \tag{34}$$

Now, let e be an event where $e = 1$ if there exists an arm $a \in \mathcal{A}$, such that $|\hat{\mu}_a - \mu_a| \geq \epsilon$. We also define event e_a where $e_a = 1$ if $|\hat{\mu}_a - \mu_a| \geq \epsilon$. Equations (31), (32), (33) and (34) imply that for every action $a \in \mathcal{A}$,

$$P(e_a = 1 | f = 0) \leq 2\exp\big(-\epsilon^2 \frac{B}{4M'n(\mathbf{q})}\big).$$

Applying the union bound on the above equation implies

$$P(e=1|f=0) \leq (4N+2)\exp\big(-\epsilon^2\frac{B}{4M'n(\mathbf{q})}\big) \leq 6N\exp\big(-\epsilon^2\frac{B}{4M'n(\mathbf{q})}\big).$$

Using the above inequalities and substituting $\epsilon = \sqrt{\frac{4M'n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}}$ yield

$$\begin{aligned}
\mathbb{E}[R_s(B)|f=0] &= \mathbb{E}[R(B)|e=0]P(e=0) + \mathbb{E}[R(B)|e=1]P(e=1) \\
&\leq \mathbb{E}[R(B)|e=0] + P(e=1) \\
&\leq 2\epsilon + 6N\exp(-\epsilon^2\frac{B}{4M'n(\mathbf{q})}) \\
&= 2\sqrt{\frac{4M'n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}} + 6\frac{n(\mathbf{q})}{B} \\
&\leq \sqrt{\frac{36M'n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}}.
\end{aligned} \tag{35}$$

Finally, Lemma 23 and Equation (35) imply

$$\begin{aligned}
\mathbb{E}[R_s(B)] &= \mathbb{E}[R(B)|f=0]P(f=0) + \mathbb{E}[R(B)|f=1]P(f=1) \\
&\leq \mathbb{E}[R(B)|f=0] + P(f=1) \\
&\leq \sqrt{\frac{36M'n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}} + 4N\mathcal{Z}\exp\big(-\frac{1}{16}(1-2^{-1/k})^2(p')^2B\big) \\
&\leq 2\sqrt{\frac{36M'n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}}.
\end{aligned}$$

Therefore, $\mathbb{E}[R_s(B)] \in \mathcal{O}(\sqrt{\frac{n(\mathbf{q})}{B}\log\frac{NB}{n(\mathbf{q})}})$. ■

Appendix E. Discussion on simple regret bounds in no-backdoor graphs

Remark 25 *Consider the special case of no-backdoor graphs (causal graphs with no unblocked backdoor paths from intervenable variables to the reward variable Y). This graphical constraint ensures that for all $X \in \mathbf{X}$, $\mathbb{E}[Y|do(X=x)] = \mathbb{E}[Y|X=x]$. This is due to the second rule of do-calculus (Pearl, 1995). For causal MABs with no-backdoor graphs, $\mu_{i,x}$ can be estimated using observation as follows $\sum_{t=1}^{B/2} y^t \mathbb{1}\{x_i^t = x\}/\sum_{t=1}^{B/2}\mathbb{1}\{x_i^t=x\}$. When the interventions have non-uniform costs, redefining $\hat{q}_{i,x} = \frac{2}{B}\sum_{t=1}^{B/2}\mathbb{1}\{x_i^t=x\}$ yields drastically lower regrets. This special case and our improvements are discussed in Appendix E.*

Proof In general graphs, we estimate $q_{i,x} = \min_{\mathbf{z}} P(X_i = x, \widetilde{\mathbf{Pa}}(X_i) = \mathbf{z})$ by

$$\hat{q}_{i,x} = \frac{2}{B}\min_{\mathbf{z}}\Big\{\sum_{t=1}^{B/2}\mathbb{1}\big\{x_i^t = x, \widetilde{\mathbf{Pa}}^t(x_i) = \mathbf{z}\big\}\Big\}.$$

In no-backdoor graphs, we redefine $q_{i,x} = P(X_i = x)$ and $\hat{q}_{i,x} = \frac{2}{B}\sum_{t=1}^{B/2} \mathbb{1}\{x_i^t = x\}$. Through out this section, we denote the redefined $q_{i,x}$ and $\hat{q}_{i,x}$ by $q_{i,x}^{new}$ and $\hat{q}_{i,x}^{new}$.

Following the same procedure as in the proof of Theorem 7, it is straightforward to show that Algorithm 2, using $q_{i,x}^{new}$ and $\hat{q}_{i,x}^{new}$, achieves $R_s(B) \in \mathcal{O}\Big(\sqrt{\frac{n(\mathbf{q}^{new})}{B}\log\frac{NB}{n(\mathbf{q}^{new})}}\Big)$. Herein, we show that $n(\mathbf{q}^{new}) \leq n(\mathbf{q})$ which implies that in the no-backdoor setting, the new definitions guarantee better regret bound for Algorithm 2.

Note that for every $i \in [N]$, $x \in \{0,1\}$ and $\mathbf{z}$ in the domain of $\widetilde{\mathbf{Pa}}(X_i)$, we have

$$P(X_i = x, \widetilde{\mathbf{Pa}}(X_i) = \mathbf{z}) = P(\widetilde{\mathbf{Pa}}(X_i) = \mathbf{z}|X_i = x)P(X_i = x) \leq P(X_i = x).$$

Therefore, $q_{i,x} = \min_{\mathbf{z}} P(X_i = x, \widetilde{\mathbf{Pa}}(X_i) = \mathbf{z}) \leq P(X_i = x) = q_{i,x}^{new}$ for every i, x. From definition of $n(\mathbf{q})$ we get

$$\sum_{i,x} c_{i,x}\mathbb{1}\big\{q_{i,x} < (\frac{1}{n(\mathbf{q})})^{1/k_i}\big\} \leq n(\mathbf{q}),$$

and since $q_{i,x} \leq q_{i,x}^{new}$,

$$\sum_{i,x} c_{i,x}\mathbb{1}\big\{q_{i,x}^{new} < (\frac{1}{n(\mathbf{q})})^{1/k_i}\big\} \leq \sum_{i,x} c_{i,x}\mathbb{1}\big\{q_{i,x} < (\frac{1}{n(\mathbf{q})})^{1/k_i}\big\}.$$

Therefore, the following inequality holds for $\tau = n(\mathbf{q})$

$$\sum_{i,x} c_{i,x}\mathbb{1}\big\{q_{i,x}^{new} < (\frac{1}{\tau})^{1/k_i}\big\} \leq \tau.$$

On the other hand, from the definition, we know that $n(\mathbf{q}^{new})$ is the smallest τ which holds in the above equation. Therefore, $n(\mathbf{q}^{new}) \leq n(\mathbf{q})$. ■

Remark 26 *Nair et al. (2021) studies the causal MAB problem with no-backdoor graphs and an additional constraint on the costs that is $c_{i,x} = c > 1$ for all i and x and $c_0 = 1$. Note that this setting does not satisfy the non-budgeted assumption in Maiti et al. (2022). Moreover, their algorithm uses a different exploration set than $\mathcal{A}'$ that seems to result in both worse performance and theoretical bound. Specifically, the threshold for determining the infrequent arms in Nair et al. (2021) is given by $m'(\mathbf{q}) := \min\{\tau | \sum_{i,x} \mathbb{1}\{p_{i,x} < \frac{1}{\tau}\} \leq \tau\}$. As we show in Appendix E, in this setting, $n(\mathbf{q}) \leq cm'(\mathbf{q})$ for all $c > 1$ and $\mathbf{q}$. Nair et al. (2021) shows that the expected simple regret of their algorithm is bounded by $\mathcal{O}\Big(\sqrt{\frac{cm'(\mathbf{q})}{B}\log\frac{NB}{cm'(\mathbf{q})}}\Big)$. Given that $n(\mathbf{q}) \leq cm'(\mathbf{q})$ for all $c > 1$, even in the special setting of Nair et al. (2021), our algorithm achieves better expected simple regret. This is also shown empirically in our experiment in Section 6.2.*

Proof Recall $p_{i,x} = P(X_i = x)$. Similar to Remark 6, we denote $p_{i,x}$ by $q_{i,x}^{new}$. By assuming $c_{i,x} = c$ for all $i \in [N]$ and $x \in \{0,1\}$, and all variables are observable in the underlying causal graph, we aim to show that $n(\mathbf{q}^{new}) \leq m'(\mathbf{q}^{new})$. Note that since all the variables are observable,

the size of the c-component for every X_i is equal to one, i.e., $k_i = 1$. By the definition of $m'(\mathbf{q}^{new})$, we have

$$\sum_{i,x} \mathbb{1}\{q^{new}_{i,x} < \frac{1}{m'(\mathbf{q}^{new})}\} \leq m'(\mathbf{q}^{new}).$$

Moreover, we have

$$\sum_{i,x} \mathbb{1}\{p_{i,x} < \frac{1}{cm'(\mathbf{q}^{new})}\} \leq \sum_{i,x} \mathbb{1}\{q^{new}_{i,x} < \frac{1}{m'(\mathbf{q}^{new})}\}.$$

Using the above equations, we can write $c\sum_{i,x} \mathbb{1}\{q^{new}_{i,x} < \frac{1}{cm'(\mathbf{q}^{new})}\} \leq c.m'(\mathbf{q}^{new})$. Finally, using the definition of $n(\mathbf{q}^{new})$ implies that $n(\mathbf{q}^{new}) \leq c.m'(\mathbf{q}^{new})$. ■

Appendix F. Cumulative Regret Lower Bound:

We need the following technical lemmas.

Lemma 27 (Bretagnolle–Huber inequality) *Let P and Q be probability measures on the same measurable space, and let A be an arbitrary event. Then*

$$P(A) + Q(A^c) \geq \frac{1}{2}e^{-D(P||Q)},$$

where D and A^c denote the KL-divergence and the complement of A, respectively.

Lemma 28 (Lattimore and Szepesvári (2020), Lemma 15.1) *Let $v = (P_1, ..., P_k)$ be the reward distributions associated with one k-armed bandit, and let $v' = (P'_1, ..., P'_k)$ be the reward distributions associated with another k-armed bandit. Choose a particular policy π and let P_π and P'_π be the probability measures on the canonical bandit model induced by the n-round interconnection of π and v and v', respectively. Then,*

$$D(P_\pi||P'_\pi) = \sum_{i=1}^{k} \mathbb{E}_{\pi,v}[N_i(n)]D(P_i||p'_i).$$

Lemma 29 *Let $\{P_\theta : \theta \in \mathbb{R}\}$ be a parametric family of distributions such that P_θ has mean θ. Suppose that the densities are twice continuously differentiable. Then, there exists x_0 such that $|x_0| \leq |\delta|$ and*

$$D(P_\theta||P_{\theta+\delta}) = \frac{I(\theta + x_0)}{2}\delta^2,$$

where $I(x)$ denotes the Fisher information of the family P_θ at x.

Proof Using a Taylor expansion of $h(x) := D(P_\theta||P_{\theta+x})$ around $x = 0$, we get

$$h(x) = h(0) + h'(0)x + h''(x_0)\frac{x^2}{2},$$

where $|x_0| \leq |x|$. We have $h(0) = 0$, $h''(x_0) = I(\theta + x_0)$, and

$$h'(0) = \frac{\partial}{\partial x} D(P_\theta || P_{\theta+x})\Big|_{x=0} = \int \frac{\partial}{\partial x} \log \frac{dP_\theta}{dP_{\theta+x}}\Big|_{x=0} dP_\theta = -\int \frac{\partial}{\partial x} \log \frac{dP_{\theta+x}}{dP_\theta}\Big|_{x=0} dP_\theta$$
$$= -\int \frac{\frac{\partial}{\partial x}\left(\frac{dP_{\theta+x}}{dP_\theta}\right)}{\frac{dP_{\theta+x}}{dP_\theta}}\Big|_{x=0} dP_\theta = -\int \frac{\partial}{\partial x}\left(\frac{dP_{\theta+x}}{dP_\theta}\right)\Big|_{x=0} dP_\theta = -\frac{\partial}{\partial x}\int \left(\frac{dP_{\theta+x}}{dP_\theta}\right) dP_\theta\Big|_{x=0}$$
$$= -\frac{\partial}{\partial x}\int dP_{\theta+x}\Big|_{x=0} = -\frac{\partial}{\partial x} 1\Big|_{x=0} = 0.$$

■

To establish the result, we use an analogous argument as in the classical multi-arm bandits and how any algorithm suffers $\Omega(\sqrt{\lfloor B/c \rfloor KN})$ regret on a specific causal graph $\mathcal{G}_0$ depicted in Figure 5 with predefined reward distributions P_1 and P_2 and the uniform costs $\mathcal{C}_0 := \{c_{i,x} = c\}$.

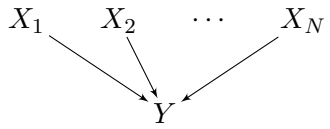

Figure 5: The ADMAG $\mathcal{G}_0$ over $\mathbf{V} = \{X_1, ..., X_N, Y\}$.

In this causal bandits setting, we assume that all variables can take values in $[K] = \{1, ..., K\}$ and consider two different distributions over the reward variable Y belonging to a parametric family of distributions that have twice differentiable density functions, e.g., Gaussian distributions. Distribution P_1 is selected such that $\mathbb{E}_1[Y|do(X_i = x)] = b$ for constant b, for all $i \in [N]$, and all $x \in [K]$ and $\mathbb{E}_1[Y] = b + \sqrt{\frac{KN}{w\lfloor B/c \rfloor}}$ for constant w to be defined later. Hence, the best action to play in this setting would be a_0, i.e., no intervention.

Let A_B be an arbitrary adaptive algorithm that selects the actions possibly based on its previous interactions with the problem and the total budget B and let $P_{A,1}$ be the resulting distribution of applying this algorithm when the rewards are distributed according to P_1. In order to design the second distribution, we select the least played action by A_B and assign a higher expected reward to it. To be more precise, let

$$a_{i^*,x^*} := \arg\min_{a_{i,x}} \mathbb{E}_{A,1}[N^B_{a_{i,x}}],$$

where $N^B_{a_{i,x}}$ denotes the number of times that arm $a_{i,x}$ is played by the algorithm using all its budget B (since all arms have the same cost c, that is equivalent to playing the arms over a time horizon $\lfloor B/c \rfloor$) and the expectation is taken with respect to $P_{A,1}$. Note that $\sum_{a_{i,x} \in \mathcal{A}} N^B_{a_{i,x}} = \lfloor B/c \rfloor$, the total number of actions are $|\mathcal{A}| = NK + 1$, and thus

$$\mathbb{E}_{A,1}[N^B_{a_{i^*,x^*}}] \leq \frac{\lfloor B/c \rfloor}{NK}. \tag{36}$$

Now, we can design the second distribution P_2 that is identical to P_1 except at index a_{i^*,x^*} and at this index it has $\mathbb{E}_2[Y|do(X_{i^*} = x^*)] = b + 2\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}$. Therefore, the optimal arm under this

distribution is a_{i^*,x^*}. We have

$$R_c(A_B, \mathcal{G}_0, P_1, \mathcal{C}_0) = \sum_{a_{i,x} \in \mathcal{A}} \mathbb{E}_{A,1}[N^B_{a_{i,x}}]\delta_{i,x} = \mathbb{E}_{A,1}[\lfloor B/c \rfloor - N^B_{a_0}]\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}$$
$$\geq P_{A,1}(N^B_{a_0} \leq \lfloor B/c \rfloor/2)\frac{\lfloor B/c \rfloor}{2}\sqrt{\frac{KN}{w\lfloor B/c \rfloor}},$$
$$R_c(A_B, \mathcal{G}_0, P_2, \mathcal{C}_0) = \sum_{a_{i,x} \in \mathcal{A}} \mathbb{E}_{A,2}[N^B_{a_{i,x}}]\delta_{i,x} \geq \mathbb{E}_{A,2}[N^B_{a_0}]\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}$$
$$\geq P_{A,2}(N^B_{a_0} > \lfloor B/c \rfloor/2)\frac{\lfloor B/c \rfloor}{2}\sqrt{\frac{KN}{w\lfloor B/c \rfloor}},$$

where $R_c(A_B, \mathcal{G}, P, \mathcal{C})$ denotes the cumulative regret of algorithm A on a causal graph $\mathcal{G}$ with the distribution P over the rewards and the cost set $\mathcal{C}$. Combining the above inequalities and using the Bretagnolle–Huber inequality (Lattimore and Szepesvári, 2020), we have

$$2\max\{R_c(A_B, \mathcal{G}_0, P_1, \mathcal{C}_0), R_c(A_B, \mathcal{G}_0, P_2, \mathcal{C}_0)\} \geq R_c(A_B, \mathcal{G}_0, P_1, \mathcal{C}_0) + R_c(A_B, \mathcal{G}_0, P_2, \mathcal{C}_0)$$
$$\geq \left(P_{A,1}(N^B_{a_0} \leq \lfloor B/c \rfloor/2) + P_{A,2}(N^B_{a_0} > \lfloor B/c \rfloor/2)\right)\frac{\lfloor B/c \rfloor}{2}\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}$$
$$\geq \frac{\lfloor B/c \rfloor}{4}\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}e^{-D(P_{A,1}||P_{A,2})}.$$

From the definition of P_1 and P_2 and using Lemma 28, we get

$$D(P_{A,1}||P_{A,2}) = \mathbb{E}_{A,1}[N^B_{a_{i^*,x^*}}]D(P_1(a_{i^*,x^*})||P_2(a_{i^*,x^*})).$$

Using (36) and Lemma 29, we get

$$D(P_{A,1}||P_{A,2}) \leq \frac{\lfloor B/c \rfloor}{NK}\frac{I(b+\epsilon)}{2}\left(2\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}\right)^2 = \frac{2I(b+\epsilon)}{w},$$

where $\epsilon \leq 2\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}$. By selecting w large enough, we can ensure that $2I(b+\epsilon) < w$. Note that $I(b+\epsilon)$ is a constant and depends on the family distribution. For instance, for the family of Gaussian distributions with unit variance and mean θ, we have $I(b+\epsilon) = 1$. Combining the above inequalities leads to

$$\max_{\mathcal{G}_N, P, \mathcal{C}} R_c(A_B, \mathcal{G}_N, P, \mathcal{C}) \geq \max\{R_c(A_B, \mathcal{G}_0, P_1, \mathcal{C}_0), R_c(A_B, \mathcal{G}_0, P_2, \mathcal{C}_0)\} \geq \frac{\lfloor B/c \rfloor}{8e}\sqrt{\frac{KN}{w\lfloor B/c \rfloor}}.$$

This inequality holds for any arbitrary adaptive algorithm A_B.

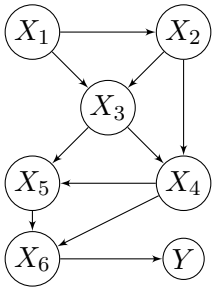

Figure 6: Causal graph of experiments in Section 6.1.

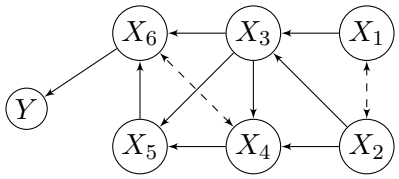

Figure 7: Causal graph of experiments in Figure 8.

Appendix G. Additional Experiments

G.1. Additional Experiments on Cumulative Regret in General Graphs

Figure 6 illustrates the underlying causal graph of the experiment in Section 6.1, which we used to compare the performance of Algorithm 1 with *CRM* and *F-KUBE*.

As mentioned before, since *CRM* is not designed for graphs with hidden variables, we did not include *CRM* for comparison in graphs with hidden variables.

Herein, we compare Algorithm 1 with *F-KUBE* when the model is constructed as explained in Section 6.1. The underlying graphs are demonstrated in Figure 7, which has 2 hidden variables. The left plot in Figure 8 illustrates the performance of algorithms in terms of cumulative regret versus budget when the cost of pulling each arm was selected randomly from $\{2, 3\}$. By increasing the budget, the cumulative regret of all of the algorithms increases. Although, our algorithm has a lower growth rate than *F-KUBE* and *CRM*.

Moreover, the right plot compares the performance of the algorithms when the budget is fixed to 1000, and the cost of all interventional arms is equal to c, such that $c \in \{2, 3, \ldots, 20\}$ and the cost of the observational arm is equal to 1. As shown in this Figure, the cumulative regret of Algorithm 1 grows substantially slower than others.

G.2. Additional Experiments on Simple Regret in No-backdoor Graphs

Since γ-NB is designed for settings with uniform costs over the arms, to have a fair comparison between Algorithm 2 and γ-NB, we included an extra experiment in this section. In this experiment, the cost of pulling each interventional arm $a_{i,x}$ for $i \in [N], x \in \{0, 1\}$ is set to be 4, and the cost of the observational arm is 1. The other setting of this experiment is similar to the one in Section 6.2.

Figure 9 illustrates the result of this experiment and it shows that when the cost of pulling interventional arms is uniform, by increasing the budget, our proposed algorithm converges quicker to zero than the others.

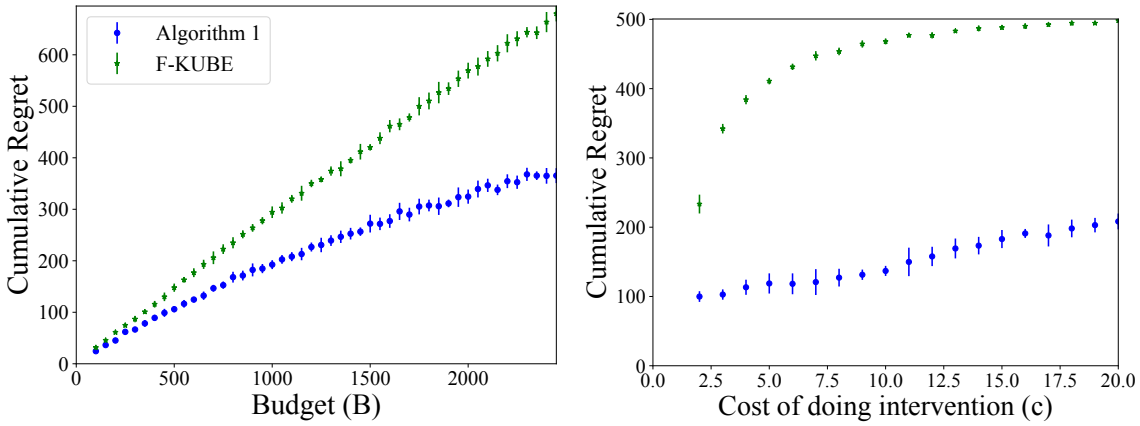

Figure 8: Performances of different algorithms on a general graphs with $N = 6$ depicted in Figure 7.

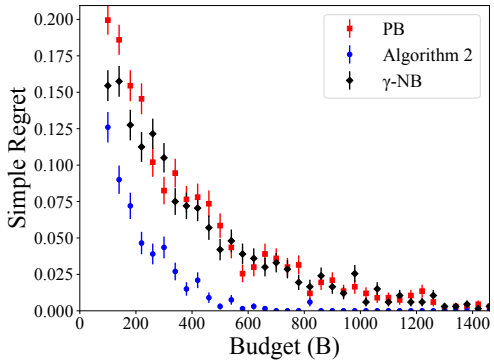

Figure 9: Performance of different algorithms on parallel graphs with $N = 50$.

We also included an experiment on a smaller parallel graph (with 7 intervenable variables) to be able to compare our algorithm with *Successive Rejects*. To construct the underlying model, we used the same setting as in Section 6.2. Figure 10 illustrates the performance of different algorithms in terms of their simple regret. For simple regret vs. budget, we set the cost of each interventional arm randomly from $\{2, 3, 4, 5\}$. This figure shows that Algorithm 2 convergence is faster to zero than the other algorithms. For simple regret vs. the cost of doing an intervention, the budget is set to 1500, and the cost of all interventional arms is equal to $c \in \{1, 2, \ldots, 20\}$. This figure demonstrates that by increasing c, Algorithm 2 has a slower growth rate than others.

G.3. Additional Experiments on Simple Regret in General Graphs

Herein, we present the underlying causal graph of the experiments in Section 6.3. As shown, this graph is not a no-backdoor graph as it has unblocked backdoor paths from intervenable variables to the reward variable. Furthermore, it has two hidden confounders.

We also provided additional experiments on a different graph illustrated in Figure 12 with $N = 5$ and one hidden confounder. Note that this graph also includes unblocked backdoor paths from intervenable variables to the reward variable. We constructed the underlying model similar to the

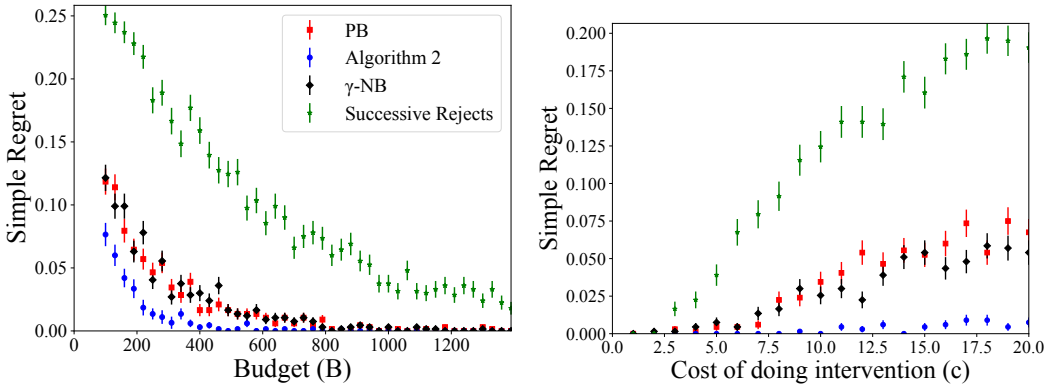

Figure 10: Performance of different algorithms on a parallel graph with $N = 7$.

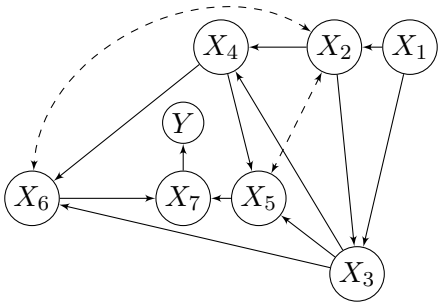

Figure 11: Causal graph of experiments in Section 6.3 with $N = 7$.

one in Section 6.1. The left plot in Figure 13 compares the performance of algorithms in terms of their simple regret when the cost of pulling each interventional arm was selected randomly from $\{5, 6, 7\}$. As depicted, by increasing the budget, Algorithm 2 converges to zero faster. The right plot in Figure 13 shows that when the cost of all interventional arms is equal to c, the simple regret increases by increasing c as expected, but our algorithm's regret grows at a lower rate.

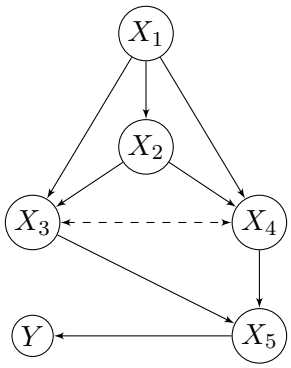

Figure 12: Causal graph of the additional experiments.

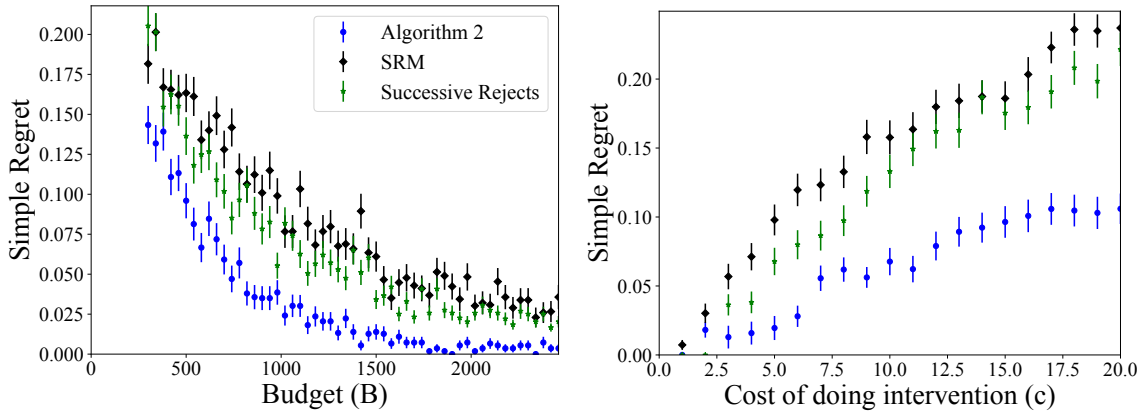

Figure 13: Performance of different algorithms on the general graphs depicted in Figure 12.

Appendix H. Discussion on the previous work

In (Nair et al., 2021), Eq. (17) in the proof of Lemma B.6 (crucial for Theorem 3 pertaining to cumulative regret bound) reads as follows

$$N_T^{i,x} \leq \max(0, l - \sum_{t\in[T]} \mathbb{1}\{a(t) = a_0, X_i = x\}) + \sum_{t\in T} \mathbb{1}\{a(t) = a_{i,x}, E_t^{i,x} \geq l\}$$

which is wrong. As a counterexample, suppose that $T = 3$ and the pulled arms are $\{a(0) = a_{i,x}, a(1) = a_{j,x'}, a(2) = a_0, a(3) = a_0\}$, where $j \neq i$ and observed $X_i = x$ at both times $t = 2, 3$. In this case, for $l = 2$, the inequality becomes $1 \leq 0$.