Graph Learning Is Suboptimal in Causal Bandits

Mohammad Shahverdikondori
College of Management of Technology, EPFL

Jalal Etesami
Department of Computer Science, TU Munich

Negar Kiyavash
College of Management of Technology, EPFL

Abstract

We study regret minimization in causal bandits under causal sufficiency where the underlying causal structure is not known to the agent. Previous work has focused on identifying the reward's parents and then applying classic bandit methods to them, or jointly learning the parents while minimizing regret. We investigate whether such strategies are optimal. Somewhat counterintuitively, our results show that learning the parent set is suboptimal. We do so by proving that there exist instances where regret minimization and parent identification are fundamentally conflicting objectives. We further analyze both the known and unknown parent set size regimes, establish novel regret lower bounds that capture the combinatorial structure of the action space. Building on these insights, we propose nearly optimal algorithms that bypass graph and parent recovery, demonstrating that parent identification is indeed unnecessary for regret minimization. Experiments confirm that there exists a large performance gap between our method and existing baselines in various environments.

1 INTRODUCTION

Multi-armed bandit (MAB) settings are a fundamental framework for modeling sequential decision-making in stochastic environments [BCB+12, LS20], with applications ranging from recommendation systems [LCLS10, BR19, SEG25], clinical trials [VBW15, LSN+20], to A/B tests [Sco10]. In this framework, an agent repeatedly selects actions (arms), observes the resulting rewards, and aims to maximize the expected cumulative reward. In the classic MAB problem, arms are assumed to be independent. This limits its applicability in more structured environments. To remedy this limitation, various extensions have introduced dependencies among arms, including linear bandits [DHK08, AYPS11], bandits with graph feedback [ACBDK15, ACBG+17], bandits with interference [JFK24, AAMW24, JSK25], and causal bandits [LLR16, YHS+18, LB18].

In causal bandits, the focus of this paper, dependencies are captured by a causal graph over the variables, with one node designated as the reward. Actions correspond to interventions on subsets of variables, after which the agent observes the reward along with the values of non-intervened variables. Exploiting this causal structure and the additional observations has been shown to significantly accelerate learning [LLR16, LMTY20].

A common assumption in much of the causal bandit literature is knowledge of the full causal graph. Although progress has been made in developing scalable causal discovery methods [GZS19, VCB22, LAR22, SMK24, MEAK25], such methods typically require full knowledge of observational or interventional distributions and remain imperfect. Therefore, in practice it is more likely that the structure of the causal graph is unknown. Under causal sufficiency (i.e., no unobserved variables) and for agents that can intervene on subsets of nodes, it was shown that the optimal action is an intervention on all parents of the reward node [LB18]. Consequently, prior work has focused on identifying the parent set and then applying standard regret-minimization algorithms to it, or attempting to simultaneously learn the parents and minimize regret [LMT21, KEK25, PZM25, EGK24].

In this work, we ask whether identifying the parent set is indeed optimal for regret minimization. We show that, with no assumptions on the underlying generating model and for the worst-case regret minimization, even when the graph over non-reward variables is fully known, parent identification is sub-optimal.

It is important to note that our analysis does not impose any structural assumptions (e.g., linearity) on the

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

causal model and focuses on hard interventions, where the agent sets fixed values for a subset of variables. Our results are specific to this setting. In contrast, some related works consider other classes of interventions, such as those that modify conditional distributions [SSDS17, SFDvO25], and our conclusions may not directly extend to those scenarios.

Our main contributions are as follows:

- We demonstrate that parent identification and regret minimization can be fundamentally at odds. That is, there exist instances where the set of high-reward actions is disjoint from the set of informative actions to identify the parents.
- When the number of parents of the reward node is known, we establish a worst-case regret lower bound that reflects the combinatorial structure of the action space and holds even if the agent has complete knowledge about the graph over the non-reward variables. We further propose a simple algorithm that, under a mild assumption on the problem parameters, achieves regret matching this lower bound up to logarithmic factors, without requiring any prior knowledge.
- When the number of parents is unknown, we prove a lower bound, showing that no algorithm can attain the same regret rates as in the known case uniformly over all parent set sizes. In addition, we introduce an adaptive algorithm that adjusts to the unknown parent size. This algorithm is Pareto optimal, up to logarithmic factors, when the agent can intervene on all variables in each round, and is nearly Pareto optimal in more general settings.
- Our experiments in diverse environments show that our algorithms outperform the existing baselines and reduce regret by up to a factor of 20.

1.1 Related Work

Causal Bandits. The causal bandit problem was first introduced in [LLR16], where the authors considered simple regret minimization with access to causal background knowledge, showing that such knowledge can significantly improve learning efficiency. Since then, causal bandits have been studied under a variety of settings and assumptions. Examples include assuming access to the distribution of parents of the reward node under each intervention [LMTY20, BWR22, LAR, SSDS17], restricting the causal model to be linear [VSST23, YMVT24, YT24], incorporating action costs into a budget constraint [NPS21, JEK24], studying combinatorial causal bandits with binary generalized linear models (BGLM) [FC23, XC], and addressing best-arm identification [SARK25, FXC25]. Another line of work proposed offline approaches to reduce the action space by identifying so-called possibly optimal intervention sets (POMIS) before the learning process begins [LB18, LB19].

Causal bandits with an *unknown graph structure* have also been investigated. In [LMT21, KEK25], the authors proposed algorithms in the atomic intervention setting that first identify the parents of the reward node and later run a standard regret minimization algorithm on the identified set. These algorithms are inherently limited to atomic interventions. [LMT21] proved a lower bound that showed when the reward node has a single parent and only atomic interventions are allowed, without additional distributional assumptions, no algorithm can achieve regret better than that of standard bandit algorithms applied to the full action set. The combinatorial causal bandits with an unknown graph under the BGLM model was studied in [FXC25]. [MAC23] studied causal bandits with an unknown graph under an additive outcome assumption and general interventions, showing that the problem can be reduced to an additive combinatorial linear bandit with full-bandit feedback. For continuous models, [YT24] studied unknown-graph causal bandits under linear structural equations. More recently, [PZM25] considered unknown-graph causal bandits with soft interventions under a linear structural equation model and proposed a sub-graph learning UCB method that controls false negative graph errors. Similarly, [ZZ25] studied unknown-graph causal bandits under a Gaussian linear DAG, using backdoor adjustment to combine observational and experimental data in a UCB algorithm. Finally, [EGK24] demonstrated that, when the graph is unknown, partial causal discovery suffices to identify the set of POMISs.

In contrast to the aforementioned works, we do not assume any particular structure on the graph or the distribution. We allow interventions of size m, thus generalizing previous studies that only consider atomic or general interventions.

Adaptivity to Unknown Parameters. Since we study the case where the size of the reward's parent set is unknown and provide an algorithm that adapts to it, our work is also related to the broader literature on adaptivity to unknown parameters in bandit problems. The closest work is [ZN20], which considers bandits with multiple optimal arms where the number of optimal arms is unknown to the learner. Other examples include continuum-armed bandits [Agr95, LC18, Had19], where algorithms are designed to adapt to an unknown smoothness parameter.

2 PRELIMINARIES AND PROBLEM SETUP

In this section, we formally introduce the problem setup for the causal bandits problem with an unknown graph and fixed-sized interventions, adapting a terminology similar to prior work.

A causal graph $\mathcal{G}$ over n random variables $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$ is a directed acyclic graph (DAG), where an edge $X_i \to X_j$ indicates that X_i directly causes X_j, i.e., changes in the value of X_i while keeping all other variables fixed may alter the distribution of X_j. We assume causal sufficiency holds, i.e., there are no unobserved variables. Let $\mathrm{Pa}_{\mathcal{G}}(X_i)$ denote the set of parents of node X_i in $\mathcal{G}$. We assume that each variable in $\mathcal{G}$ takes values in $[\ell] \coloneqq \{1, \dots, \ell\}$ for some positive integer ℓ.[1]

The causal bandits problem is a sequential game between an agent and an environment. The environment selects a causal graph $\mathcal{G}$ on variables $X_1, \dots, X_n$ and a reward variable Y, together with the conditional dependencies consistent with the graph. Let $\mathrm{Pa}_Y \subseteq \mathcal{X}$ denote the set of parents of the reward node in the causal graph with cardinality $k \coloneqq |\mathrm{Pa}_Y|$. We assume that the agent does not know Pa_Y but may have knowledge of the graph $\mathcal{G}$ over $\mathcal{X}$ or the value of k.

Action: At each round, the agent selects a subset of variables of size at most m and fixes their values through intervention. Such a selection is called an *action* or playing an *arm*. We use these terms interchangeably. Let $P([n])$ denote the power set of $[n]$. Each action a is defined by a pair (p_a, s_a), where $p_a \in P([n])$ is a subset of $[n]$ indicating the indices of the intervened variables with $|p_a| \le m$, and $s_a \in [\ell]^{|p_a|}$ is the corresponding vector of their assigned values. Let $\mathcal{A}$ denote the set of all such possible actions. We also define $\mathcal{A}_i \coloneqq \{a \in \mathcal{A} \mid |p_a| = i\}$ as the set of actions with intervention size i. Then, $|\mathcal{A}_i| = \binom{n}{i}\ell^i$ and $|\mathcal{A}| = \sum_{i=0}^{m}\binom{n}{i}\ell^i$. We assume that p_a is sorted in increasing order and that s_a is aligned with this ordering: the i-th entry of s_a specifies the value assigned to the variable whose index is the i-th element of p_a. Formally, at round t, the agent chooses an action $a_t = (p_{a_t}, s_{a_t}) \in \mathcal{A}$. With a slight abuse of notation, for any set S and integer r, we denote by $\binom{S}{r}$, the family of all subsets of S with r elements.

Reward: After performing action a_t at round t, the agent observes a vector $\left(\mathbf{x}^{(t)}, y_t\right) = (x_{1,t}, \dots, x_{n,t}, y_t)$, which is a sample from the post-interventional distribution $\mathbb{P}\left(X_1, \dots, X_n, Y \mid do(\mathbf{X}_{p_{a_t}} = s_{a_t})\right)$. We denote the expected reward of this action by $\mu_{a_t} \coloneqq \mathbb{E}[Y \mid a_t] = \mathbb{E}[Y \mid do(\mathbf{X}_{p_{a_t}} = s_{a_t})]$. Let $a^* \in \arg\max_{a\in\mathcal{A}} \mu_a$ be an optimal action. We assume that $\mu_a \in [0, 1]$ for all $a \in \mathcal{A}$ and that the distribution of Y under each action is 1-sub-Gaussian, which is a common assumption in the bandit literature [BCB+12, LS20].

Environments: We denote by $\mathcal{E}(n, \ell, k)$ the class of environments with n variables, each taking values in $[\ell]$, a reward node with k parents and 1-sub-Gaussian reward distributions with means in $[0, 1]$. When the parameters are fixed, we write $\mathcal{E}$.

Regret: Agent's policy, denoted by π, represents a sequence of actions over time i.e., $\pi = (a_1, a_2, ...)$. Accordingly, the cumulative regret of a policy π in an environment $\mathcal{V} \in \mathcal{E}$ over T rounds is defined as

$$R_T(\pi, \mathcal{V}) \coloneqq T\mu_{a^*} - \sum_{t\in[T]} \mu_{a_t}.$$

This corresponds to the cumulative gap between the expected reward of the optimal action and that of the actions chosen by π.

Agent's objective: To encode the different assumptions about the agent's prior knowledge, we introduce an *information set* $\mathcal{I}$ that specifies what the agent knows in advance $\big($e.g., $\mathcal{I} = \emptyset$, $\{k\}$, $\{\mathcal{G}\}$, or $\{k, \mathcal{G}\}\big)$ which correspond to the agent having no side information, knowing the number of parents of the reward node, the causal graph over the random variables $\mathcal{X}$, or both, respectively). We define $\Pi(\mathcal{I})$ as the set of all policies that can utilize prior information $\mathcal{I}$. The agent's goal is to design a policy $\pi \in \Pi(\mathcal{I})$ that minimizes the worst-case cumulative regret over the set of possible instances defined by

$$R_T(\pi, \mathcal{E}) \coloneqq \max_{\mathcal{V}\in\mathcal{E}} R_T(\pi, \mathcal{V}).$$

Next lemma shows that there always exists an optimal action in $\mathcal{A}_m$. This allows us to design algorithms that only explore the actions in $\mathcal{A}_m$.

Lemma 2.1. *Under causal sufficiency, for any values of n, ℓ, k, m and any instance $\mathcal{V} \in \mathcal{E}$, there exists an optimal action with the maximum intervention size, that is* $\max_{a\in\mathcal{A}_m} \mu_a = \mu_{a^*}$.

3 GRAPH LEARNING MIGHT BE SUBOPTIMAL

Under causal sufficiency and for action sizes at least as large as the reward node's parent set ($m \ge k$), prior work [LMT21, KEK25] establishes that the optimal action corresponds to an intervention on the parent set. In the presence of unobserved variables, the parent set is replaced by the set of possibly optimal minimal intervention sets (POMIS) [LB18]. Thus, in causal bandits,

[1] The extension to finite domains with different cardinalities is straightforward.

if the causal graph is known, this knowledge can be used to reduce the exploration set. However, when the causal graph is unknown, the following fundamental question arises: *does regret minimization necessarily require identification of the parent set?*

In this section, we show that, under causal sufficiency and without any distributional assumptions, identifying the parent set is not necessarily optimal for minimizing cumulative regret. In fact, we shall see that there are instances where the two objectives, regret minimization and parent set identification, are at odds. Thus, in such cases, no algorithm that achieves the optimal regret rate can simultaneously identify the parent set with high probability. To formalize this, we define the probability of parent misidentification for a given algorithm.

Definition 3.1 (Parent-Identification)**.** In an environment $\mathcal{V} \in \mathcal{E}$, a parent identification algorithm uses the interventions generated by a policy $\pi \in \Pi(\mathcal{I})$ for T rounds along with a decision rule $\bar{d}$ to output an estimated parent set of the reward node, denoted by $\widehat{Pa}_T(\pi, \bar{d}, \mathcal{V}) \subseteq \mathcal{X}$. For a subclass $\mathcal{E}_0 \subseteq \mathcal{E}$ of instances, we define $\delta_T\left(\pi, \bar{d}, \mathcal{E}_0\right)$ as the maximum probability of misidentifying the true parent set after T rounds over instances in $\mathcal{E}_0$, i.e.,

$$\delta_T\left(\pi, \bar{d}, \mathcal{E}_0\right) \coloneqq \max_{\mathcal{V} \in \mathcal{E}_0} \mathbb{P}\left(\widehat{Pa}_T(\pi, \bar{d}, \mathcal{V}) \neq \mathrm{Pa}_Y(\mathcal{V})\right).$$

To demonstrate the trade-off between the probability of identifying a reward's parent set and regret, we introduce a subclass of instances $\mathcal{E}_0$, in which any parent identification algorithm with $m = k$ that achieves low error rates (in terms of T) necessarily incurs suboptimal regret. To show that $\mathcal{E}_0$ is not a collection of degenerate instances where, for example, identifying the parent set is impossible, we propose a simple uniform sampling algorithm and prove that it achieves good performance in the parent identification task over $\mathcal{E}_0$.

Uniform Sampling. Consider the case $m = k$. The uniform sampling policy π_{Unif}, on any instance $\mathcal{V}$, plays each action $a \in \mathcal{A}_m$ equally often, that is, $\frac{T}{\binom{n}{m} v^m}$ times.

Theorem 3.2 (Identification-Regret Trade-Off)**.** *There exists a subclass $\mathcal{E}_0 \subseteq \mathcal{E}(n, \ell, k)$ such that for $m = k$, the two following statements hold:*

1. *There exists a decision rule $\bar{d}_{Unif}$ that combined with the uniform sampling policy achieves $\delta_T\left(\pi_{Unif}, \bar{d}_{Unif}, \mathcal{E}_0\right) \in \mathcal{O}\left(\exp\left(-T\right)\right)$.*
2. *The regret of any policy π for which there exists a decision rule $\bar{d}$ such that $\delta_T\left(\pi, \bar{d}, \mathcal{E}_0\right) \in \mathcal{O}\left(\exp\left(-T^\alpha\right)\right)$ grows as $R_T\big(\pi, \mathcal{E}_0, \{k\}\big) \in \Omega\left(T^\alpha\right)$.*

Proof Sketch: The key step in the proof of this result is the construction of a class of instances in which the set of high-reward actions and the set of informative actions for identifying the set of parents are disjoint. Consequently, at each round the learner must decide to either minimize the regret by playing an optimal action or to play an action that is informative for learning the parents. To illustrate the construction, consider an instance with n binary variables and let $m = k$. The causal model is defined such that every variable is equal to 0 unless intervened upon, and the first k variables are the parents of all others. If all of these k variables are set to 1, then all remaining variables also become 1 deterministically. The reward is defined to have mean 0 for all actions except the one that sets the first k variables to 1, which yields the expected reward of 1.

In this setting, the unique high-reward action sets the first k variables to 1, but this simultaneously forces all other variables to be 1, making it impossible to distinguish which variables actually influence the reward. To identify the parent set, the learner must instead intervene on arbitrary subsets of k variables and test whether setting them to 1 increases the reward. However, such actions are suboptimal and incur regret. This complete separation between high-reward and informative actions is the fundamental source of the trade-off.

The construction used in the proof is intentionally degenerate to highlight this phenomenon clearly. While similar lower bounds might be established for non-degenerate models, doing so requires more involved constructions, and the trade-off might be weaker. For instance, suppose we impose a non-degeneracy condition such that each variable takes any possible value with probability at least $\epsilon > 0$ under any parent configuration. In the regime $T \gg 1/\epsilon$, the situation becomes more nuanced. Even when the learner plays the optimal (low-regret) action, it observes alternative configurations of other variables with non-zero probability. These observations can reveal that changes in those variables do not affect the reward, effectively providing information about the parent set "for free." This suggests that, in such settings, the trade-off would be weaker. More details are in Appendix 9.3.

Theorem 3.2 shows the fundamental trade-off between the cumulative regret minimization and the parent identification. More precisely, it implies that any policy achieving the same parent-identification performance as the one using $\bar{d}_{\text{Unif}}$ and uniform sampling (i.e., $\alpha = 1$) necessarily suffers linear cumulative regret. It is noteworthy that there are other pure-exploration objectives in bandit literature without any fundamental trade-off with the cumulative regret. For instance, *simple regret* objective seeks to minimize the expected difference between the estimated and true

best arm [BMS09, ZSSJ23]. For simple regret, it is established that there is no fundamental trade-off with cumulative regret: algorithms that are optimal for cumulative regret can, up to constant factors, also achieve optimal simple regret. For further details, see [LS20, Section 33].

4 KNOWN PARENT SIZE

In this section, we address the cumulative regret minimization problem in causal bandits with an unknown graph, where the number of parents of the reward node k is known to the agent, i.e., $k \in \mathcal{I}$. However, the exact parent set Pa_Y is not known. We consider two regimes based on the relation between the intervention size (m) and the number of parents (k): $m \geq k$ and $m < k$. First, we establish lower bounds on the worst-case regret for both regimes, and then propose an algorithm together with an upper bound on its regret that is close to the lower bounds in both regimes.

4.1 Lower Bounds

The results of this section highlight the difficulty of worst-case regret minimization in causal bandits when the agent knows only the number of parents k but not their identities. The bounds are information-theoretic and hold for any policy π.

An important feature of both lower bounds is that they remain valid even under the additional assumption that the agent has full knowledge of $\mathcal{G}$, the causal graph over $\mathcal{X}$, the set of non-reward variables. Moreover, the bounds hold for any such causal graph. Therefore, the crucial information to minimize the regret is the knowledge of the parent set Pa_Y as knowing $\mathcal{G}$ does not improve the worst-case regret guarantees.

Theorem 4.1 ($m \geq k$ Lower Bound)**.** *For any policy $\pi \in \Pi(\{k, \mathcal{G}\})$, any values $n \geq m \geq k$, and any causal graph $\mathcal{G}$:*

$$R_T(\pi, \mathcal{E}) \in \Omega\left(\sqrt{T \max\left((\ell-1)^k \frac{\binom{n}{k}}{\binom{m}{k}},\ \ell^k\right)}\right). \quad (1)$$

Proof Sketch: To prove this theorem, we establish a novel and more general statement. Let $\mathcal{V}_0 \in \mathcal{E}$ denote the neutral instance where the graph $\mathcal{G}$ is the empty graph, meaning that the variables are random and independent such that the value of any non-intervened variable is always 1, and the reward distribution is always $\mathcal{N}(0,1)$ regardless of the values of its parents. Fix a policy $\pi \in \Pi(\{k, \mathcal{G}\})$. For any pair (p, s) with p denoting a subset of $[n]$ of size k and $s \in [\ell]^k$, let $w_{p,s}(\pi)$ denote the expected fraction of times t in T rounds of interaction between π and $\mathcal{V}_0$ such that $\mathbf{x}_p^{(t)} = s$. Let $\mathcal{P}_k$ denote the set of all such pairs (p, s). We prove the following lower bound, which holds for any graph $\mathcal{G}$:

$$R_T(\pi, \mathcal{E}) \in \Omega\left(\max_{\mathcal{J} \subseteq \mathcal{P}_k} (1 - c_{\mathcal{J}})^2 \sqrt{T \frac{|\mathcal{J}|}{\sum_{(p,s)\in\mathcal{J}} w_{p,s}(\pi)}}\right),$$

where $c_{\mathcal{J}}$ is the maximum value of the fraction $\frac{\sum_{(p,s)\in\mathcal{J}} w_{p,s}(\pi)}{|\mathcal{J}|}$ over all policies. We then focus on a collection of candidate sets $\mathcal{S}$ with $|\mathcal{S}| = k+1$, and show that maximizing the expression over $\mathcal{J} \in \mathcal{S}$ yields the regret lower bound stated in the theorem.

In the second regime, $m < k$, the agent cannot intervene on all parents simultaneously. The bound is obtained by considering instances in which there are k true parents, but only m of them affect the reward.

Theorem 4.2 ($m < k$ Lower Bound)**.** *For any policy $\pi \in \Pi(\{k, \mathcal{G}\})$, any values $n \geq k > m$, and any graph $\mathcal{G}$, we have*

$$R_T(\pi, \mathcal{E}) \in \Omega\left(\sqrt{T \max\left((\ell-1)^m \binom{n}{m},\ \ell^m\right)}\right). \quad (2)$$

Remark 4.3 (Interpretation of the Lower Bounds). To interpret the bounds more concretely, consider the case where $\ell \in \Omega(k)$. This is by assuming that the number of reward's parents in the causal graph is small. In this case, $\frac{\ell^k}{(\ell-1)^k} \in \mathcal{O}(1)$ and for any $\pi \in \Pi(\{k, \mathcal{G}\})$, the following holds:

$$R_T(\pi, \mathcal{E}) \in \Omega\left(\sqrt{T\, \ell^k \frac{\binom{n}{k}}{\binom{m}{k}}}\right), \quad \text{for } m \geq k, \quad (3)$$

$$R_T(\pi, \mathcal{E}) \in \Omega\left(\sqrt{T\, \ell^m \binom{n}{m}}\right), \quad \text{for } m < k. \quad (4)$$

Note that when $m < k$, the quantity $\ell^m \binom{n}{m}$ is exactly the number of actions in $\mathcal{A}_m$, i.e., the set of all interventions on m variables. Therefore, no algorithm can achieve a better worst-case performance than running a standard UCB algorithm over the action set $\mathcal{A}_m$.

4.2 Algorithm and Upper Bound

We are now ready to propose a simple algorithm that achieves near-optimal performance on broad classes of instances. The pseudo-code is presented in Algorithm 1. Note that this algorithm does not attempt to identify the parent set Pa_Y or to recover any other causal relation among the variables. Moreover, it does not require knowledge of the causal graph $\mathcal{G}$; its information set is $\mathcal{I} = \{k\}$, i.e., the number of reward's parents. The algorithm is solely designed for regret minimization and may find the optimal action without explicitly inferring which variables are reward's parents.

For the case $m \geq k$, every action that intervenes on the set Pa_Y and assigns it the optimal values achieves the same expected reward, since the reward distribution depends only on the values of the parents. Leveraging this observation, Algorithm 1 samples a random subset $\mathcal{A}' \subseteq \mathcal{A}_m$ of appropriate size such that with high probability $\mathcal{A}'$ contains an optimal action. The algorithm then runs the standard UCB algorithm on the actions in $\mathcal{A}'$. For the case $m < k$, the algorithm directly runs UCB on the entire action set $\mathcal{A}_m$.

Algorithm 1

1: **Input.** The integers m, n, k, ℓ, T. ($\mathcal{G}$ is unknown)
2: **if** $m \geq k$ **then**
3: Set $n_0 = \min\left(\ell^k \frac{\binom{n}{k}}{\binom{m}{k}} \ln\sqrt{T}\ ,\ \ell^m \binom{n}{m}\right)$.
4: **else**
5: Set $n_0 = \ell^m \binom{n}{m}$.
6: **end if**
7: Construct a uniformly random subset $\mathcal{A}' \subseteq \mathcal{A}_m$ with $|\mathcal{A}'| = n_0$.
8: Run the standard Upper Confidence Bound (UCB) algorithm on the arms in $\mathcal{A}'$ for T rounds.

Theorem 4.4 (Known k Upper Bound)**.** *The worst-case regret of Algorithm 1 with input k is bounded by*

$$R_T\big(\texttt{Alg.1}[k], \mathcal{E}(n,\ell,k)\big) \in \begin{cases} \tilde{\mathcal{O}}\left(\sqrt{T\ell^k \frac{\binom{n}{k}}{\binom{m}{k}}}\right), & m \geq k, \\ \tilde{\mathcal{O}}\left(\sqrt{T\ell^m \binom{n}{m}}\right), & m < k, \end{cases} \tag{5}$$

where $\tilde{\mathcal{O}}$ hides constants and logarithmic factors.

Comparing the lower bounds for when $\ell \in \Omega(k)$ given in (3) and (4) with the regret bound of Algorithm 1 in (5) shows that Algorithm 1 is optimal in both regimes up to logarithmic factors. Moreover, since this algorithm does not require knowledge of $\mathcal{G}$, whereas the lower bounds apply even to algorithms with access to $\mathcal{G}$, we conclude that prior knowledge of $\mathcal{G}$ does not improve the worst-case regret bounds.

Remark 4.5. In the case of $m = n$, the derived lower and upper bounds match (up to logarithmic factors), yielding a tight bound of $\tilde{\mathcal{O}}\left(\sqrt{T\,\ell^k}\right)$. This is precisely the regret bound of the setting where the agent knows the exact set of Pa_Y and applies standard regret minimization algorithms over the ℓ^k possible arms. Hence, surprisingly, when the agent can intervene on all variables in each round, the regret with full knowledge of the graph (including the parents of the reward) is equal to that with no knowledge.

5 UNKNOWN PARENT SIZE

In this section, we drop the assumption that the agent knows k, the number of reward's parents. We first establish a lower bound showing that no algorithm can fully adapt to the unknown value of k without incurring a penalty: specifically, it is impossible to achieve the same regret bounds as in the known-k case uniformly over all values of k. Afterwards, we propose an algorithm with an empty information set, $\mathcal{I} = \{\}$, whose regret differs from our lower bound by only a small margin.

Change of performance measure: When an algorithm does not know the value of k, its performance may vary significantly across instances with different k. Therefore, its performance should be assessed for all possible values of k, i.e., $k \in [n]$ rather than a single k. However, as indicated by the upper and lower bounds from the previous section, for $k > m$, any action could be optimal, meaning the algorithm must explore all actions regardless of k. Therefore, we focus on the regime $k \leq m \leq n$ and evaluate the performance of a policy π using the following vector, which we call the *regret vector* of the policy:

$$\Big[R_T\big(\pi, \mathcal{E}(n,\ell,k)\big)\Big]_{k\in[m]}.$$

Recall that $R_T\big(\pi, \mathcal{E}(n,\ell,k)\big)$ denotes the worst-case regret of π over all instances with n variables and k reward parents.

One might consider the maximum worst-case regret over k (i.e., the largest entry of the above vector) as a performance measure, but this is unsatisfactory: since for any $k_1 < k_2$, we have $\mathcal{E}(n,\ell,k_1) \subseteq \mathcal{E}(n,\ell,k_2)$, which implies

$$R_T\big(\pi, \mathcal{E}(n,\ell,k_1)\big) \leq R_T\big(\pi, \mathcal{E}(n,\ell,k_2)\big).$$

Hence, the maximum entry of the regret vector always occurs at $k = m$. A trivial strategy that optimizes for $k = m$ is to run standard UCB on all actions; however, this is not necessarily the optimal approach. On the other hand, the next theorem shows that any algorithm tailored to instances with $k = k_1$ necessarily exhibits sub-optimal performance on instances with $k > k_1$. This highlights the need for algorithms that adapt to the unknown value of k and perform well across all $k \in [m]$. To compare different algorithms, we use the notions of Pareto domination and Pareto optimality for their regret vectors.

Definition 5.1 (Rate-Pareto Domination and Optimality)**.** Let $\mathbf{r}, \mathbf{s} \in \mathbb{R}_+^m$ be two regret vectors, where each entry is a function of the parameters T, n, m, ℓ. We say that $\mathbf{r}$ *rate-Pareto dominates* $\mathbf{s}$ if there exists a

universal constant $C > 0$ such that

$$\forall k \in [m]: \quad r_k \;\le\; C\, s_k,$$

and the reverse does not hold (i.e., there is no constant $C' > 0$ such that $\forall k \in [m]: \; s_k \le C' r_k$). In other words, $\mathbf{r}$ is not worse than $\mathbf{s}$ in every coordinate (up to constant factors) and is strictly better in at least one coordinate.

A regret vector $\mathbf{r}$ is said to be *rate-Pareto optimal* for a set of policies Π if no regret vector corresponding to a policy in Π rate-Pareto dominates $\mathbf{r}$. A policy π is *rate-Pareto optimal* if its regret vector is rate-Pareto optimal.

5.1 Lower Bound

Herein, we present a lower bound on the product of two distinct entries of the regret vector for any policy $\pi \in \Pi(\mathcal{G})$.

Theorem 5.2 (Unknown k Lower Bound)**.** *For any causal graph $\mathcal{G}$, any policy $\pi \in \Pi(\mathcal{G})$, and any values $n \ge m \ge k_2 > k_1$, we have*

$$R_T\big(\pi, \mathcal{E}(n,\ell,k_1)\big) \times R_T\left(\pi, \mathcal{E}(n,\ell,k_2)\right) \in \Omega\left(T \max\left((\ell-1)^{k_2} \frac{\binom{n-k_1}{k_2-k_1}}{\binom{m-k_1}{k_2-k_1}},\; \ell^{k_2}\right)\right). \tag{6}$$

This result shows a fundamental trade-off: for any policy in $\Pi(\{\mathcal{G}\})$, improving performance on instances with a fixed value of $k = k_1$ necessarily worsens performance on instances with a larger value of $k = k_2 > k_1$.

5.2 Algorithm and Upper Bound

As discussed earlier, if the goal were to minimize the maximum worst-case regret (i.e., minimizing the largest entry of the regret vector), Algorithm 1 could be applied with the input $k = m$. In this case, we obtain the following for each $k \in [m]$,

$$R_T\big(\texttt{Alg.1}[m], \mathcal{E}(n,v,k)\big) \in \mathcal{O}\left(\sqrt{T\,\ell^m \binom{n}{m}}\right).$$

However, based on the result of Theorem 5.2, the incurred regret by Algorithm 1 is not optimal when the actual number of parents is less than m. We now propose an algorithm that adapts to the unknown value of k and incurs lower regret when k is small. The design is adapted from the setting of bandits with multiple optimal arms [ZN20], which in turn is inspired by the literature on continuum-armed bandits [Agr95, LC18, Had19].

The algorithm proceeds in phases. In each phase i, it randomly selects a subset $S_i \subseteq \mathcal{A}_m$ of arms of size q_i, and then runs the standard UCB algorithm for ΔT_i rounds on the arms in S_i together with the mixture arms constructed in earlier phases. At the end of phase i, the algorithm defines a *mixture arm* $\tilde{a}_i$ based on the actions played during that phase. Formally, if the actions played in phase i are $a_1, ..., a_{\Delta T_i}$, then $\tilde{a}_i$ is the randomized arm that, when played, samples an action a uniformly from $\{a_1, ..., a_{\Delta T_i}\}$ and plays it. Intuitively, $\tilde{a}_i$ summarizes the exploration of phase i into a single representative action, while preserving the empirical distribution of actions observed in that phase.

The schedule of the algorithm is such that q_i is halved at each new phase, while the phase length ΔT_i is doubled. Pseudocode for this procedure is given in Algorithm 2, and the theorem below provides its regret upper bound on the class of instances with k parents, for any $k \in [n]$. It is important to note that Algorithm 2 requires neither knowledge of $\mathcal{G}$ nor of k.

Algorithm 2

1: **Input.** The integers m, n, ℓ, T. ($\mathcal{G}$ is unknown)
2: **Initialization.** Set $i_f = \left\lceil \log_2 \sqrt{T\frac{m}{\ell n}} \right\rceil, \forall i \in [i_f]:$ $q_i = 2^{\lceil \log_2 \sqrt{T} \rceil - i + 1}, \Delta T_i = \left\lceil \frac{\ell n}{m} \right\rceil 2^{\lceil \log_2 \sqrt{T} \rceil + i}$, and $M = \emptyset$.
3: **for** i in $1, 2, \ldots, i_f$ **do**
4: Construct $S_i \subseteq \mathcal{A}_m$ consisting if q_i uniform random actions selected with replacement.
5: Run UCB on actions in $S_i \cup M$ for ΔT_i rounds.
6: Construct the mixture arm $\tilde{a}_i$ from the UCB actions and add it to M.
7: **end for**

Theorem 5.3 (Unknown k Upper Bound)**.** *For any k, the worst-case regret of Algorithm 2 on all the instances in $\mathcal{E}(n,\ell,k)$ is upper bounded as follows*

$$R_T\left(\texttt{Alg.2}, \mathcal{E}(n,\ell,k)\right) \in \begin{cases} \tilde{\mathcal{O}}\left(\sqrt{T\frac{m}{n}\ell^{k-\frac{1}{2}}\frac{\binom{n}{k}}{\binom{m}{k}}}\right), m \ge k, \\ \tilde{\mathcal{O}}\left(\sqrt{T\frac{m}{n}\ell^{m-\frac{1}{2}}\binom{n}{m}}\right), m < k, \end{cases} \tag{7}$$

where $\tilde{\mathcal{O}}$ hides constants and logarithmic factors.

Next lemma uses the results of Theorems 5.3 and 5.2 to show that Algorithm 2 is near rate-Pareto optimal.

Lemma 5.4 (Pareto Optimality of Algorithm 2)**.** *The following statements hold:*

- *When $m = n$, Algorithm 2 is rate-Pareto optimal for $\Pi(\{\mathcal{G}\})$, up to logarithmic factors.*
- *In the general case, if $\ell \in \Omega(m)$, then the regret*

vector

$$\Big[R_T\big(\texttt{Alg.2}, \mathcal{E}(n,\ell,1)\big),\ R_T\big(\texttt{Alg.2}, \mathcal{E}(n,\ell,2)\big)\tfrac{m}{n},\ \ldots,\ R_T\big(\texttt{Alg.2}, \mathcal{E}(n,\ell,m)\big)\tfrac{m}{n}\Big]$$

is rate-Pareto optimal for $\Pi(\{\mathcal{G}\})$*. Up to logarithmic terms, this vector coincides with the regret vector of Algorithm 2 when* $k = 1$*, and exhibits a multiplicative gap of* $\frac{m}{n}$ *for larger values of* k*.*

Note that, similar to the known parent-size setting, the lower bound in Theorem 5.2 applies to any policy in $\Pi(\{\mathcal{G}\})$. In contrast, Algorithm 2 operates without knowledge of $\mathcal{G}$, yet, as shown in Lemma 5.4, it remains close to Pareto optimal within $\Pi(\{\mathcal{G}\})$. This shows that even in the case of an unknown k, knowing the graph $\mathcal{G}$ does not significantly improve the rate of the worst-case regret.

6 EXPERIMENTS

In this Section, we evaluate the proposed algorithms on a variety of instances and compare their performance with existing methods. For all experiments, we generate Erdős–Rényi random graphs with edge probability $p = \frac{2}{n}$. The reward is modeled as a binary random variable whose mean is chosen uniformly at random from $[0, 1]$ for each possible combination of parent values. Each experiment is repeated 100 times; solid lines in the plots represent averages over these runs, while shaded regions indicate one standard deviation above and below the mean. Additional details on the setup of the experiments, such as the algorithm implementations, and further results are provided in the Appendix 10. The code is available at https://github.com/ban-epfl/Unknown-Graph-Causal-Bandits.

We compare against the following four algorithms:

EmpKnownUCB+. This is an empirical variant of Algorithm 1. In the experiments, if the algorithm chooses action $a_t = (p_t, s_t)$ at round t and observes $(\mathbf{x}^{(t)}, y_t)$, then for any other action $a = (p, s) \in \mathcal{A}_m$ such that $\mathbf{x}_p^{(t)} = s$, we also treat y_t as a reward sample for a. While this modification may lead to high regret in certain instances, our experiments indicate that it improves the average performance on random instances.

EmpUnknownUCB+. This is an empirical variant of Algorithm 2 with two modifications: (i) samples collected in each phase are reused to construct confidence bounds for the arms in subsequent phases, and (ii) for phases $i > 1$, the algorithm selects q_i arms with the highest empirical means, instead of choosing them uniformly at random.

RAPS. This algorithm, proposed in [KEK25], represents the most recent approach for causal bandits with an unknown graph. It proceeds in two phases. Phase one is purely focused on parent identification, without regard for regret, and involves a sequential search procedure, only on atomic interventions, that identifies parents one by one. Phase two runs a standard UCB algorithm on the identified parent nodes. We additionally reuse the samples collected during phase one in phase two.

Standard UCB. This baseline algorithm simply runs the standard UCB algorithm on the entire set of actions.

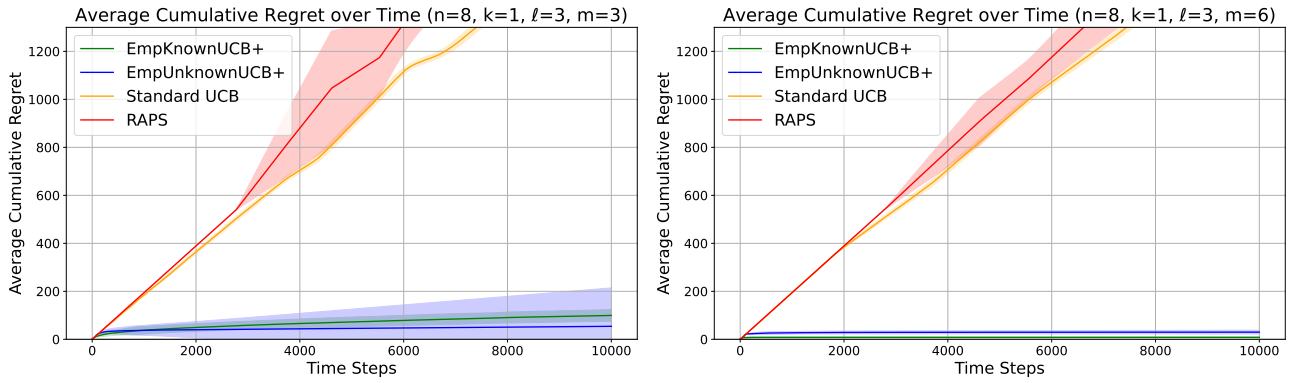
In the following two experiments, we set $k = 1$ and compare the regret of the algorithms over time and across different numbers of nodes. We fix $k = 1$ to provide a favorable setting for the RAPS algorithm, since its first phase, which is dedicated entirely to parent identification, typically requires a large number of rounds and is repeated once per parent, thereby significantly increasing regret in practice. Although this is the most advantageous case for RAPS, our results demonstrate a substantial performance gap between our algorithms and both RAPS and Standard UCB. More experimental results are provided in Appendix 10.

Figures 1a and 1b report the average regret of the algorithms on an instance with parameters $n = 8$, $k = 1$, $\ell = 3$, and intervention sizes $m = 3, 6$, for a horizon of $T = 10000$. Our two proposed algorithms perform nearly identically and achieve more than a $20\times$ improvement compared to the baselines. This close performance aligns with our theoretical analysis: for $k = 1$, the regret bound of Algorithm 2 coincides with that of Algorithm 1, and the additional adaptation cost appears only when $k > 1$.

Figure 2 shows the average final regret of the algorithms on instances with parameters $k = 1$, $\ell = 3$, for varying values of $n \in \{3, 5, 8, 15\}$, with action size $m = 3$ and horizon $T = 10000$. The results again show that EmpKnownUCB+ and EmpUnknownUCB+ achieve very similar performance, with EmpUnknownUCB+ showing higher variance, while both methods substantially outperform the two baseline algorithms.

7 DISCUSSION AND FUTURE WORK

We studied causal bandits with unknown causal structure and showed that, under no distributional assumptions, worst-case regret minimization does not require identifying the reward's parents. Our results establish the existence of a trade-off between structure learning and regret minimization, while our algorithms achieve



(a) (b)

Figure 1: Average regret of the algorithms over time. (a) Corresponds to the setting with action size $m = 3$. (b) Corresponds to the setting with action size $m = 6$.

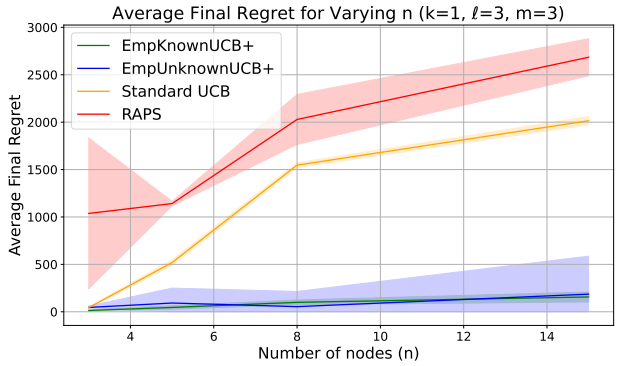


Figure 2: Average cumulative regret of algorithms at time T for instances with varying numbers of nodes.

nearly optimal rates without recovering the parent set. In both the known and unknown parent-size regimes, our algorithms provably outperform existing baselines and demonstrate strong empirical performance, highlighting that regret minimization can be addressed directly without explicit causal discovery.

While we prove that regret minimization and parent identification can be conflicting objectives in certain instances, characterizing this trade-off remains an interesting direction for future work. For example, one could study the settings in which these objectives are aligned and can be optimized simultaneously, or characterize achievable pairs of parent identification error rates and regret rates, and design algorithms that achieve such trade-offs.

We established that graph learning is not optimal for regret minimization without distributional assumptions. Still it would be valuable to investigate whether there exist realistic assumptions on the causal model that improve the regret. Since structural information alone does not significantly improve regret rates, future work should focus on distributional assumptions that could meaningfully enhance learning performance.

References

[AAMW24] Abhineet Agarwal, Anish Agarwal, Lorenzo Masoero, and Justin Whitehouse. Mutli-armed bandits with network interference. *Advances in Neural Information Processing Systems*, 37:36414–36437, 2024.

[ACBDK15] Noga Alon, Nicolo Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Conference on Learning Theory*, pages 23–35. PMLR, 2015.

[ACBG+17] Noga Alon, Nicolo Cesa-Bianchi, Claudio Gentile, Shie Mannor, Yishay Mansour, and Ohad Shamir. Nonstochastic multi-armed bandits with graph-structured feedback. *SIAM Journal on Computing*, 46(6):1785–1826, 2017.

[Agr95] Rajeev Agrawal. The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951, 1995.

[AYPS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

[BCB+12] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and

nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[BMS09] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[BR19] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

[BWR22] Blair Bilodeau, Linbo Wang, and Dan Roy. Adaptively exploiting d-separators with causal bandits. *Advances in Neural Information Processing Systems*, 35:20381–20392, 2022.

[DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory*, number 101, pages 355–366, 2008.

[EGK24] Muhammad Qasim Elahi, Mahsa Ghasemi, and Murat Kocaoglu. Partial structure discovery is sufficient for no-regret learning in causal bandits. *Advances in Neural Information Processing Systems*, 37:109066–109100, 2024.

[FC23] Shi Feng and Wei Chen. Combinatorial causal bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7550–7558, 2023.

[FXC25] Shi Feng, Nuoya Xiong, and Wei Chen. Combinatorial causal bandits without graph skeleton. In *Asian Conference on Machine Learning*, pages 271–286. PMLR, 2025.

[GZS19] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[Had19] Hédi Hadiji. Polynomial cost of adaptation for x-armed bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[JEK24] Fateme Jamshidi, Jalal Etesami, and Negar Kiyavash. Confounded budgeted causal bandits. In *Causal Learning and Reasoning*, pages 423–461. PMLR, 2024.

[JFK24] Su Jia, Peter Frazier, and Nathan Kallus. Multi-armed bandits with interference. *arXiv preprint arXiv:2402.01845*, 2024.

[JSK25] Fateme Jamshidi, Mohammad Shahverdikondori, and Negar Kiyavash. Graph-dependent regret bounds in multi-armed bandits with interference. *arXiv preprint arXiv:2503.07555*, 2025.

[KEK25] Mikhail Konobeev, Jalal Etesami, and Negar Kiyavash. Causal bandits without graph learning. In *Causal Learning and Reasoning*, pages 31–63. PMLR, 2025.

[LAR] Ziyi Liu, Idan Attias, and Daniel M Roy. Causal bandits: The pareto optimal frontier of adaptivity, a reduction to linear bandits, and limitations around unknown marginals. In *Forty-first International Conference on Machine Learning*.

[LAR22] Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In *Uncertainty in Artificial Intelligence*, pages 1052–1062. PMLR, 2022.

[LB18] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? *Advances in neural information processing systems*, 31, 2018.

[LB19] 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.

[LC18] Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in x-armed bandits. In *Conference on Learning Theory*, pages 1463–1492. PMLR, 2018.

[LCLS10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[LLR16] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. *Advances in neural information processing systems*, 29, 2016.

[LMT21] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. *Advances in*

Neural Information Processing Systems, 34:24817–24828, 2021.

[LMTY20] 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.

[LS20] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[LSN+20] Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of medical Internet research*, 22(7):e18477, 2020.

[MAC23] Alan Malek, Virginia Aglietti, and Silvia Chiappa. Additive causal bandits with unknown graph. In *International Conference on Machine Learning*, pages 23574–23589. PMLR, 2023.

[MEAK25] Ehsan Mokhtarian, Sepehr Elahi, Sina Akbari, and Negar Kiyavash. Recursive causal discovery. *Journal of Machine Learning Research*, 26(61):1–65, 2025.

[NPS21] 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.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[PZM25] Chen Peng, Di Zhang, and Urbashi Mitra. Asymmetric graph error control with low complexity in causal bandits. *IEEE Transactions on Signal Processing*, 2025.

[SARK25] Mohammad Shahverdikondori, Amir Mohammad Abouei, Alireza Rezaeimoghadam, and Negar Kiyavash. Optimal best arm identification with post-action context. *arXiv preprint arXiv:2502.03061*, 2025.

[Sco10] Steven L Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.

[SEG25] Suryanarayana Sankagiri, Jalal Etesami, and Matthias Grossglauser. Recommendations from sparse comparison data: Provably fast convergence for nonconvex matrix factorization. *International Conference on Machine Learning*, 2025.

[SFDvO25] Francisco NFQ Simoes, Itai Feigenbaum, Mehdi Dastani, and Thijs van Ommen. The minimal search space for conditional causal bandits. *arXiv preprint arXiv:2502.06577*, 2025.

[SMK24] Mohammad Shahverdikondori, Ehsan Mokhtarian, and Negar Kiyavash. QWO: Speeding up permutation-based causal discovery in LiGAMs. *Advances in Neural Information Processing Systems*, 37, 2024.

[SSDS17] Rajat Sen, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *International Conference on Machine Learning*, pages 3057–3066. PMLR, 2017.

[VBW15] Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

[VCB22] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

[VSST23] Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *Journal of Machine Learning Research*, 24(297):1–59, 2023.

[XC] Nuoya Xiong and Wei Chen. Combinatorial pure exploration of causal bandits. In *The Eleventh International Conference on Learning Representations*.

[YHS+18] Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. In *International Conference on Machine Learning*, pages 5512–5520. PMLR, 2018.

[YMVT24] Zirui Yan, Arpan Mukherjee, Burak Varıcı, and Ali Tajer. Robust causal bandits for linear models. *IEEE Journal on Selected Areas in Information Theory*, 5:78–93, 2024.

[YT24] Zirui Yan and Ali Tajer. Linear causal bandits: Unknown graph and soft interventions. *Advances in Neural Information Processing Systems*, 37:23939–23987, 2024.

[ZN20] Yinglun Zhu and Robert Nowak. On regret with multiple best arms. *Advances in Neural Information Processing Systems*, 33:9050–9060, 2020.

[ZSSJ23] Yao Zhao, Connor Stephens, Csaba Szepesvári, and Kwang-Sung Jun. Revisiting simple regret: Fast rates for returning a good arm. In *International Conference on Machine Learning*, pages 42110–42158. PMLR, 2023.

[ZZ25] Yijia Zhao and Qing Zhou. Causal bandits with backdoor adjustment on unknown gaussian dags. *arXiv preprint arXiv:2502.02020*, 2025.

8 AISTATS Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

9 OMITTED PROOFS

9.1 Additional Notation and Helper Lemmas

In this subsection, we introduce notation and a few helper quantities used in the proofs of the main theorems.

For any pair of probability measures P, Q, we denote their Kullback-Leibler divergence by $\mathrm{KL}(P, Q)$. For any integer $r \geq 1$, define

$$\mathcal{P}_r \coloneqq \left\{(p, s) \;\middle|\; p \in \tbinom{[n]}{r},\ s \in [\ell]^r\right\},$$

where $\binom{[n]}{r}$ denotes the set of all subset of $[n]$ of size r. We use $\mathcal{P}_r$ to encode pairs (p, s) where p specifies the indices of r variables (out of n) and s lists, in increasing index order, the special values assigned to those variables. In particular, the set of size m actions $\mathcal{A}_m$ can be identified with $\mathcal{P}_m$.

For $r \geq 1$, any $(p, s) \in \mathcal{P}_r$, any policy π, and any horizon $t \in T$, let

$$N_{p,s}(t, \pi) \coloneqq \sum_{u=1}^{t} \mathbb{1}_{\left\{\mathbf{x}_p^{(u)} = s\right\}},$$

be the number of rounds up to time t in which the coordinates indexed by p take the value pattern s, and define the corresponding empirical mean reward

$$\hat{\mu}_{p,s}(t, \pi) \coloneqq \begin{cases} \dfrac{1}{N_{p,s}(t, \pi)} \displaystyle\sum_{u=1}^{t} Y_u \, \mathbb{1}_{\left\{\mathbf{x}_p^{(u)} = s\right\}} & \text{if } N_{p,s}(t, \pi) > 0, \\ 0 & \text{if } N_{p,s}(t, \pi) = 0. \end{cases}$$

Here $(\mathbf{x}^{(u)}, Y_u)$ denotes the observation at round u along the trajectory induced by policy π (we omit the instance for brevity). Intervened variables are counted with their assigned values. We similarly write $N_a(t, \pi)$ and $\hat{\mu}_a(t, \pi)$ for any action $a \in \mathcal{A}$.

When clear from context, we drop the explicit dependence on π in each of these notations.

For any instance with $|\mathrm{Pa}_{\mathcal{G}}(Y)| = k$, let α_k denotes the fraction of optimal arms in $\mathcal{A}_m$. By Lemma 2.1, the optimal action lies in $\mathcal{A}_m$. For $m > k$, in $\mathcal{A}_m$ there are at least $\ell^{m-k}\binom{n-k}{m-k}$ arms achieving the optimal mean reward, since the reward depends only on the values of the k parents and there are exactly this many interventions that set the values of all parents to their optimal combination. Thus,

$$\alpha_k = \frac{\ell^{m-k}\binom{n-k}{m-k}}{\ell^m \binom{n}{m}} = \frac{\binom{m}{k}}{\ell^k \binom{n}{k}}. \tag{8}$$

For values of k with $m < k$, we know one of the actions in $\mathcal{A}_m$ is optimal, thus $\alpha_k = \frac{1}{|\mathcal{A}_m|} = \frac{1}{\ell^m \binom{n}{m}}$.

Now we provide two definitions from the literature on causal inference, which we will use in the proof of Lemma 2.1.

Definition 9.1 (Blocked Path)**.** Let $\mathcal{G}$ be a directed acyclic graph (DAG). A path between two nodes X and Y is said to be *blocked* given a set of nodes Z if at least one of the following holds:

- The path contains a chain $X_i \to X_j \to X_k$ or a fork $X_i \leftarrow X_j \to X_k$ such that the middle node $X_j \in Z$.
- The path contains a collider $X_i \to X_j \leftarrow X_k$ such that neither X_j nor any of its descendants belong to Z.

Definition 9.2 (d-separation)**.** In a DAG $\mathcal{G}$, two disjoint sets of variables A and B are said to be *d-separated* given a set of variables Z if every path between a node in A and a node in B is blocked given Z. We denote this d-separation by $(A \perp\!\!\!\perp B|Z)_{\mathcal{G}}$.

Lemma 9.3 (Pinsker Inequality)**.** *For measures P and Q on the same probability space $(\Omega, \mathcal{F})$*

$$\delta(P,Q) \;\coloneqq\; \sup_{A\in\mathcal{F}} \big(P(A) - Q(A)\big) \;\le\; \sqrt{\frac{1}{2}\,\mathrm{KL}\,(P,Q)}.$$

Lemma 9.4 ([LS20], Section 14)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let $P, Q : \mathcal{F} \to [0,1]$ be probability measures. Let $a < b$ and $X : \Omega \to [a,b]$ be an $\mathcal{F}$-measurable random variable, we have*

$$\left|\int_\Omega X(\omega)\, dP(\omega) - \int_\Omega X(\omega)\, dQ(\omega)\right| \;\le\; (b-a)\,\delta(P,Q),$$

where $\delta(P,Q)$ is as defined in Lemma 9.3.

Theorem 9.5 (Bretagnolle–Huber Inequality)**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$P(A) + Q(A^c) \ge \frac{1}{2}\exp\big(-\mathrm{KL}\,(P,Q)\big),$$

where $A^c = \Omega \setminus A$.

Lemma 9.6. *Let $X_1, X_2, \ldots, X_k$ be sub-Gaussian random variables with parameter $\sigma^2 = 1$ and means $\mathbb{E}[X_i] = \mu_i \in [a,b]$. Define the mixture random variable X such that $X = X_i$ with probability p_i. Then X is sub-Gaussian with parameter*

$$\sigma_X^2 = 1 + \frac{(b-a)^2}{4}.$$

Proof. We can decompose X as $X = Y + \epsilon$, where Y is a discrete random variable taking value μ_i with probability p_i, and ϵ is a 1-sub-Gaussian random variable independent of Y. The random variable Y is $\frac{(b-a)}{2}$-sub-Gaussian because it is supported on an interval of length $(b-a)$. Since ϵ is 1-sub-Gaussian, and the sum of independent σ_1- and σ_2-sub-Gaussian random variables is $\sqrt{\sigma_1^2 + \sigma_2^2}$-sub-Gaussian, it follows that

$$\sigma_X^2 = 1 + \frac{(b-a)^2}{4}.$$

□

Lemma 9.7. *Let α denote the fraction of optimal arms in $\mathcal{A}_m$, i.e., the number of optimal arms divided by $|\mathcal{A}_m|$. Then, if we choose a random subset of arms of size $\frac{1}{\alpha}\ln\sqrt{T}$, either with or without replacement, the probability that the subset contains no optimal arm is less than $\frac{1}{\sqrt{T}}$.*

Proof. Let $N = |\mathcal{A}_m|$ be the total number of arms, so the number of optimal arms is αN. Consider sampling with replacement. In each draw, the probability of picking a non-optimal arm is $1-\alpha$. After

$$q \;\coloneqq\; \frac{1}{\alpha}\ln\sqrt{T}$$

independent draws, the probability that all sampled arms are non-optimal is

$$(1-\alpha)^q \;\le\; \exp(-\alpha q) \;=\; \exp\big(-\ln\sqrt{T}\big) \;=\; \frac{1}{\sqrt{T}},$$

where the inequality comes from the fact that $\forall x \in \mathbb{R} : 1 - x \le \exp(-x)$. Thus, with replacement, the probability that the sampled set contains no optimal arm is at most $1/\sqrt{T}$.

Sampling without replacement can only decrease the probability of missing all optimal arms (it is stochastically dominated by the with-replacement model), so the same bound applies. □

9.2 Proof of Lemma 2.1

In this section, we present the proof of Lemma 2.1.

Lemma 2.1. *Under causal sufficiency, for any values of n, ℓ, k, m and any instance $\mathcal{V} \in \mathcal{E}$, there exists an optimal action with the maximum intervention size, that is* $\max_{a \in \mathcal{A}_m} \mu_a = \mu_{a^*}$.

Proof. To prove this lemma, we separate two cases depending on whether $m \geq k$ or $m < k$.

Case 1: $m \geq k$**.** In this case, for any action $a = (p, s) \in \mathcal{A}$, we have

$$\begin{aligned}\mu_a = \mathbb{E}\big[Y \mid do(\mathbf{X}_p = s)\big] &= \sum_{\mathbf{z} \in [\ell]^k} \mathbb{P}\big(\mathrm{Pa}_{\mathcal{G}}(Y) = \mathbf{z} \mid do(\mathbf{X}_p = s)\big)\, \mathbb{E}\big[Y \mid \mathrm{Pa}_{\mathcal{G}}(Y) = \mathbf{z}\big] \\ &\leq \max_{\mathbf{z} \in [\ell]^k} \mathbb{E}\big[Y \mid \mathrm{Pa}_{\mathcal{G}}(Y) = \mathbf{z}\big] = \max_{\mathbf{z} \in [\ell]^k} \mathbb{E}\big[Y \mid do(\mathrm{Pa}_{\mathcal{G}}(Y) = \mathbf{z})\big] \\ &= \mu_{a_{\mathbf{z}}},\end{aligned}$$

where $a_{\mathbf{z}}$ is any of the actions in $\mathcal{A}_m$ that intervenes on all the reward parents and sets their values to $\mathbf{z}$.

Case 2: $m < k$**.** This case requires a more involved argument. We show that for any action $a = (p, s)$, there exists an intervention that intervenes on one more variable than a and achieves mean reward at least μ_a. Repeating this process yields an action of size m with mean reward at least as large as μ_a.

Fix any $r \notin p$. Then

$$\begin{aligned}\mu_a &= \mathbb{E}\big[Y \mid do(\mathbf{X}_p = s)\big] \\ &= \sum_{i \in [\ell]} \mathbb{P}\big(X_r = i \mid do(\mathbf{X}_p = s)\big)\, \mathbb{E}\big[Y \mid do(\mathbf{X}_p = s), X_r = i\big].\end{aligned}$$

If we can show that

$$\mathbb{P}\big(Y \mid do(\mathbf{X}_p = s), X_r = i\big) = \mathbb{P}\big(Y \mid do(X_p = s, X_r = i)\big) \quad \forall i \in [\ell],$$

then it follows that

$$\begin{aligned}\mu_a &= \sum_{i \in [\ell]} \mathbb{P}\big(X_r = i \mid do(\mathbf{X}_p = s)\big)\, \mathbb{E}\big[Y \mid do(X_p = s, X_r = i)\big] \\ &\leq \max_{i \in [\ell]} \mathbb{E}\big[Y \mid do(X_p = s, X_r = i)\big] = \mu_{a'},\end{aligned}$$

where a' is an action that extends a by also intervening on X_r. Thus, it suffices to show that such an X_r exists.

By the second rule of Pearl's do-calculus [Pea09], a sufficient condition for the above equality is that

$$(Y \perp\!\!\!\perp X_r \mid X_p)_{\mathcal{G}'_{\overline{X_p}\underline{X_r}}}, \tag{9}$$

where $\mathcal{G}' = \mathcal{G} \cup \{Y\}$, and $\mathcal{G}'_{\overline{X_p}\underline{X_r}}$ denotes the graph $\mathcal{G}'$ with all incoming edges to X_p removed and all outgoing edges from X_r removed.

To prove (9), let $\sigma : [n] \to [n]$ be a topological ordering of the DAG $\mathcal{G}$, i.e., in the sequence $X_{\sigma(1)}, X_{\sigma(2)}, \ldots, X_{\sigma(n)}$ all edges point forward. Choose r as the first index in this order such that $r \notin p$. We claim this choice of X_r satisfies (9).

Consider any path $X_r - X_{i_1} - X_{i_2} - \ldots - X_{i_t} - Y$ in $\mathcal{G}'_{\overline{X_p}\underline{X_r}}$. This path has the following properties:

1. There must be at least one intermediate variable other than X_r and Y (i.e., $t > 0$), because all outgoing edges from X_r are removed and Y has no children. Thus $X_r \to Y$ and $X_r \leftarrow Y$ are impossible.
2. The first edge must be oriented $X_r \leftarrow X_{i_1}$, since no outgoing edges from X_r remain in $\mathcal{G}'_{\overline{X_p}\underline{X_r}}$.
3. By the causal order and the choice of r, we must have $X_{i_1} \in X_p$ or there is no path between X_r and Y.

Therefore, X_{i_1} lies on the path, is conditioned on, and blocks the path regardless of the orientation of the next edge. Hence all such paths are blocked, establishing (9).

This proves that for any action a there exists an extended action a' with $\mu_{a'} \geq \mu_a$. Repeating this argument iteratively shows that there exists an optimal action in $\mathcal{A}_m$, completing the proof. □

9.3 Proofs of Section 3

Herein, we present the proofs for Section 3 of the main text.

Theorem 3.2 (Identification-Regret Trade-Off)**.** *There exists a subclass $\mathcal{E}_0 \subseteq \mathcal{E}(n, \ell, k)$ such that for $m = k$, the two following statements hold:*

1. *There exists a decision rule $\bar{d}_{Unif}$ that combined with the uniform sampling policy achieves $\delta_T\left(\pi_{Unif}, \bar{d}_{Unif}, \mathcal{E}_0\right) \in \mathcal{O}\left(\exp\left(-T\right)\right)$.*
2. *The regret of any policy π for which there exists a decision rule $\bar{d}$ such that $\delta_T\left(\pi, \bar{d}, \mathcal{E}_0\right) \in \mathcal{O}\left(\exp\left(-T^\alpha\right)\right)$ grows as $R_T\big(\pi, \mathcal{E}_0, \{k\}\big) \in \Omega\left(T^\alpha\right)$.*

Proof. To prove this result, we first introduce the class of instances $\mathcal{E}_0$. The definition of $\mathcal{E}_0$ proceeds in three steps: we specify the common graph $\mathcal{G}$ over $\mathcal{X}$, then the conditional distributions of variables given their parents, and finally the parent set Pa_Y together with the reward distribution. Throughout, we assume all variables are binary.

The graph $\mathcal{G}$ shared among all instances in $\mathcal{E}_0$ is defined as follows. For every $i \in \{1, 2, \ldots, k\}$ and every $j \in \{k+1, \ldots, n\}$, X_i is a parent of X_j, and there are no other edges between pairs of nodes. This structure is illustrated in Figure 3.

For the conditional distributions, the variables X_i with $i \in [k]$ have no parents and are independent. We fix $\mathbb{P}(X_i = 0) = 1$ for each such variable, meaning they always take the value 0 unless intervened upon. For any $i > k$, we define

$$\mathbb{P}\left(X_i = 1 \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\right) = \begin{cases} 0 & \text{if } x_1 x_2 \cdots x_k = 0, \\ 1 & \text{if } x_1 x_2 \cdots x_k = 1. \end{cases}$$

That is, these variables are always equal to 0, except in the case where all nodes $\{X_1, X_2, \ldots, X_k\}$ equal 1, in which case they take the value 1.

The reward variable Y has distribution $\mathcal{N}(0, 1)$ for any combination of parents except the case where all parents equal 1, in which case its distribution is $\mathcal{N}(1, 1)$. Thus, the mean reward is 0 in all cases except when every parent is set to 1, where the mean is 1.

All instances in $\mathcal{E}_0$ share the same graph $\mathcal{G}$, the distributions of all variables, and the reward distribution. The only difference between them is the identity of the parent set Pa_Y. For each set $p \in \binom{[n]}{k}$, we construct an instance $\mathcal{V}_p \in \mathcal{E}_0$ where $\mathrm{Pa}_Y = p$. Hence, the total number of instances is $|\mathcal{E}_0| = \binom{n}{k}$.

We now prove that in the setting $m = k$, uniform sampling, combined with a suitable suggestion rule $\bar{d}_{\mathrm{Unif}}$, achieves parent identification error $\mathcal{O}\left(\exp\left(-T\right)\right)$ on instances in $\mathcal{E}_0$.

Decision Rule $\bar{d}_{\mathbf{Unif}}$. Given the empirical means $\hat{\mu}_a(T)$ for $a \in \mathcal{A}_k$ (policy/instance dependence suppressed), the rule is:

1. If there exists an action $a = (p_a, s_a)$ such that
   - $\hat{\mu}_a(T) > 0.5$, and
   - $p_a \neq (1, 2, \ldots, k)$,

   then output $\bar{d}_{\mathrm{Unif}} = p_a$ (if there are multiple such actions, choose one of them arbitrarily).
2. Otherwise, output $\bar{d}_{\mathrm{Unif}} = [k]$.

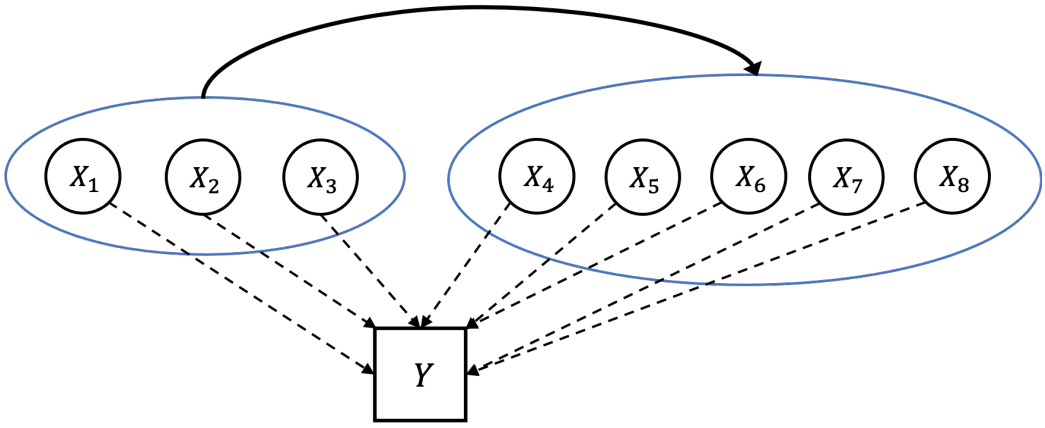


Figure 3: Graph structure for $n = 8$, $k = 3$. Nodes X_1, X_2, X_3 act as parents of all other variables. The reward node Y has dashed incoming edges from all variables, indicating that any subset of size k can form the true parent set Pa_Y.

Consider the uniform sampling policy that samples each arm in $\mathcal{A}_m$ uniformly, with $m = k$, i.e., each $a \in \mathcal{A}_k$ is pulled exactly $T/|\mathcal{A}_k|$ times (up to ± 1). We show that applying $\bar{d}_{\mathrm{Unif}}$ after T rounds yields an error probability $\mathcal{O}\left(\exp\left(-T\right)\right)$ on $\mathcal{E}_0$.

Proposition 9.8. *For the class $\mathcal{E}_0$ described above and $m = k$, the uniform-sampling policy on $\mathcal{A}_k$ combined with the rule $\bar{d}_{Unif}$ satisfies*

$$\delta\left(\pi_{unif}, \bar{d}_{Unif}, \mathcal{E}_0\right) \leq C \exp\left(-cT\right),$$

for some constants $C, c > 0$ that may depend on n, k but not on T.

Proof. Fix an instance $\mathcal{V}_p \in \mathcal{E}_0$ with true parent set $\mathrm{Pa}_Y = p$. Recall from the definition of $\mathcal{E}_0$ (binary variables) that the reward has mean 1 if and only if *all* parents in Pa_Y equal 1, and mean 0 otherwise. Let

$$a_p^\star \;=\; \big(p, \mathbf{1}_k\big), \qquad a_0 \;=\; \big([k], \mathbf{1}_k\big),$$

where $\mathbf{1}_k$ is the all-ones vector in $\{0,1\}^k$. Under $\mathcal{E}_0$, the only actions with $\mu_a = 1$ are $a_p^\star$ and a_0; every other $a \in \mathcal{A}_k$ has $\mu_a = 0$. Indeed, setting $\mathbf{X}_{[k]}$ to 1 forces all X_j, $j > k$, to 1, so a_0 ensures that *any* parent set Pa_Y is all ones; conversely, if an action does not set all coordinates in p to 1 and does not set $\mathbf{X}_{[k]}$ to 1, then at least one parent remains 0 (since $X_i = 0$ for each i unless intervened or the action is a_0), hence $\mu_a = 0$.

Under uniform sampling,

$$N_a(T) \in \left\{ \left\lfloor \tfrac{T}{|\mathcal{A}_k|} \right\rfloor, \left\lceil \tfrac{T}{|\mathcal{A}_k|} \right\rceil \right\} \quad \text{so in particular} \quad N_{\min} \coloneqq \min_{a \in \mathcal{A}_k} N_a(T) \;\geq\; \frac{T - |\mathcal{A}_k|}{|\mathcal{A}_k|}.$$

Since rewards are 1-sub-Gaussian with means $\mu_a \in \{0,1\}$, by Hoeffding's inequality, for any $\epsilon \in (0, 1/2)$,

$$\mathbb{P}\big(\big|\hat{\mu}_a(T) - \mu_a\big| > \epsilon\big) \;\leq\; 2\exp\big(-c_0 N_a \epsilon^2\big) \quad \text{for all } a \in \mathcal{A}_k$$

for some absolute constant $c_0 > 0$. By a union bound over $\mathcal{A}_k$ and the bound on $N_{\min}$,

$$\mathbb{P}\big(\exists a \in \mathcal{A}_k : \big|\hat{\mu}_a(T) - \mu_a\big| > \epsilon\big) \;\leq\; 2|\mathcal{A}_k| \exp\big(-c_0 N_{\min} \epsilon^2\big) \;\leq\; 2|\mathcal{A}_k| \exp\left(-c_0 \epsilon^2 \, \frac{T - |\mathcal{A}_k|}{|\mathcal{A}_k|}\right).$$

Fix $\epsilon = \frac{1}{4}$. For all $T \geq 2|\mathcal{A}_k|$ we have $(T - |\mathcal{A}_k|)/|\mathcal{A}_k| \geq T/(2|\mathcal{A}_k|)$, hence

$$\Pr\big(\exists a \in \mathcal{A}_k : \big|\hat{\mu}_a(T) - \mu_a\big| > \tfrac{1}{4}\big) \;\leq\; C_1 \exp(-c_1 T),$$

with $C_1 = 2|\mathcal{A}_k|$ and $c_1 = c_0/(32|\mathcal{A}_k|)$. (For $T < 2|\mathcal{A}_k|$ the bound can be absorbed into constants.)

Now we show that under the complement of this event, i.e., $\forall a \in \mathcal{A}_k : \left|\hat{\mu}_a(T) - \mu_a\right| \leq \frac{1}{4}$, the parent set is identified correctly by $\bar{d}_{\text{Unif}}$. Under this good event, we have

$$\hat{\mu}_{a_p^\star}(T) \geq 1 - \tfrac{1}{4} > \tfrac{1}{2}, \qquad \hat{\mu}_{a_0}(T) \geq 1 - \tfrac{1}{4} > \tfrac{1}{2}, \qquad \hat{\mu}_a(T) \leq \tfrac{1}{4} < \tfrac{1}{2} \;\; \forall a \notin \{a_p^\star, a_0\}.$$

Case 1: $p \neq [k]$. Since $p_{a_p^\star} = p \neq [k]$, $\hat{\mu}_{a_p^\star}(T) > \frac{1}{2}$, and $\hat{\mu}_a(T) < \frac{1}{2} \;\; \forall a \notin \{a_p^\star, a_0\}$, the decision rule outputs $\bar{d}_{\text{Unif}} = p$.

Case 2: $p = [k]$. Then $a_p^\star = a_0$, and under the good event there is no action with $\hat{\mu}_a(T) > \frac{1}{2}$ and $p_a \neq [k]$. The rule therefore outputs $\bar{d}_{\text{Unif}} = [k] = p$.

In both cases, under the good event the output is correct; thus

$$\mathbb{P}\big(\bar{d}_{\text{Unif}} \neq p\big) \;\leq\; \mathbb{P}\big(\exists a \in \mathcal{A}_k : \left|\hat{\mu}_a(T) - \mu_a\right| > \tfrac{1}{4}\big) \;\leq\; C_1 \exp(-c_1 T)\,.$$

Renaming constants gives the claimed bound with some $C, c > 0$ depending at most on n, k (through $|\mathcal{A}_k| = \binom{n}{k} 2^k$), but not on T. □

We now prove the second statement of the theorem. Fix any policy π. For any set $p \in \binom{[n]}{k}$, recall that $\mathbb{P}_p^\pi$ denotes the probability measure induced by T rounds of interaction between π and $\mathcal{V}_p$. Define the event $\mathcal{E}$ as the event that the estimated parent set of π, denoted by $\hat{Pa}(\pi)$, is equal to $[k]$. Then, by the definition of $\delta(\pi, \bar{d}, \mathcal{E}_0)$, for some decision rule notation $\bar{d}$, we have

$$\forall p \neq [k] : \quad 2\delta(\pi, \bar{d}, \mathcal{E}_0) \;\geq\; \mathbb{P}_{[k]}^\pi\left(\mathcal{E}^c\right) + \mathbb{P}_p^\pi(\mathcal{E}).$$

We omit the decision rule for simpler representation. By Bretagnolle–Huber's inequality 9.5, it follows that

$$\mathbb{P}_{[k]}^\pi\left(\mathcal{E}^c\right) + \mathbb{P}_p^\pi(\mathcal{E}) \;\geq\; \tfrac{1}{2} \exp\left(-\mathrm{KL}\left(\mathbb{P}_{[k]}^\pi, \mathbb{P}_p^\pi\right)\right),$$

and therefore

$$4\delta(\pi, \mathcal{E}_0) \;\geq\; \exp\left(-\mathrm{KL}\left(\mathbb{P}_{[k]}^\pi, \mathbb{P}_p^\pi\right)\right). \tag{10}$$

To compute the KL divergence between $\mathbb{P}_{[k]}^\pi$ and $\mathbb{P}_p^\pi$, observe that the only difference between these two instances arises when the action is $a_p = (p, \mathbf{1}_k)$, i.e., the intervention that sets all variables in p to 1. In this case, the reward distribution is $\mathcal{N}(1,1)$ in $\mathcal{V}_p$ but $\mathcal{N}(0,1)$ in $\mathcal{V}_{[k]}$. For any other action a, the joint distribution of $(X_1, \ldots, X_n, Y)$ is identical under both instances, so the contribution to the KL divergence vanishes.

In this setting, when the algorithm is playing on $\mathcal{V}_{[k]}$, the only informative action for the parent identification task is a_p, since any other action provides no information to distinguish between $\mathcal{V}_{[k]}$ and $\mathcal{V}_p$. However, in $\mathcal{V}_{[k]}$ this action incurs a regret of 1. On the other hand, the only optimal action in $\mathcal{V}_{[k]}$ is $a_0 = ([k], \mathbf{1}_k)$, which provides no information about the true parent set. Since the set of optimal actions and the set of informative actions for parent identification are disjoint, this illustrates the fundamental trade-off between regret minimization and parent identification.

By the divergence decomposition lemma (Lemma 15.1 in [LS20]), we obtain

$$\mathrm{KL}\left(\mathbb{P}_{[k]}^\pi, \mathbb{P}_p^\pi\right) \;=\; \mathbb{E}_{[k]}\big[N_{a_p}(T)\big]\, \mathrm{KL}\left(\mathcal{N}(0,1), \mathcal{N}(1,1)\right),$$

where $\mathbb{E}_{[k]}$ denotes expectation with respect to $\mathbb{P}_{[k]}^\pi$. Since

$$\mathrm{KL}\left(\mathcal{N}(1,1), \mathcal{N}(0,1)\right) = \tfrac{1}{2},$$

we conclude that

$$\forall p \neq [k] : \quad \mathrm{KL}\left(\mathbb{P}_{[k]}^\pi, \mathbb{P}_p^\pi\right) \;=\; \tfrac{1}{2}\, \mathbb{E}_{[k]}\big[N_{a_p}(T)\big].$$

Now suppose $\delta(\pi, \mathcal{E}_0) \in \mathcal{O}\left(\exp(-T^\alpha)\right)$, meaning that there exist constants $C, c > 0$ such that

$$\delta(\pi, \mathcal{E}_0) \leq C \exp(-cT^\alpha).$$

By (10), we then have

$$\begin{aligned} C\exp(-cT^{\alpha}) \;&\geq\; \exp\!\big(-\tfrac{1}{2}\,\mathbb{E}_{[k]}\big[N_{a_p}(T)\big]\big) \\ &\Longrightarrow\; \ln(C) - cT^{\alpha} \;\geq\; -\tfrac{1}{2}\,\mathbb{E}_{[k]}\big[N_{a_p}(T)\big] \\ &\Longrightarrow\; \mathbb{E}_{[k]}\big[N_{a_p}(T)\big] \;\geq\; C'T^{\alpha}, \end{aligned}$$

for some constant $C' > 0$.

Now, the key point is that each time the algorithm plays a_p in $\mathcal{V}_{[k]}$, it incurs a regret of 1, since a_p is strictly suboptimal there. Therefore, the total expected regret is proportional to the number of pulls of a_p. Combining this with the bound above, we obtain

$$R_T(\pi, \mathcal{E}_0) \;\geq\; R_T(\pi, \mathcal{V}_{[k]}) \;\geq\; \mathbb{E}_{[k]}\big[N_{a_p}(T)\big] \times 1 \;\geq\; C'T^{\alpha},$$

which completes the proof.

□

9.4 Proofs of Section 4

Theorem 4.1 ($m \geq k$ Lower Bound)**.** *For any policy $\pi \in \Pi(\{k, \mathcal{G}\})$, any values $n \geq m \geq k$, and any causal graph $\mathcal{G}$:*

$$R_T\big(\pi, \mathcal{E}\big) \in \Omega\left(\sqrt{T \max\left((\ell-1)^k \frac{\binom{n}{k}}{\binom{m}{k}},\; \ell^k\right)}\right). \tag{1}$$

Proof. First note that the empty graph is a subgraph of any graph $\mathcal{G}$. Moreover, for any graph $\mathcal{G}$ over variables $X_1, \ldots, X_n$, we can specify a data-generating distribution that *ignores* any given edge (i.e., the conditional distribution of a child either does not change with its parent or changes arbitrary small), effectively removing that edge. Hence, to prove the desired statement for arbitrary $\mathcal{G}$, it suffices to show it for the empty graph on $\{X_1, \ldots, X_n\}$; the result then extends to any $\mathcal{G}$ by appropriately "turning off" edges. We therefore assume from now on that the graph is empty, so all variables are independent.

Let $\mathcal{V}$ denote the neutral instance in which the graph $\mathcal{G}$ is empty, the reward distribution is $\mathcal{N}(0,1)$ for any values of its parents (indeed, the distribution is constant so the choice of the parent set can be any k-subset), and

$$\forall i \in [n]: \quad \Pr(X_i = 1) = 1,$$

so that each X_i equals 1 unless it is intervened upon. Fix any policy π interacting with $\mathcal{V}$ for T rounds, and let $\mathbb{P}_{\mathcal{V}}$ denote the probability measure induced by this sequential interaction.

For any pair (p, s) with $p \in \binom{[n]}{k}$ and $s \in [\ell]^k$, define a perturbed instance $\mathcal{V}_{p,s}$ which is identical to $\mathcal{V}$ in the graph and in the distributions of the non-reward variables. In $\mathcal{V}_{p,s}$, the reward's parent set is p, and for any assignment $\mathbf{b} \in [\ell]^k$ of the parents we set

$$Y \mid \mathbf{X}_p = \mathbf{b} \sim \mathcal{N}\big(\mu(\mathbf{b}),\, 1\big), \qquad \text{where} \quad \mu(\mathbf{b}) \;=\; \begin{cases} \Delta, & \mathbf{b} = s, \\ 0, & \text{otherwise}, \end{cases}$$

where Δ is a positive real number to be chosen later.

Let $\mathbb{P}_{p,s}$ denote the probability measure obtained by T rounds of interaction between π and $\mathcal{V}_{p,s}$. By setting $P = \mathbb{P}_{\mathcal{V}}, Q = \mathbb{P}_{p,s}$ in Lemma 9.4, and letting $T_{p,s}$ denote the random variable equal to the number of times that $\mathbf{x}_p = s$ during the T rounds, i.e. $T_{p,s} = \sum_{t\in[T]} \mathbb{1}_{\left\{\mathbf{x}_p^{(t)} = s\right\}}$, we obtain

$$\big|\mathbb{E}_{\mathcal{V}}[T_{p,s}] - \mathbb{E}_{p,s}[T_{p,s}]\big| \;\leq\; T\,\delta(\mathbb{P}_{\mathcal{V}}, \mathbb{P}_{p,s}),$$

where we use the fact that $T_{p,s} \in [0, T]$.

Then, by Lemma 9.3, it follows that

$$\mathbb{E}_{p,s}[T_{p,s}] \;\le\; \mathbb{E}_{\mathcal{V}}[T_{p,s}] \;+\; T\sqrt{\tfrac{1}{2}\,\mathrm{KL}\left(\mathbb{P}_{\mathcal{V}}, \mathbb{P}_{p,s}\right)}. \tag{11}$$

Next, note that the observational distribution of all variables is identical in $\mathcal{V}$ and $\mathcal{V}_{p,s}$, except when $\mathbf{X}_p = s$. Therefore, by the divergence decomposition lemma (Lemma 15.1 in [LS20]), we obtain

$$\mathrm{KL}\left(\mathbb{P}_{\mathcal{V}}, \mathbb{P}_{p,s}\right) \;=\; \mathbb{E}_{\mathcal{V}}[T_{p,s}]\,\mathrm{KL}\left(\mathcal{N}(0,1), \mathcal{N}(\Delta,1)\right).$$

Since $\mathrm{KL}\left(\mathcal{N}(0,1), \mathcal{N}(\Delta,1)\right) = \frac{\Delta^2}{2}$, inequality (11) yields

$$\mathbb{E}_{p,s}[T_{p,s}] \;\le\; \mathbb{E}_{\mathcal{V}}[T_{p,s}] \;+\; \frac{T}{2}\sqrt{\mathbb{E}_{\mathcal{V}}[T_{p,s}]\Delta^2}.$$

Fix any arbitrary subset $\mathcal{J} \subseteq \mathcal{P}_k$ consisting of pairs (p,s). Let

$$w_{p,s} \;=\; \frac{1}{T}\,\mathbb{E}_{\mathcal{V}}[T_{p,s}]$$

denote the fraction of time that policy π observes $\mathbf{X}_p = s$ during the process. Then, the above inequality implies

$$\begin{aligned}\sum_{(p,s)\in\mathcal{J}} \mathbb{E}_{p,s}[T_{p,s}] &\le T\sum_{(p,s)\in\mathcal{J}} w_{p,s} \;+\; \frac{T\Delta}{2}\sum_{(p,s)\in\mathcal{J}}\sqrt{T w_{p,s}} \\ &\le T\sum_{(p,s)\in\mathcal{J}} w_{p,s} \;+\; \frac{T\Delta}{2}\sqrt{T\,|\mathcal{J}|\sum_{(p,s)\in\mathcal{J}} w_{p,s}},\end{aligned}$$

where the last inequality follows from the Cauchy–Schwarz inequality.

For any pair (p,s), note that

$$R_T(\pi, \mathcal{V}_{p,s}) \;=\; \Delta\big(T - \mathbb{E}_{p,s}[T_{p,s}]\big).$$

Therefore,

$$\begin{aligned}\sum_{(p,s)\in\mathcal{J}} R_T(\pi, \mathcal{V}_{p,s}) &= \Delta T|\mathcal{J}| - \Delta\sum_{(p,s)\in\mathcal{J}} \mathbb{E}_{p,s}[T_{p,s}] \\ &\ge \Delta T|\mathcal{J}| - \Delta T\sum_{(p,s)\in\mathcal{J}} w_{p,s} - \frac{\Delta^2 T}{2}\sqrt{T\,|\mathcal{J}|\sum_{(p,s)\in\mathcal{J}} w_{p,s}}.\end{aligned}$$

By dividing both sides by $|\mathcal{J}|$, we obtain

$$\exists (p,s)\in\mathcal{J}: \quad R_T(\pi, \mathcal{V}_{p,s}) \ge \Delta T\left(1 - \frac{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}{|\mathcal{J}|} - \frac{\Delta}{2}\sqrt{T\,\frac{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}{|\mathcal{J}|}}\right).$$

Since $\forall (p,s): w_{p,s} \le 1$, we have $\frac{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}{|\mathcal{J}|} \le 1$. For any $\mathcal{J}$, let $c_{\mathcal{J}}$ denote the maximum possible value of $\frac{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}{|\mathcal{J}|}$ over all policies. By setting

$$\Delta \;=\; (1 - c_{\mathcal{J}})\sqrt{\frac{|\mathcal{J}|}{T\sum_{(p,s)\in\mathcal{J}} w_{p,s}}},$$

we obtain

$$\begin{aligned}\exists (p,s)\in\mathcal{J}: \quad R_T(\pi, \mathcal{V}_{p,s}) &\ge (1 - c_{\mathcal{J}})\sqrt{\frac{T|\mathcal{J}|}{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}}\cdot\frac{1 - c_{\mathcal{J}}}{2} \\ &\in \Omega\left((1 - c_{\mathcal{J}})^2\sqrt{\frac{T|\mathcal{J}|}{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}}\right).\end{aligned}$$

Since this holds for any subset $\mathcal{J} \subseteq \mathcal{P}_k$, we arrive at the following general lower bound, which to our knowledge is novel in the literature:

Lemma 9.9 (General Lower Bound). *For any graph $\mathcal{G}$ and any policy $\pi \in \Pi(\{k, \mathcal{G}\})$, we have*

$$R_T(\pi, \mathcal{E}) \;\in\; \Omega\left(\max_{\mathcal{J} \subseteq \mathcal{P}_k} (1 - c_{\mathcal{J}})^2 \sqrt{\frac{T|\mathcal{J}|}{\sum_{(p,s)\in\mathcal{J}} w_{p,s}}}\right), \tag{12}$$

where $c_{\mathcal{J}}$ denotes the maximum possible value of $\frac{1}{|\mathcal{J}|}\sum_{(p,s)\in\mathcal{J}} w_{p,s}$ attainable by any policy.

In the following, we propose specific candidate subsets $\mathcal{J}$ that make the bound in Lemma 9.9 large while keeping $1 - c_{\mathcal{J}}$ constant.

Recall that $w_{p,s}$ is the fraction of time that policy π observes $\mathbf{X}_p = s$ in the interaction with instance $\mathcal{V}$. Due to the construction of this instance, at each round t, the vector $\mathbf{x}^{(t)}$ can have at most m (intervention size) entries different from 1, because any non-intervened variable is always equal to 1. Let

$$\mathcal{B}_m \coloneqq \Big\{\mathbf{b} \in [\ell]^n \;\Big|\; \sum_{i\in[n]} \mathbb{1}_{\{b_i \neq 1\}} \leq m\Big\}$$

denote the set of all possible realizations that can be observed by any algorithm interacting with $\mathcal{V}$.

Then for any set $\mathcal{J}$, we define

$$M(\mathcal{J}) \coloneqq \max_{\mathbf{b}\in\mathcal{B}_m} \sum_{(p,s)\in\mathcal{J}} \mathbb{1}_{\{\mathbf{b}_p = s\}}, \tag{13}$$

which represents the maximum number of pairs $(p, s) \in \mathcal{J}$ that can simultaneously be observed in a single round under any policy. By this definition, we have

$$\sum_{(p,s)\in\mathcal{J}} w_{p,s} = \frac{1}{T}\sum_{t\in[T]}\sum_{(p,s)\in\mathcal{J}} \mathbb{E}\left[\mathbb{1}_{\{\mathbf{x}_p^{(t)} = s\}}\right] \;\leq\; M(\mathcal{J}). \tag{14}$$

For any integer $i \in [k+1]$, define the set $\mathcal{J}_i \subseteq \mathcal{P}_k$ as

$$\mathcal{J}_i \coloneqq \Big\{(p, s) \;\Big|\; p \in \tbinom{[n]}{k},\; \sum_{j\in[k]} \mathbb{1}_{\{s_j = 1\}} < i\Big\}.$$

Thus $\mathcal{J}_i$ consists of pairs (p, s) where p is any k-subset of $[n]$ and s contains fewer than i entries equal to 1. The following lemma provides the values of $|\mathcal{J}_i|$ and $M(\mathcal{J}_i)$ for any i.

Lemma 9.10. *For any $i \in [k]$, the set $\mathcal{J}_i$ has the following properties:*

$$|\mathcal{J}_i| = \binom{n}{k} \ell^k\, \mathbb{P}\big(B(k, 1/\ell) < i\big),$$
$$M(\mathcal{J}_i) = \binom{n}{k} \mathbb{P}\left(HG(n,\, n-m,\, k) < i\right),$$

where $B(k, 1/\ell)$ denotes a Binomial random variable with parameters $(k, 1/\ell)$, and $HG(n, n-m, k)$ denotes a Hypergeometric random variable obtained by drawing k items without replacement from a population of size n with $n - m$ successes.

Proof. For $|\mathcal{J}_i|$, note that p can be any k-subset, so there are $\binom{n}{k}$ possibilities for p. For each p, the number of valid vectors $s \in [\ell]^k$ is

$$\ell^k\, \mathbb{P}\big(\mathrm{B}(k, 1/\ell) < i\big),$$

where $\mathrm{B}(k, 1/\ell)$ denotes a Binomial random variable with parameters $(k, 1/\ell)$. This holds because each entry of s equals 1 with probability $1/\ell$, and we require fewer than i entries equal to 1. Therefore,

$$|\mathcal{J}_i| = \binom{n}{k} \ell^k\, \mathbb{P}\big(\mathrm{B}(k, 1/\ell) < i\big).$$

We now calculate $M(\mathcal{J}_i)$. For any $i \in [k]$, our claim is that the maximizer of (13) corresponds to an action that sets exactly m variables to values different from 1.

Consider an action a that sets r variables to values $\neq 1$ and $m - r$ variables to 1 (the environment sets other variables also equal to 1). Then under this action, the number of pairs (p, s) with fewer than i ones in s is

$$\sum_{j=0}^{i-1} \binom{n-r}{j} \binom{r}{k-j}.$$

This is because to form a pair (p, s) observed under this action: (i) we must choose j coordinates among the $n - r$ coordinates fixed to 1 to appear as 1 in s, (ii) and simultaneously choose $k - j$ coordinates among the r coordinates fixed to non-1 values to appear as non-1 in s. Thus the total number of such pairs is given by the above sum.

Finally, we show that the maximum of the above sum occurs at the maximum possible value of r, which is m.

Claim. For fixed integers n, k, i with $0 \le k \le n$ and $i \in \{1, \dots, k+1\}$, the function

$$S(r) = \sum_{j=0}^{i-1} \binom{n-r}{j} \binom{r}{k-j}, \qquad r = 0, 1, \dots, n,$$

is increasing in r.

Proof of claim. Interpret the sum via a hypergeometric law. Consider a population of n items with r *successes* and $(n - r)$ *failures.* Draw k items uniformly at random without replacement, and let I_r denote the number of failures in the sample. Then

$$\mathbb{P}(I_r = j) = \frac{\binom{n-r}{j}\binom{r}{k-j}}{\binom{n}{k}}, \qquad j = 0, 1, \dots, k,$$

so

$$S(r) = \binom{n}{k} \mathbb{P}(I_r < i)\,.$$

We prove $\mathbb{P}(I_r < i)$ is increasing in r by a coupling argument. Move from r to $r + 1$ by converting one failure in the population into a success (leaving all other items unchanged). Use the *same* random k-subset of indices to form both samples. If the converted item is not selected, the sample composition is unchanged, hence $I_{r+1} = I_r$. If it *is* selected, one failure becomes a success, hence $I_{r+1} = I_r - 1$. In all cases

$$I_{r+1} \;\le\; I_r \quad \text{almost surely.}$$

Therefore, for every threshold t,

$$\mathbb{P}(I_{r+1} < t) \;\ge\; \mathbb{P}(I_r < t).$$

Taking $t = i$ and multiplying by $\binom{n}{k}$ gives $S(r + 1) \ge S(r)$, i.e., $S(r)$ is increasing in r.

Therefore,

$$M(\mathcal{J}_i) = S(m) = \binom{n}{k} \mathbb{P}\left(\mathrm{HG}(n,\, n-m,\, k) < i\right).$$

□

By Lemma 9.10 and Equation (14) we obtain

$$\frac{|\mathcal{J}_i|}{\sum_{(p,s)\in\mathcal{J}_i} w_{p,s}} \;\ge\; \frac{\ell^k\, \mathbb{P}\big(\mathrm{B}(k, 1/\ell) < i\big)}{\mathbb{P}\left(\mathrm{HG}(n,\, n-m,\, k) < i\right)}.$$

It can be shown that the right-hand side as a function of i first decreases and, after some point, increases again; thus the maximizer

$$i^* \;=\; \mathrm{argmax}_{i\in[k+1]} \frac{|\mathcal{J}_i|}{\sum_{(p,s)\in\mathcal{J}_i} w_{p,s}}$$

lies in $\{1,\, k+1\}$. Hence, only these two candidates are sufficient to prove the lower bound.

For $i = 1$, we have

$$\frac{|\mathcal{J}_1|}{\sum_{(p,s)\in\mathcal{J}_1} w_{p,s}} \;\geq\; \frac{\ell^k\ \mathbb{P}\big(\mathrm{B}(k, 1/\ell) < 1\big)}{\mathbb{P}\left(\mathrm{HG}(n,\, n-m,\, k) < 1\right)} \;=\; \frac{\ell^k \left(\frac{\ell-1}{\ell}\right)^k}{\binom{m}{k}/\binom{n}{k}} \;=\; (\ell-1)^k\, \frac{\binom{n}{k}}{\binom{m}{k}}.$$

For $i = k+1$, we obtain

$$\frac{|\mathcal{J}_{k+1}|}{\sum_{(p,s)\in\mathcal{J}_{k+1}} w_{p,s}} \;\geq\; \frac{\ell^k\ \mathbb{P}\big(\mathrm{B}(k, 1/\ell) < k+1\big)}{\mathbb{P}\left(\mathrm{HG}(n,\, n-m,\, k) < k+1\right)} \;=\; \frac{\ell^k \cdot 1}{1} \;=\; \ell^k.$$

Hence, for i^*,

$$\frac{|\mathcal{J}_{i^*}|}{\sum_{(p,s)\in\mathcal{J}_{i^*}} w_{p,s}} \;\geq\; \max\left\{(\ell-1)^k\, \frac{\binom{n}{k}}{\binom{m}{k}},\ \ell^k\right\}.$$

In particular, since $\ell \geq 2$ in our setting, $\ell^k \geq 2$, and therefore

$$1 - c_{\mathcal{J}_{i^*}} \;=\; 1 - \frac{\sum_{(p,s)\in\mathcal{J}_{i^*}} w_{p,s}}{|\mathcal{J}_{i^*}|} \;=\; 1 - \frac{1}{\frac{|\mathcal{J}_{i^*}|}{\sum_{(p,s)\in\mathcal{J}_{i^*}} w_{p,s}}} \;\geq\; 1 - \frac{1}{2} \;=\; \frac{1}{2}.$$

Substituting this result in the bound of Lemma 9.9 yields the desired regret lower bound for any policy $\pi \in \Pi(\{k, \mathcal{G}\})$:

$$R_T\big(\pi, \mathcal{E}\big) \;\in\; \Omega\left(\sqrt{T\, \max\left((\ell-1)^k\, \frac{\binom{n}{k}}{\binom{m}{k}},\ \ell^k\right)}\right),$$

which completes the proof.

□

Theorem 4.2 ($m < k$ Lower Bound)**.** *For any policy $\pi \in \Pi(\{k, \mathcal{G}\})$, any values $n \geq k > m$, and any graph $\mathcal{G}$, we have*

$$R_T\big(\pi, \mathcal{E}\big) \in \Omega\left(\sqrt{T \max\left((\ell-1)^m \binom{n}{m},\ \ell^m\right)}\right). \tag{2}$$

Proof. This theorem follows directly from the previous result. For any $k > m$, consider instances in which $k - m$ of the reward's parents are *neutral*, that is, variables whose values do not affect the distribution of the reward (or whose effect is so small that it cannot be detected within T samples). In such cases, the effective number of true parents is m, and the instance is statistically indistinguishable from one with exactly m parents. Consequently, the worst-case regret of any algorithm on instances with k reward parents must be at least as large as the regret for instances with m parents. Therefore, the lower bound stated in the theorem also holds for all $k > m$. □

Theorem 4.4 (Known k Upper Bound)**.** *The worst-case regret of Algorithm 1 with input k is bounded by*

$$R_T\big(\texttt{Alg.1}[k], \mathcal{E}(n, \ell, k)\big) \in \begin{cases} \tilde{\mathcal{O}}\left(\sqrt{T\ell^k \frac{\binom{n}{k}}{\binom{m}{k}}}\right), & m \geq k, \\ \tilde{\mathcal{O}}\left(\sqrt{T\ell^m \binom{n}{m}}\right), & m < k, \end{cases} \tag{5}$$

where $\tilde{\mathcal{O}}$ hides constants and logarithmic factors.

Proof. We first recall the standard UCB regret bound, which will be used in the analysis of both algorithms.

For any horizon T and any bandit environment $\mathcal{V}$ with N arms, UCB satisfies

$$R_T(\pi_{\mathrm{UCB}}, \mathcal{V}) \;\leq\; C_{\mathrm{UCB}} \sqrt{TN \ln(T)}, \tag{15}$$

where $C_{\text{UCB}} > 0$ is a universal constant. While this bound was originally proved for independent arms, it also holds when dependencies exist among arms. For the proof, see [LS20, BCB+12].

To analyze Algorithm 1, define the event

$$\mathcal{E}_0 = \{\mu^*_{\mathcal{A}'} < \mu_{a^*}\},$$

where $\mathcal{A}'$ is the subset of arms sampled in the algorithm, $\mu^*_{\mathcal{A}'}$ is the maximum mean reward among arms in $\mathcal{A}'$, and a^* is the globally optimal arm.

By the definition of α_k as the fraction of optimal arms in $\mathcal{A}_m$, and the calculation of this value in (8), the algorithm samples exactly $\frac{1}{\alpha_k} \ln \sqrt{T}$ random arms (or all arms in $\mathcal{A}_m$ if this number exceeds $|\mathcal{A}_m|$). By Lemma 9.7, the probability that none of them are optimal satisfies

$$\mathbb{P}(\mathcal{E}_0) \;\le\; \frac{1}{\sqrt{T}}.$$

Putting these together, for any instance $\mathcal{V} \in \mathcal{E}$, we can bound the regret as

$$\begin{aligned} R_T(\texttt{Alg.1}[k], \mathcal{V}) &= \mathbb{P}(\mathcal{E}_0)\; R_T(\texttt{Alg.1}[k], \mathcal{V}) \mid \mathcal{E}_0 + \mathbb{P}(\mathcal{E}_0^c)\; R_T(\texttt{Alg.1}[k], \mathcal{V}) \mid \mathcal{E}_0^c \\ &\le T \cdot \mathbb{P}(\mathcal{E}_0) + R_T(\texttt{Alg.1}[k], \mathcal{V}) \mid \mathcal{E}_0^c \\ &\le \sqrt{T} + C_{\text{UCB}} \sqrt{T\,|\mathcal{A}'| \ln(T)}, \end{aligned}$$

where with a slight abuse of notation, we use $R_T(\texttt{Alg.1}[k], \mathcal{V}) \mid \mathcal{E}_0$ to denote the expected regret under the event $\mathcal{E}_0$.

Finally, substituting the size of $\mathcal{A}'$, i.e., n_0 in Algorithm 1, yields

$$R_T(\texttt{Alg.1}[k], \mathcal{V}) \;\le\; \begin{cases} 2C_{\text{UCB}} \sqrt{\frac{1}{2}\, T\, \ell^k \frac{\binom{n}{k}}{\binom{m}{k}}\, \ln^2(T)} \;\in\; \tilde{\mathcal{O}}\left(\sqrt{T\, \ell^k \frac{\binom{n}{k}}{\binom{m}{k}}}\right), & m \ge k, \\ 2C_{\text{UCB}} \sqrt{T\, \ell^m \binom{n}{m} \ln(T)} \;\in\; \tilde{\mathcal{O}}\left(\sqrt{T\, \ell^m \binom{n}{m}}\right), & m < k, \end{cases}$$

which matches the theorem statement and completes the proof. □

9.5 Proofs of Section 5

Theorem 5.2 (Unknown k Lower Bound)**.** *For any causal graph $\mathcal{G}$, any policy $\pi \in \Pi(\mathcal{G})$, and any values $n \ge m \ge k_2 > k_1$, we have*

$$\begin{aligned} &R_T\big(\pi, \mathcal{E}(n, \ell, k_1)\big) \times R_T\left(\pi, \mathcal{E}(n, \ell, k_2)\right) \\ &\in \Omega\left(T \max\left((\ell - 1)^{k_2} \frac{\binom{n-k_1}{k_2-k_1}}{\binom{m-k_1}{k_2-k_1}},\; \ell^{k_2}\right)\right). \end{aligned} \tag{6}$$

Proof. Analogous to the proof of the lower-bound in the previous section, since the empty graph is a subgraph of any $\mathcal{G}$, it suffices to establish the result when $\mathcal{G}$ is empty; the general case then follows immediately.

Fix values k_1, k_2 such that $k_1 < k_2 \le m \le n$. Consider the instance $\mathcal{V} \in \mathcal{E}(n, \ell, k_1)$ defined as follows. The graph $\mathcal{G}$ over the non-reward variables is the empty graph, and the parent set of the reward is $\text{Pa}_{\mathcal{G}}(Y) = \{X_1, X_2, \ldots, X_{k_1}\}$. Since $\mathcal{G}$ is empty, the variables X_i are mutually independent. Moreover, each variable X_i satisfies $\mathbb{P}(X_i = 1) = 1$, meaning that every variable equals 1 unless it is intervened upon.

For any vector $\mathbf{b} \in [\ell]^{k_1}$ representing the values of the reward's parents, the reward distribution is

$$Y \mid \text{Pa}_{\mathcal{G}}(Y) = \mathbf{b} \;\sim\; \mathcal{N}(\mu_{\mathbf{b}}, 1),$$

where

$$\mu_{\mathbf{b}} = \begin{cases} \Delta, & \text{if } \mathbf{b} = (1, 1, \ldots, 1), \\ 0, & \text{otherwise}, \end{cases}$$

and $\Delta > 0$ is a constant that will be specified later.

To define the family of alternative instances, we define $\mathcal{M}$ to be the set of all pairs (p, s) satisfying:

(i) $p \in \binom{[n]}{k_2}$,

(ii) $[k_1] \subset p$,

(iii) $s \in [\ell]^{k_2}$,

(iv) $\sum_{i=1}^{k_1} \mathbb{1}_{s_i=1} < k_1$ (i.e., at least one of the first k_1 entries of s is not 1),

Here, p represents a set of k_2 indices containing the first k_1 parent indices, and s assigns values to those indices such that at least one of the first k_1 entries is not 1.

For each $(p, s) \in \mathcal{M}$, define an alternative instance $\mathcal{V}_{p,s} \in \mathcal{E}(n, \ell, k_2)$ as follows. The graph and the distributions of all non-reward variables are identical to those in $\mathcal{V}$. For any $\mathbf{b} \in [\ell]^{k_2}$ representing the values of the parents of Y, define

$$Y \mid \mathrm{Pa}_{\mathcal{G}}(Y) = \mathbf{b} \;\sim\; \mathcal{N}(\mu_{\mathbf{b}}, 1),$$

where

$$\mu_{\mathbf{b}} = \begin{cases} \Delta, & \text{if } b_1 = b_2 = \ldots = b_{k_1} = 1, \\ 2\Delta, & \text{if } \mathbf{b} = s, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $\mathcal{V}_{p,s}$ differs from $\mathcal{V}$ only in the set of parents and only in the reward distribution corresponding to the interventions that set $\mathbf{X}_p = s$. Formally, for all other interventions, the joint distribution of all variables $(X_1, \ldots, X_n, Y)$ are identical, while for actions setting $\mathbf{X}_p = s$, they differ only in the reward distribution.

Fix a policy π, and let $\mathbb{P}_{\mathcal{V}}$ and $\mathbb{P}_{p,s}$ denote the probability measures induced by T rounds of interaction between π and $\mathcal{V}$, $\mathcal{V}_{p,s}$, respectively. Define $T_{p,s}$ as the random variable counting the number of rounds t during which $\mathbf{x}_p^{(t)} = s$, and let $w_{p,s} = \mathbb{E}_{\mathcal{V}}[T_{p,s}]/T$ denote its expectation fraction under $\mathbb{P}_{\mathcal{V}}$.

By the divergence decomposition lemma (Lemma 15.1 in [LS20]), we obtain

$$\mathrm{KL}\left(\mathbb{P}_{\mathcal{V}}, \mathbb{P}_{p,s}\right) \;=\; \mathbb{E}_{\mathcal{V}}[T_{p,s}]\,\mathrm{KL}\left(\mathcal{N}(0,1), \mathcal{N}(2\Delta, 1)\right) \;=\; 2\,T w_{p,s}\Delta^2,$$

where we used $\mathrm{KL}\left(\mathcal{N}(0,1), \mathcal{N}(2\Delta,1)\right) = 2\Delta^2$ in the last equality.

Now, we define the event E as

$$E = \left\{ \sum_{t \in [T]} \mathbb{1}_{\{\mathbf{x}_{[k_1]}^{(t)} = (1,1,\ldots,1)\}} \geq \frac{T}{2} \right\}.$$

which captures the case where, in at least half of the rounds, the first k_1 variables are all equal to 1. Then

$$\begin{aligned} R_T(\pi, \mathcal{V}_{p,s}) | E &\geq \frac{T\Delta}{2} \qquad \forall (p,s) \in \mathcal{M}, \\ R_T(\pi, \mathcal{V}) | E^c &\geq \frac{T\Delta}{2}, \end{aligned}$$

where with a slight abuse of notation, $R_T(\pi, \mathcal{V})|E$ represents the expected regret given the event E. This implies that

$$R_T(\pi, \mathcal{V}_{p,s}) + R_T(\pi, \mathcal{V}) \geq \frac{T\Delta}{2}\left(\mathbb{P}_{p,s}(E) + \mathbb{P}_{\mathcal{V}}(E^c)\right).$$

Then, by Bretagnolle-Huber inequality 9.5, we obtain

$$\mathbb{P}_{p,s}(E) + \mathbb{P}_{\mathcal{V}}(E^c) \geq \frac{1}{2}\exp\left(\mathrm{KL}\left(\mathbb{P}_{\mathcal{V}}, \mathbb{P}_{p,s}\right)\right) = \frac{1}{2}\exp\left(-2\,T w_{p,s}\Delta^2\right).$$

Now since $\mathcal{E}(n,\ell,k_1) \subseteq \mathcal{E}(n,\ell,k_2)$, we have for any policy π, $R_T(\pi,\mathcal{E}(n,\ell,k_1)) \le R_T(\pi,\mathcal{E}(n,\ell,k_2))$, then

$$\begin{aligned}
2R_T(\pi,\mathcal{E}(n,\ell,k_2)) &\ge R_T(\pi,\mathcal{E}(n,\ell,k_1)) + R_T(\pi,\mathcal{E}(n,\ell,k_2)) \\
&\ge R_T(\pi,\mathcal{V}) + R_T(\pi,\mathcal{V}_{p,s}) \ge \frac{T\Delta}{4}\exp\left(-2\,T w_{p,s}\Delta^2\right) \\
&\implies \ln\left(8R_T(\pi,\mathcal{E}(n,\ell,k_2))\right) - \ln\left(T\Delta\right) \ge -2\,T w_{p,s}\Delta^2 \\
&\implies 2\,T w_{p,s}\Delta^2 \ge \ln\left(\frac{T\Delta}{8R_T(\pi,\mathcal{E}(n,\ell,k_2))}\right).
\end{aligned}$$

Then letting $\Delta = \frac{16R_T(\pi,\mathcal{E}(n,\ell,k_2))}{T}$, we obtain

$$\begin{aligned}
&\frac{2w_{p,s}R_T(\pi,\mathcal{E}(n,\ell,k_2))^2}{T} \ge \ln(2), \\
&\implies w_{p,s} \ge \ln(\sqrt{2})\frac{T}{R_T(\pi,\mathcal{E}(n,\ell,k_2))^2}.
\end{aligned} \tag{16}$$

Thus we have derived a lower bound on $w_{p,s}$, the expected fraction of rounds (under $\mathcal{V}$) in which π observes the low-reward configuration $\mathbf{x}_p = s$. Note that without loss of generality, we can assume that $\Delta \le 1$. This is because when $\delta > 1$, then the regret of π exceeds $T/16$, which clearly satisfies the proposed lower bound in this theorem.

Now, to bound the regret using this lower bound, let $Q = \{2,3,\ldots,\ell\}^{k_1}$. For any $\mathbf{q} \in Q$, let $w_{\mathbf{q}}$ denote the expected fraction of rounds t in which $\mathbf{x}^{(t)}_{[k_1]} = \mathbf{q}$ during the interaction between π and $\mathcal{V}$. In this case, the regret can be bounded as

$$R_T(\pi,\mathcal{V}) \;\ge\; \Delta \sum_{\mathbf{q}\in Q} T\,w_{\mathbf{q}}. \tag{17}$$

To bound $w_{\mathbf{q}}$, for each $\mathbf{q} \in Q$ define $\mathcal{M}_{\mathbf{q}}$ as the set of pairs $(p,s) \in \mathcal{M}$ with $s_{[k_1]} = \mathbf{q}$. Then

$$|\mathcal{M}_{\mathbf{q}}| \;=\; \binom{n-k_1}{k_2-k_1}(\ell-1)^{\,k_2-k_1},$$

because the indices $\{1,2,\ldots,k_1\}$ must be contained in p and their values are fixed to $\mathbf{q}$, leaving $\binom{n-k_1}{k_2-k_1}$ choices for the remaining indices and $(\ell-1)^{k_2-k_1}$ choices for their non-one values.

On the other hand, for any $\mathbf{b} \in [\ell]^n$, the maximum number of pairs $(p,s) \in \mathcal{M}_{\mathbf{q}}$ such that $\mathbf{b}_p = s$ is $\binom{m-k_1}{k_2-k_1}$. Indeed, under $\mathcal{V}$, any non-intervened variable equals 1, while for $(p,s) \in \mathcal{M}$ all entries of s beyond the first k_1 are non-one. Thus, the only way to realize $\mathbf{b}_p = s$ is to (i) intervene to set the first k_1 variables to $\mathbf{q}$ (none of which is 1) and (ii) set $m-k_1$ additional variables to non-one values, yielding at most $\binom{m-k_1}{k_2-k_1}$ distinct matches.

Combining this counting argument with the lower bound in (16) gives, for each $\mathbf{q} \in Q$,

$$w_{\mathbf{q}} \;\ge\; \frac{\sum_{(p,s)\in\mathcal{M}_{\mathbf{q}}} w_{p,s}}{\max_{\mathbf{b}\in[\ell]^n}\sum_{(p,s)\in\mathcal{M}_{\mathbf{q}}}\mathbb{1}_{\{\mathbf{b}_p=s\}}} \;\ge\; \frac{\frac{\ln(\sqrt{2})\,T\,|\mathcal{M}_{\mathbf{q}}|}{R_T(\pi,\mathcal{E}(n,\ell,k_2))^2}}{\binom{m-k_1}{k_2-k_1}} \;\ge\; c\,\frac{T\,(\ell-1)^{\,k_2-k_1}\binom{n-k_1}{k_2-k_1}}{R_T(\pi,\mathcal{E}(n,\ell,k_2))^2\binom{m-k_1}{k_2-k_1}}, \tag{18}$$

for a universal constant $c > 0$. Plugging (18) into (17) yields

$$\begin{aligned}
R_T(\pi,\mathcal{E}(n,\ell,k_1)) \ge R_T(\pi,\mathcal{V}) \;&=\; T\Delta\sum_{\mathbf{q}\in Q} w_{\mathbf{q}} \\
&\ge 16c\,|Q|\,T\,\frac{R_T(\pi,\mathcal{E}(n,\ell,k_2))}{T}\,\frac{T\,(\ell-1)^{\,k_2-k_1}\binom{n-k_1}{k_2-k_1}}{R_T(\pi,\mathcal{E}(n,\ell,k_2))^2\binom{m-k_1}{k_2-k_1}} \\
&\ge 16c\,\frac{T\,(\ell-1)^{\,k_2}\binom{n-k_1}{k_2-k_1}}{R_T(\pi,\mathcal{E}(n,\ell,k_2))\binom{m-k_1}{k_2-k_1}} \\
\implies R_T(\pi,\mathcal{E}(n,\ell,k_1))R_T(\pi,\mathcal{E}(n,\ell,k_2)) \;&\in\; \Omega\left(T\,(\ell-1)^{\,k_2}\,\frac{\binom{n-k_1}{k_2-k_1}}{\binom{m-k_1}{k_2-k_1}}\right)
\end{aligned} \tag{19}$$

In the above, we use the fact that $|Q| = (\ell - 1)^{k_1}$. This proves the first term inside the max in the theorem.

We now show that the product is also lower bounded by the second term, namely $T\,\ell^{k_2}$. For this, let

$$R = \{(p_0, s) \in \mathcal{M} \mid p_0 = [k_2]\} \subseteq \mathcal{M}.$$

Then, each instance $\mathcal{V}_{p_0,s}$ in R corresponds to the setting where the reward's parents are the first k_2 nodes, and the reward mean is 2Δ whenever their values are equal to s. Since $R \subseteq \mathcal{M}$, for each $(p_0, s) \in R$, the lower bound in (16) also holds:

$$w_{p_0,s} \;\geq\; c\,\frac{T}{R_T(\pi, \mathcal{E}(n,\ell,k_2))^2}.$$

Now, because each $\mathcal{V}_{p_0,s}$ differs from $\mathcal{V}$ only when $\mathbf{x}_{[k_2]} = s$, and there are $|R| = (\ell^{k_1} - 1)\ell^{k_2 - k_1}$ such distinct configurations s with $(p_0, s) \in \mathcal{M}$, the total expected fraction of rounds in which π observes one of these configurations when interacting with $\mathcal{V}$ is at least

$$\sum_{(p_0,s)\in R} w_{p_0,s} \;\geq\; c\,|R|\,\frac{T}{R_T(\pi, \mathcal{E}(n,\ell,k_2))^2} \geq c\,\frac{\ell^{k_2}}{2}\,\frac{T}{R_T(\pi, \mathcal{E}(n,\ell,k_2))^2}$$

Moreover, by construction of $\mathcal{V}$, every such configuration $\mathbf{x}_{[k_2]} = s$ with at least one non-one value in the first k_1 entries corresponds to a suboptimal mean reward. Thus, the regret on $\mathcal{V}$ can be written as

$$R_T(\pi, \mathcal{V}) = \Delta\, T \sum_{(p_0,s)\in R} w_{p_0,s} \;\geq\; \frac{c}{2}\,\Delta\, T^2\,\frac{\ell^{k_2}}{R_T(\pi, \mathcal{E}(n,\ell,k_2))^2}.$$

Finally, substituting $\Delta = \frac{16 R_T(\pi, \mathcal{E}(n,\ell,k_2))}{T}$ as before yields

$$R_T(\pi, \mathcal{E}(n,\ell,k_1)) \;\geq\; R_T(\pi, \mathcal{V}) \;\geq\; c'\,\frac{T\,\ell^{k_2}}{R_T(\pi, \mathcal{E}(n,\ell,k_2))},$$

which implies

$$R_T(\pi, \mathcal{E}(n,\ell,k_1))\, R_T(\pi, \mathcal{E}(n,\ell,k_2)) \;\in\; \Omega\big(T\,\ell^{k_2}\big).$$

This establishes the second term in the max expression of the theorem and completes the proof.

□

Theorem 5.3 (Unknown k Upper Bound)**.** *For any k, the worst-case regret of Algorithm 2 on all the instances in $\mathcal{E}(n,\ell,k)$ is upper bounded as follows*

$$R_T\left(\texttt{Alg.2}, \mathcal{E}(n,\ell,k)\right) \in \begin{cases} \tilde{\mathcal{O}}\left(\sqrt{T\tfrac{m}{n}}\,\ell^{k-\frac{1}{2}}\,\frac{\binom{n}{k}}{\binom{m}{k}}\right), m \geq k, \\ \tilde{\mathcal{O}}\left(\sqrt{T\tfrac{m}{n}}\,\ell^{m-\frac{1}{2}}\binom{n}{m}\right), m < k, \end{cases} \tag{7}$$

where $\tilde{\mathcal{O}}$ hides constants and logarithmic factors.

Proof. We first introduce a few notations. Recall the definition of α_k as the fraction of optimal arms in $\mathcal{A}_m$ when $|\mathrm{Pa}_{\mathcal{G}}(Y)| = k$, whose value is given in (8). We define

$$N_k \;=\; \frac{1}{\alpha_k}\ln\sqrt{T}.$$

Fix a value of k with $k \leq m$ and an arbitrary instance $\mathcal{V} \in \mathcal{E}(n,\ell,k)$. Let R_i denote the regret of Algorithm 2 on $\mathcal{V}$ during phase i, $T_i = \sum_{j=1}^{i} \Delta T_j$, and let $r = \left\lceil \log_2 \sqrt{T} \right\rceil$.

Note that, since we introduce the mixture arms as new arms after phase one, the assumptions that (i) the reward distributions under different actions are independent and (ii) each is 1-sub-Gaussian, no longer hold. Indeed, the

reward of a mixture arm is a mixture of several sub-Gaussian variables corresponding to previously played arms, and these mixture components may overlap across different arms, introducing correlations.

To address this, we make two observations. First, the regret upper bound of the UCB algorithm remains valid even when the arms are correlated. Second, by Lemma 9.6, the reward distribution of any mixture arm is itself sub-Gaussian with parameter $\sigma^2 = \frac{5}{4}$. This increased sub-Gaussian parameter scales the regret bound of UCB by a factor of $\frac{5}{4}$, which we absorb into the UCB constant C_{UCB} introduced in (15).

We first show that Algorithm 2 is well-defined, i.e., it executes all T rounds. Since it runs ΔT_i rounds in phase i, the total number of rounds is

$$\begin{aligned}\sum_{i\in[i_f]} \Delta T_i &= 2^r \left\lceil \frac{\ell n}{m} \right\rceil \sum_{i\in[i_f]} 2^i \\ &= 2^r \left\lceil \frac{\ell n}{m} \right\rceil (2^{i_f+1} - 2) \;\geq\; 2^{2i_f} \left\lceil \frac{\ell n}{m} \right\rceil \;\geq\; T \frac{m}{\ell n} \left\lceil \frac{\ell n}{m} \right\rceil \;\geq\; T,\end{aligned}$$

where we used $r \geq i_f$ and $i_f = \left\lceil \log_2 \sqrt{T \frac{m}{\ell n}} \right\rceil$.

For each phase i, let $\mathcal{F}_{i-1}$ denote the information available at the start of phase i (i.e., all previous observations and the random arms selected in S_i). We decompose the regret in phase i into two components:

$$\begin{aligned}R_i = \Delta T_i \mu^* - \sum_{t=T_{i-1}+1}^{T_i} \mathbb{E}[\mu_{a_t}] &= \Delta T_i(\mu^* - \mu_i^*) \;+\; \left(\Delta T_i \mu_i^* - \sum_{t=T_{i-1}+1}^{T_i} \mathbb{E}[\mu_{a_t}]\right) \\ &= R_i^{(1)} + R_i^{(2)},\end{aligned}$$

where $\mu_i^* = \max_{a\in S_i \cup M} \mu_a$ is the best mean reward among arms considered in phase i.

Bounding $R_i^{(2)}$. Since each phase runs a standard UCB subroutine, the learning term $R_i^{(2)}$ is bounded by the UCB regret upper bound (15):

$$\begin{aligned}\mathbb{E}\left[R_i^{(2)}\right] &\leq C_{\mathrm{UCB}} \sqrt{\Delta T_i |S_i \cup M| \ln(\Delta T_i)} \\ &\leq C_{\mathrm{UCB}} \sqrt{2^{r+i} \left\lceil \frac{\ell n}{m} \right\rceil (2^{r-i+1} + i - 1) \ln(T)} \\ &\leq C_{\mathrm{UCB}} \sqrt{2^{2r+1} \left\lceil \frac{\ell n}{m} \right\rceil \ln(T) \log_2(T)} \\ &\leq C_{\mathrm{UCB}} \sqrt{16\, T \left\lceil \frac{\ell n}{m} \right\rceil \ln^2(T)} \;\leq\; 4\sqrt{2}\, C_{\mathrm{UCB}} \sqrt{T \left(\frac{\ell n}{m}\right) \ln^2(T)} \;\in\; \tilde{\mathcal{O}}\left(\sqrt{T\left(\frac{\ell n}{m}\right)}\right), \qquad (20)\end{aligned}$$

where in the second inequality, we use the fact that $|S_i \cup M| \leq 2^{r-i+1} + (i-1)$ and in the third inequality, we used

$$\begin{aligned}\forall i \in [i_f]: \quad 2^{r+i}(2^{r-i+1} + i - 1) < 2^{2r+1} + 2^{2r} i_f \;&\leq\; 2^{2r+1} + 2^{2r} \log_2(T) \\ &\leq 2^{2r+1} \log_2(T),\end{aligned}$$

using $i_f = \left\lceil \log_2 \sqrt{T \frac{m}{\ell n}} \right\rceil \leq \log_2(T)$, which holds for $T > 3$. The next inequality follows from

$$\begin{aligned}2^{2r} &\leq 2^{2(\log_2 \sqrt{T}+1)} = 4T, \\ \log_2(T) &\leq 2 \ln(T),\end{aligned}$$

and finally, we use $\lceil x \rceil \leq 2x$ for $x \geq 1$.

Bounding $R_i^{(1)}$. To bound the term $R_i^{(1)}$, let

$$i^* = \max\left\{ i \in [i_f] \;\middle|\; q_i \geq N_k \right\}.$$

If no such i exists, we will have

$$N_k = \frac{1}{\alpha_k} \ln \sqrt{T} = \frac{\ell^k \binom{n}{k}}{\binom{m}{k}} \ln \sqrt{T} > q_1 = 2^r \geq \sqrt{T},$$

which implies

$$\frac{\ell^k \binom{n}{k}}{\binom{m}{k}} \sqrt{T \frac{m}{\ell n}} > T \frac{\sqrt{\frac{m}{\ell n}}}{\ln \sqrt{T}}.$$

In the above inequality, the left-hand side corresponds to the target regret upper bound, while the right-hand side grows linearly in T, ignoring the logarithmic factor. In this case, the bound trivially holds. Thus, for the remainder, we may assume that i^* exists.

For each i, define the event

$$\mathcal{E}_i = \{\text{no optimal arm is contained in } S_i\}.$$

By Lemma 9.7, for any $i \leq i^*$, since $q_i \geq N_k$ (recall that q_i is decreasing in i), we have $\mathbb{P}(\mathcal{E}_i) \leq \frac{1}{\sqrt{T}}$. Under the complement event $\mathcal{E}_i^c$, at least one optimal arm is included in S_i, implying that $\mu_i^* = \mu^*$ and therefore $R_i^{(1)} = 0$.

Consequently, the expected contribution of $R_i^{(1)}$ for phases $i \leq i^*$ can be bounded as

$$\mathbb{E}\left[R_i^{(1)}\right] = \Delta T_i \, \mathbb{E}[\mu^* - \mu_i^*] \leq \Delta T_i \, (\mu^* - \mu_{\min}) \, \mathbb{P}(\mathcal{E}_i) \leq \Delta T_i \, \mathbb{P}(\mathcal{E}_i) \;\leq\; \frac{\Delta T_i}{\sqrt{T}} \;\leq\; \sqrt{T}, \tag{21}$$

where $\mu_{\min}$ denotes the smallest mean reward among all arms, and the second inequality follows from the boundedness of rewards in $[0, 1]$.

For each $i > i^*$, note that the mixture arm $\tilde{a}_{i^*}$, constructed at the end of phase i^*, is included in the arm set for phase i. Hence,

$$\mathbb{E}\Big[R_i^{(1)}\Big] = \Delta T_i \, \mathbb{E}[\mu^* - \mu_i^*] \;\leq\; \Delta T_i \, \mathbb{E}[\mu^* - \mu_{\tilde{a}_{i^*}}] = \Delta T_i \, (\mu^* - \mathbb{E}[\mu_{\tilde{a}_{i^*}}]) = \Delta T_i \, \frac{\mathbb{E}[R_{i^*}^{(2)}]}{\Delta T_{i^*}},$$

where the last equality holds because the regret of the mixture arm equals the average regret of the actions played during phase i^*.

By substituting the bound from (20) and using the fact that $\Delta T_i \leq T$, we obtain

$$\mathbb{E}\Big[R_i^{(1)}\Big] \leq \Delta T_i \, \frac{4 \, C_{\mathrm{UCB}} \sqrt{T \left\lceil \frac{\ell n}{m} \right\rceil \ln^2(T)}}{\left\lceil \frac{\ell n}{m} \right\rceil 2^{r+i^*}} \;\leq\; 4 \, T \, C_{\mathrm{UCB}} \sqrt{\frac{T \ln^2(T)}{\left\lceil \frac{\ell n}{m} \right\rceil 2^{2r+2i^*}}} \;\leq\; 4 \, T \, C_{\mathrm{UCB}} \sqrt{\frac{\ln^2(T)}{\left\lceil \frac{\ell n}{m} \right\rceil 2^{2i^*}}},$$

where the last inequality uses $r \geq \log_2(\sqrt{T})$, implying $2^{2r} \geq T$.

From the definition of i^*, we have

$$N_k > q_{i^*+1} = 2^{r-i^*} \quad \& \quad q_{i^*} = 2^{r-i^*+1} \geq N_k \quad \Longrightarrow \quad 2^{i^*} N_k > 2^r \geq \sqrt{T},$$

which gives

$$2^{2i^*} > \frac{T}{N_k^2} = \frac{T \alpha_k^2}{\ln(T)}.$$

Substituting this into the previous bound yields

$$\mathbb{E}\Big[R_i^{(1)}\Big] \leq 4 \, T \, C_{\mathrm{UCB}} \sqrt{\frac{\ln^3(T)}{\left\lceil \frac{\ell n}{m} \right\rceil T \alpha_k^2}} = 4 \, C_{\mathrm{UCB}} \, \frac{1}{\alpha_k} \sqrt{\frac{T \ln^3(T)}{\left\lceil \frac{\ell n}{m} \right\rceil}} \in \tilde{\mathcal{O}}\left(\sqrt{T \, \frac{m}{n}} \, \ell^{k - \frac{1}{2}} \, \frac{\binom{n}{k}}{\binom{m}{k}} \right). \tag{22}$$

Finally, combining (20), (21), and (22), we obtain

$$\forall i \in [i_f] : \quad \mathbb{E}[R_i] \in \tilde{\mathcal{O}}\left(\sqrt{T \, \frac{m}{n}} \, \ell^{k - \frac{1}{2}} \, \frac{\binom{n}{k}}{\binom{m}{k}} \right).$$

Since the number of phases i_f is logarithmic in problem parameters, the same upper bound holds for the total regret (up to logarithmic factors hidden in the $\tilde{\mathcal{O}}$ notation).

This completes the proof for the case $k \leq m$. For $k > m$, all the proof steps hold by setting $k = m$, resulting in the regret bound presented in the theorem.

□

Lemma 5.4 (Pareto Optimality of Algorithm 2)**.** *The following statements hold:*

- *When $m = n$, Algorithm 2 is rate-Pareto optimal for $\Pi(\{\mathcal{G}\})$, up to logarithmic factors.*
- *In the general case, if $\ell \in \Omega(m)$, then the regret vector*

$$\Big[R_T\big(\texttt{Alg.2}, \mathcal{E}(n,\ell,1)\big),\ R_T\big(\texttt{Alg.2}, \mathcal{E}(n,\ell,2)\big)\tfrac{m}{n},$$
$$\ldots,\ R_T\big(\texttt{Alg.2}, \mathcal{E}(n,\ell,m)\big)\tfrac{m}{n}\Big]$$

is rate-Pareto optimal for $\Pi(\{\mathcal{G}\})$. Up to logarithmic terms, this vector coincides with the regret vector of Algorithm 2 when $k = 1$, and exhibits a multiplicative gap of $\frac{m}{n}$ for larger values of k.

Proof. First, consider the case $m = n$. In this setting, ignoring logarithmic factors, the regret vector of Algorithm 2 is

$$\mathbf{R} = \Big[\sqrt{T}\,\ell^{\frac{1}{2}},\ \sqrt{T}\,\ell^{\frac{3}{2}},\ \ldots,\ \sqrt{T}\,\ell^{m-\frac{1}{2}}\Big].$$

For the sake of contradiction, assume there exists a policy π with regret vector $\mathbf{R}'$ that rate-Pareto dominates $\mathbf{R}$.

Note that for $k = 1$, the lower bound in Theorem 4.1 establishes that the optimal rate is $\Theta(\sqrt{T\ell})$, which coincides with the first entry of $\mathbf{R}$. Hence, $\mathbf{R}'$ cannot achieve a strictly better rate for $k = 1$, implying that $R'_1 = \Theta(R_1)$. This means there exists a $k' > 1$ and a constant C such that

$$R'_{k'} \leq C R_{k'} = C\sqrt{T}\,\ell^{k'-\frac{1}{2}},$$

and the inequality does not hold in the reverse direction (i.e., $\mathbf{R}'$ improves over $\mathbf{R}$ at k'). However, by Theorem 5.2, there exists a constant C' such that

$$R'_1\,R'_{k'} \ \geq\ C'T\,\ell^{k'} \quad\Longrightarrow\quad R'_{k'} \ \geq\ C''\sqrt{T}\,\ell^{k'-\frac{1}{2}},$$

for some universal constant C''. This shows that $R_{k'}$ and $R'_{k'}$ have the same rate, which contradicts the assumption that $\mathbf{R}'$ rate-Pareto dominates $\mathbf{R}$. Therefore, Algorithm 2 is rate-Pareto optimal in this setting.

Now consider the general case. If $\ell \in \Omega(m)$, then $\frac{\ell^m}{(\ell-1)^m} \in \mathcal{O}(1)$. Ignoring logarithmic and constant factors, the regret vector $\mathbf{R}$ from the theorem satisfies:

$$R_1 = \sqrt{T\frac{(\ell-1)n}{m}}, \qquad R_k = \sqrt{T\frac{m}{n}}\,\frac{m}{n}(\ell-1)^{k-\frac{1}{2}}\frac{\binom{n}{k}}{\binom{m}{k}}, \quad \text{for } k > 1.$$

Suppose again that there exists a policy π with regret vector $\mathbf{R}'$ that rate-Pareto dominates $\mathbf{R}$. Since Algorithm 2 is optimal for $k = 1$, we have $R'_1 \in \mathcal{O}(R_1)$. Therefore, there must exist a $k' > 1$ and a constant C such that $R'_{k'} \leq C R_{k'}$. By Theorem 5.2, there exists a constant C' satisfying

$$R'_1 R'_{k'} \ \geq\ C'T(\ell-1)^{k'}\frac{\binom{n-1}{k'-1}}{\binom{m-1}{k'-1}} \quad\Longrightarrow\quad R'_{k'} \ \geq\ C''\sqrt{T\frac{m}{n}}\,\frac{m}{n}(\ell-1)^{k'-\frac{1}{2}}\frac{\binom{n}{k'}}{\binom{m}{k'}},$$

where C'' is a universal constant and we used

$$\frac{\binom{n-1}{k'-1}}{\binom{m-1}{k'-1}} = \frac{\binom{n}{k'}}{\binom{m}{k'}}\frac{m}{n}.$$

Since this lower bound matches the rate of $R_{k'}$, we again reach a contradiction, proving that regret vector $\mathbf{R}$ is rate-Pareto optimal.

□

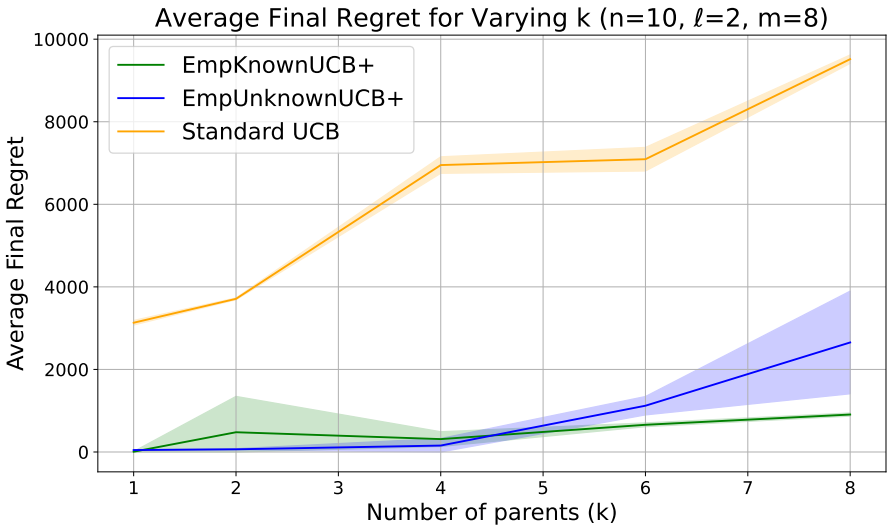


Figure 4: Average cumulative regret of algorithms at time T for varying numbers of reward parents $k \in 1, 2, 4, 6, 8$ on random instances with $n = 10$, $\ell = 2$, $m = 8$, and $T = 30{,}000$. As k increases, the environment becomes more complex, leading to higher regret. Our proposed algorithms consistently achieve lower regret across all settings.

10 ADDITIONAL EXPERIMENTS

This section presents further details on the experiments in the main text and more experimental results.

This section provides additional details about the experimental setup described in the main text and reports further empirical results.

Instance Generation. For each instance $\mathcal{V} \in \mathcal{E}(n, \ell, k)$ used in the experiments, the causal graph $\mathcal{G}$ is generated as an Erdős–Rényi random graph with edge probability $p = \frac{2}{n}$. The reward's parent set is selected uniformly at random from all subsets of size k.

For every variable $X \in \mathcal{X}$ with parent set $Pa_{\mathcal{G}}(X)$, and any $\mathbf{z} \in [\ell]^{|Pa_{\mathcal{G}}(X)|}$, the conditional distribution $\mathbb{P}(X \mid Pa_{\mathcal{G}}(X) = \mathbf{z})$ is modeled as a categorical distribution with probability vector $\mathbf{v}_X(\mathbf{z}) \in \Delta^{\ell-1}$, constructed as follows. First, we sample a base vector $\mathbf{v}_X$ from a Dirichlet$(1, 1, \ldots, 1)$ distribution, which defines a random categorical distribution shared across all parent configurations $\mathbf{z}$. Then, for each $\mathbf{z}$, we draw another random vector $\mathbf{u}$ from the same Dirichlet distribution and define

$$\mathbf{v}_X(\mathbf{z}) = (1 - \beta)\, \mathbf{v}_X + \beta\, \mathbf{u},$$

where β is the *parent-effect* parameter controlling how strongly the parents influence X. For example, $\beta = 0$ corresponds to a node whose distribution is completely independent of its parents. In all experiments, we set $\beta = 0.7$ to allow moderate parent influence while maintaining stochasticity.

The reward variable is binary, and for each combination of its parents' values, the reward mean is drawn independently and uniformly from $[0, 1]$.

RAPS. The RAPS algorithm includes a structural discovery subroutine that, when applied to any node, intervenes on that node across all ℓ possible values, performing an equal number of interventions per value, and measures the resulting changes in the distributions of other variables to identify its descendants. In our experiments, the number of interventions per value was set to $\epsilon \log(10)$, which corresponds to the number of samples required to detect a change of at least ϵ with probability 0.1. A node was identified as a descendant if the probability of at least one of its values changed by more than ϵ under these interventions. The ϵ is set to 0.05.

However, this exploration phase requires a large number of rounds and is repeated once per parent, since the discovery procedure must be repeated for each parent. Consequently, we excluded RAPS from experiments involving more than one parent ($k > 1$), as its cumulative regret becomes prohibitively large in such settings.

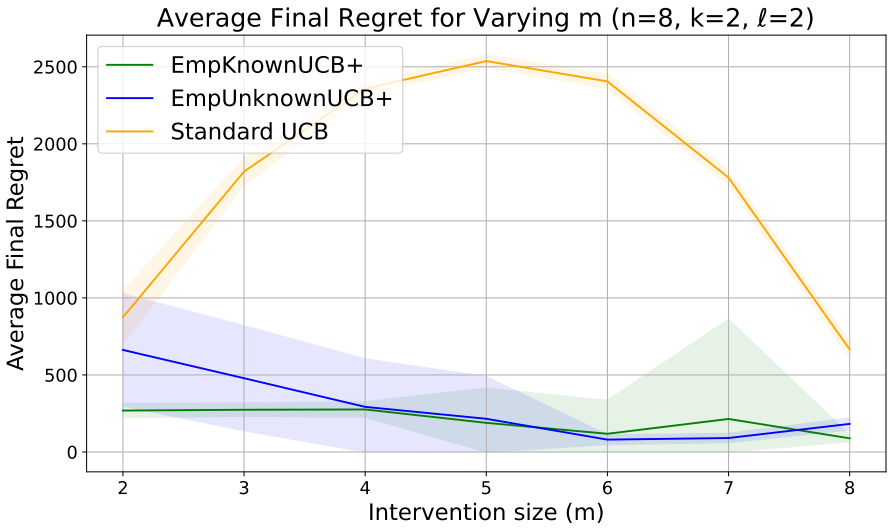


Figure 5: Average cumulative regret of algorithms at time T for varying intervention sizes $m \in 2, 3, 4, 5, 6, 7, 8$ on random instances with $n = 8$, $\ell = 2$, and $k = 2$. Larger intervention sizes enable more informative exploration, resulting in lower regret.

Effect of k. To examine the impact of the number of reward parents on algorithm performance, we evaluate all methods on a set of randomly generated instances with parameters $n = 10$, $\ell = 2$, and $m = 8$, while varying $k \in \{1, 2, 4, 6, 8\}$. Figure 4 reports the final regret for each algorithm after $T = 30{,}000$ rounds. As expected, the regret generally increases with k, reflecting the greater structural complexity and larger effective action space.

Effect of m. To analyze the influence of the intervention size, we run all algorithms on random instances with parameters $n = 8$, $\ell = 2$, and $k = 2$, varying $m \in \{2, 3, 4, 5, 6, 7, 8\}$. Figure 5 shows the final regret across algorithms after $T = 30{,}000$ rounds. Larger m values correspond to broader interventions, allowing more informative exploration and hence lower regret, consistent with our theoretical predictions.