Pure Exploration of Causal Bandits

Nuoya Xiong¹, Wei Chen²
¹ Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
² Microsoft Research, Beijing, China
xiongny20@mails.tsinghua.edu.cn, weic@microsoft.com

Abstract
Causal bandit problem integrates causal inference with multi-armed bandits. The pure exploration of causal bandits is the following online learning task: given a causal graph with unknown causal inference distributions, in each round we can choose to either intervene one variable or do no intervention, and observe the random outcomes of all random variables, with the goal that using as few rounds as possible, we can output an intervention that gives the best (or almost best) expected outcome on the reward variable $Y$ with probability at least $1 - \delta$, where $\delta$ is a given confidence level. We provide first gap-dependent fully adaptive pure exploration algorithms on three types of causal models including parallel graphs, general graphs with small number of backdoor parents, and binary generalized linear models. Our algorithms improve both prior causal bandit algorithms, which are not adaptive to reward gaps, and prior adaptive pure exploration algorithms, which do not utilize the special features of causal bandits.

1 Introduction
Stochastic multi-armed bandits (MAB) is a classical framework in sequential decision making [21, 3]. In each round, a learner selects one arm based on the reward feedback from the previous rounds, and receives a random reward of the selected arm sampled from an unknown distribution, with the goal of accumulating as much rewards as possible over $T$ rounds. This framework models the exploration-exploitation tradeoff in sequential decision making — whether one should select the best arm so far based on the past observations or one should try some arms that have not been played much. Pure exploration is an important variant of the multi-armed bandit problems, where the goal of the learner is not to accumulate reward but to identify the best arm through possibly adaptive explorations of arms. Pure exploration of MAB typically corresponds to a testing phase where we do not need to pay penalty for exploration, and it has wide applications in online recommendation, advertising, drug testing, etc.

Causal bandits, first introduced in [12], integrates causal inference [20] with multi-armed bandits. In causal bandits, we have a causal graph structure $G = (X \cup \{Y\}, E)$, where $X \cup \{Y\}$ are observable causal variables with $Y$ being a special reward variable, and $E$ is the set of causal edges between pairs of variables. For simplicity, we consider binary variables in this paper. The arms are the interventions on variables $X \in X$ together with the choice of no intervention (natural observation), i.e. the arm (action) set is $A = \{a = do(X = x) \mid X \in X, x \in \{0, 1\}\} \cup \{do()\}$, where $do(X = x)$ is the standard notation for intervening the causal graph by setting $X$ to $x$ [20]. The reward of an arm $a$ is the random outcome of $Y$ when we intervene with action $a$, and thus the expected reward is $E[Y \mid a = do(X = x)]$. In each round, one arm in $A$ is played, and the random outcomes of all variables in $X \cup \{Y\}$ are observed. Given the causal graph $G$, but without knowing its causal inference distributions among nodes, the task of pure exploration of causal bandits is to (adaptively)
select arms in each round, observe the feedback from all random variables, so that in the end the learner can identify the best arm or some near-best arm.

A number of studies have tackled pure exploration of causal bandits [12, 22, 18, 16], but most only consider the simple regret criteria, which is the expected difference in reward between the optimal arm and the arm selected by the algorithm, and the algorithms are not fully adaptive to the reward gaps. The study of [22] is the only one that discusses gap-dependent bounds for pure exploration of causal bandits in the fixed-budget setting, but it only considers soft interventions on one single node, which is quite different from the standard causal bandits defined in [12].

In this paper, we study fully adaptive pure exploration algorithms and analyze their gap-dependent sample complexity bounds in the fixed-confidence setting. More specifically, given a confidence bound \( \delta \in (0, 1) \) and an error bound \( \epsilon \), we aim at designing adaptive algorithms that output an arm such that with probability at least \( 1 - \delta \), the expected reward difference between the output and the optimal arm is at most \( \epsilon \). The algorithms should be fully adaptive in the follow two senses. First, it should adapt to the reward gaps between suboptimal and optimal arms similar to existing adaptive pure exploration bandit algorithms, such that arms with larger gaps should be explored less. Second, it should adapt to the observational data from causal bandit feedback, such that arms with enough observations already do not need further interventional rounds for exploration, similar to existing causal bandit algorithms. We are able to integrate both types of adaptivity into one algorithmic framework, and with carefully designed interaction between the two aspects, we achieve better adaptivity than either of them alone.

We first design an algorithm for the parallel graph setting where all variables in \( X \) directly pointing to \( Y \) without other causal relationships, which is also studied in [12]. We present a fully adaptive algorithm for this setting and its sample complexity result. We discuss and compare our algorithm and results with existing causal bandit algorithms or adaptive pure exploration algorithms, and show that our algorithm achieves a better sample complexity result than both classes of algorithms. Then we extend this algorithm to the general causal models where the nodes in \( X \) have a small number of backdoor parents (see Section 5 for the definition), and the binary generalized linear models.

In summary, we propose fully adaptive pure exploration algorithms for causal bandits in a number of general causal models, and provide their sample complexity analyses. Our algorithms are the first to integrate the adaptivity on reward gaps and the adaptivity on observational causal data together, and the advantage of this integration is demonstrated in our sample complexity results when comparing to the results of prior work.

2 Related work

Causal bandits problem is proposed in [12], where the authors discuss the simple regret for parallel graphs and general graphs with known probability distributions \( P(Pa(Y) \mid a) \). Nair et al. [18] extend algorithms on parallel graphs to graphs without back-door paths, and Maiti et al. [16] extend the results to the general graphs with a small number of parents. They all consider simple regret criteria that are not gap-dependent. Cumulative regret is considered in [18], [15] and [16]. To our best knowledge, [22] is the only one discussing gap-dependent bound for pure exploration of causal bandits for the fixed-budget setting, but it only considers the soft interventions (changing conditional distribution \( P(X \mid Pa(X)) \)) on one single node, which is quite different from causal bandits defined in [12]. In this paper, we provide the first gap-dependent bound in pure exploration of standard causal bandits.

Pure exploration of multi-armed bandit has been extensively studied in the fixed-confidence or fixed-budget setting [2, 8, 9, 5, 11]. PAC pure exploration is a generalized setting aiming to find the \( \epsilon \)-optimal arm instead of exactly optimal arm [4, 17, 8]. In this paper, we focus on fixed confidence setting with PAC criteria, and we utilize the adaptive LUCB algorithm in [11].

The binary generalized linear model (BGLM) is studied in [13, 6] for cumulative regret MAB problems. We borrow the maximum likelihood estimation method and its result in [13, 6] for our BGLM part, but its integration with our adaptive sampling algorithm for the pure exploration setting is new.
3 Model and Preliminaries

Causal Models From \cite{20}, a causal graph \( G = ( \mathbf{X} \cup \{ Y \}, E ) \) is a directed acyclic graph (DAG) with a set of observed random variables \( \mathbf{X} \cup \{ Y \} \), where \( \mathbf{X} = \{ X_1, \cdots, X_n \} \) is a set of variables and \( Y \) is the special reward variable. In this paper, for simplicity, we only consider \( X_i \)'s and \( Y \) are binary random variables with support \( \{ 0, 1 \} \). We consider Markovian model \( G \), which means \( G \) has no hidden variables. For any node \( X \) in \( G \), we denote the set of its parents in \( G \) as \( \text{Pa}(X) \). The set of values for \( \text{Pa}(X) \) is denoted by \( \text{pa}(X) \). The causal influence is represented by \( P(X \mid \text{Pa}(X)) \), modeling the fact that the probability distribution of a node \( X \)'s value is determined by the value of its parents. Following the rule of causal Bayesian model, the joint distribution of all nodes in the graph is given by

\[
\Pr(\mathbf{X} = x, Y = y) = \prod_{X \in \mathbf{X}} P(X = x \mid \text{Pa}(X)) \cdot P(Y = y 
\mid \text{Pa}(Y)), \tag{1}
\]

where \( x_{\text{Pa}(X)} \) is the value vector \( x \) projected on the variables in \( \text{Pa}(X) \). Henceforth, when we refer to a causal graph or a causal model, we mean both its graph structure \( (\mathbf{X} \cup \{ Y \}, E) \) and its causal influence distributions \( P(X \mid \text{Pa}(X)) \) for all \( X \in \mathbf{X} \cup \{ Y \} \). A special class of causal graphs is parallel graphs. A parallel graph \( G = (\mathbf{X} \cup \{ Y \}, E) \) is a causal graph with \( \mathbf{X} = \{ X_1, \cdots, X_n \} \), and \( E = \{ X_1 \rightarrow Y, X_2 \rightarrow Y, \cdots, X_n \rightarrow Y \} \).

An atomic intervention \( \text{do}(X = x) \) in the causal graph \( G \) means that we set node \( X \) to value \( x \), while other nodes still follow the \( P(X \mid \text{Pa}(X)) \) distributions. A null intervention \( \text{do}() \) means that we do not set any node to any value and just observe all nodes’ values following Eq.(1).

In a general causal model, distribution \( P(X \mid \text{Pa}(X)) \) may need \( |\text{Pa}(X)| \) parameters to specify. In this paper, we also study a parameterized model, the binary generalized linear model (BGLM), that only needs \( |\text{Pa}(X)| \) parameters to determine \( P(X \mid \text{Pa}(X)) \). Specifically, in BGLM, we have \( P(X = 1 \mid \text{Pa}(X) = \text{pa}(X)) = f_X(\theta_X \cdot \text{pa}(X)) + e_X \), where \( f_X \) is a monotonically non-decreasing function, \( \theta_X \in [0, 1]^{\text{pa}(X)} \) is the unknown parameter for \( X \), \( e_X \) is a zero-mean 1-sub-Gaussian noise variable that makes sure the probability is nor larger than 1. To represent the initialize probability for each node, we denote \( X_1 = 1 \) as a global variable, which is a parent of all nodes.

Pure Exploration of Causal Bandits Pure exploration of causal bandits, also called causal pure exploration, is the integration of pure exploration of multi-armed bandits with causal inference. In the causal bandit setting, variable \( Y \) is regarded as the reward variable without any outgoing edge, and the action (or arm) set of causal bandit is defined by \( A = \{ \text{do}(X_i = x) \mid i = 1, 2, \cdots, n, x = 0, 1 \} \cup \{ \text{do}() \} \). Hence \( |A| = 2n + 1 \). When we play an action \( a \in A \), we observe the sample values of all variables, and the reward is equal to \( Y \)'s value. Thus, for action \( a = \text{do}(X = x) \), we define \( \mu_a = E[Y \mid \text{do}(X = x)] \) to be the expected reward of action \( \text{do}(X = x) \). Similarly, the expected reward of action \( a = \text{do}() \) is \( \mu_a = E[Y \mid \text{do}()] = E[Y] \). Let \( \mu^* = \max_{a \in A} \mu_a \).

In each round \( t \), we perform an action in \( A \) and observe data \( \mathbf{X}_t = (X_{t,1}, X_{t,2}, \cdots, X_{t,n}) \) and \( Y_t \). The goal is to find an arm with the maximum expected reward \( \mu^* \). More precisely, we focus on the following PAC pure exploration with the gap-dependent setting in the fixed-confidence setting. In this setting, we are given a confidence parameter \( \delta \in (0, 1) \) and an error parameter \( \varepsilon \in (0, 1) \), and we want to adaptively play actions over rounds based on past observations, terminate at a certain round and output an action \( a^* \) to guarantee that \( \mu^* - \mu_{a^*} \leq \varepsilon \) with probability at least 1 - \( \delta \). The metric for this setting is sample complexity, which is the number of rounds needed to output a proper action \( a^* \).

We study the gap-dependent bounds, which means the performance measure is related to the reward gap between the optimal and suboptimal actions, as defined below. Let \( a^* \) be one of the optimal arms. For each arm \( a \), we define the gap of \( a \) as

\[
\Delta_a = \begin{cases} 
\mu_{a^*} - \max_{a \in A \setminus \{a^*\}} \{ \mu_a \}, & a = a^*; \\
\mu_{a^*} - \mu_a, & a \neq a^*. 
\end{cases} \tag{2}
\]

We further sort the gaps \( \Delta_a \)'s for all arms and assume \( \Delta^{(1)} \leq \Delta^{(2)} \leq \cdots \leq \Delta^{(2n+1)} \), where \( \Delta^{(1)} \) is also denoted as \( \Delta_{\text{min}} \).
4 Pure Exploration for Parallel Causal Graphs

We first consider the pure exploration for parallel causal graphs, which is a simple but informative case also considered by Lattimore et al. [12]. For parallel graphs we have \( P(Y \mid do(X_i = x)) = P(Y \mid X_i = x) \) for every \( i = 1, 2, \ldots, n \) and \( x \in \{0, 1\} \), which means we can estimate \( P(Y \mid do(X_i = x)) \) by only estimating the observed distribution \( P(Y \mid X_i = x) \).

Following [12], we assume that \( X_i \)'s are Bernoulli random variables when we do not intervene any variables. and define \( q_a = P(X = x) \) for action \( a = do(X = x) \), or \( q_a = 1 \) for action \( a = do() \). Intuitively, \( q_a \) represents the probability for \( X = x \) during observation. Following [12], we define the observation threshold as follows:

**Definition 1** (Observation threshold [12]). For a given causal graph \( G \) and its associated \( \{q_a \mid a \in A\} \), the observation threshold \( m \) is defined as:

\[
m = \min\{\tau \in [2n + 1] : \{a \in A \mid q_a < 1/\tau\} \leq \tau\}.
\]

The observation threshold can be equivalently defined as follows: When we sort \( \{q_a \mid a \in A\} \) as \( q^{(1)} \leq q^{(2)} \leq \cdots \leq q^{(2n+1)} \), \( m = \min\{\tau : q^{(\tau+1)} \geq \frac{1}{\tau}\} \). Note that \( m \leq n \) because \( \max\{q_{do(X_i=1)}, q_{do(X_i=0)}\} \geq \frac{1}{2} \), which implies \( \{|a \in A, q_a < \frac{1}{m}\}| \leq n \). In some cases, \( m \ll n \).

For example, when \( q_{do(X_i=1)} = q_{do(X_i=0)} = \frac{1}{2} \) for all \( i \in [n] \), \( m = 2 \ll n \). Intuitively, when we collect passive observation data without intervention, arms corresponding to \( q^{(j)} \) with \( j \leq m \) are under sampled while arms corresponding to \( q^{(j)} \) with \( j > m \) are sufficiently sampled. Thus, for convenience we name \( m \) as the observation threshold (the term is not given a name in [12]).

In this paper, we improve the definition of \( m \) to make it gap-dependent, which would lead to a better adaptive pure exploration algorithm and sample complexity bound.

**Definition 2** (Gap-dependent observation threshold). For a given causal graph \( G \) and its associated \( q_a \)'s and \( \Delta_a \)'s, the gap-dependent observation threshold \( m_{\epsilon, \Delta} \) is defined as:

\[
m_{\epsilon, \Delta} = \min\left\{ \tau \in [2n + 1] : \left\{ q_a < \max\left\{ \frac{\Delta_{\min}, \epsilon/2}{\tau}, \max\{\Delta_a, \epsilon/2\}\right\} \right\} \leq \tau \right\}.
\]

Note that \( m_{\epsilon, \Delta} \leq m \) because \( \max\{\Delta_{\min}, \epsilon/2\} \leq \max\{\Delta_a, \epsilon/2\} \) for all \( a \in A \). When \( \Delta_{\min} \) is significantly smaller than other \( \Delta_a \)'s, we may have \( m_{\epsilon, \Delta} \ll m \).

In our algorithm, \( \hat{q}_a \) and \( \hat{\mu}_a \) represent the estimates of \( q_a \) and \( \mu_a \), respectively. For \( a = do(X = x) \), we denote \( T_n(t) \) as the number of times \( X = x \) happens by the end of round \( t \). Our algorithm is based on the LUCB algorithm [11], in which the algorithm contains a lower confidence bound and an upper confidence bound for all arms. The confidence radius is defined by

\[
\beta(u, t) = \sqrt{\frac{1}{2u} \log \frac{8nt^3}{\delta}}.
\]

The confidence interval is naturally defined by

\[
[L_a, U_a] = [\hat{\mu}_a - \beta(T_n(t), t), \hat{\mu}_a + \beta(T_n(t), t)],
\]

and \( L_a \) and \( U_a \) are called the lower confidence bound and the upper confidence bound for arm \( a \). We call arm \( a \) dominates \( a' \) at round \( t \) if \( U_a < L_{a'} + \epsilon \), denoted by \( a \succeq_t a' \). Arm \( a \) dominating arm \( a' \) means that choosing \( a \) is always better than \( a' \) in terms of \( \epsilon \)-optimality.

**Pure exploration algorithm for parallel graphs** Our algorithm is given in Algorithm 1 (named Causal-PE-parallel). The algorithm contains two phases. In Phase 1, our goal is to estimate the means of arms and divide them into sets WellSampled, Uncertain, and Rejected. WellSampled contains arms whose upper and lower confidence bounds have gap at most \( \epsilon \), which means they have been well sampled and their estimates are accurate enough. Among the remaining arms, Rejected contains the arms dominated by others, and Uncertain contains the remaining arms that we are still unsure about their reward estimates.
Phase 1 operations are divided into odd-round and even-round operations. In the odd rounds (lines 4-11), we perform `do()` operation and update the estimation $\hat{q}_a$, $\hat{\mu}_a$ and the confidence intervals. This part is similar to the first phase of the parallel bandit algorithm in [12]. But we are not only estimating $\hat{q}_a$’s but also estimating the confidence intervals, which are used to eliminate some arms for further processing. The algorithm in [12] has no such mechanism in eliminating arms based on reward confidence intervals. In the even rounds (lines 12-17), we choose an arm $a = \arg\max_{a' \in A} (\hat{\mu}_a + 3\beta(T_a(t), t))$, play it and update relative parameters. This is crucial to help eliminating many suboptimal arms. When an optimal arm $a^*$ has low $q_{a^*}$, it is hard to observe data for $a^*$ in the odd rounds, and thus $a^*$ would have a large confidence interval and making it incapable of eliminating suboptimal arms. Hence, to handle this issue, in the even rounds, we select the arm $a$ with the largest $\hat{\mu}_a + 3\beta(T_a(t), t)$ to play, giving more chances to play potentially optimal arms and thus making our algorithm more effective in eliminating suboptimal arms. The reason we use a larger confidence radius $3\beta(T_a(t), t)$ in this case is to facilitate our analysis such that the optimal arm has a sufficient chance to be played. At the end of each round, we call the `Partition()` routine (Algorithm 2) to
partition arms into three sets WellSampled, Uncertain, and Rejected. The set Uncertain is used in the stopping rule for Phase 1, as given in line 3: Intuitively, the first rule in the while condition requires that all arms in the Uncertain set has small $q_a$, except the optimal arm, which may not be eliminated into the Rejected set. The second rule is to ensure that we have enough rounds for Phase 1 so that the estimates of $q$’s are reasonably accurate.

In Phase 2, we call the adaptive LUCB algorithm [11] to efficiently identify the best arm among the remaining arms in Uncertain $\cup \{a''\}$, where $a''$ is the arm with highest upper confidence bound in WellSampled. For completeness, we provide the pseudo-code of LUCB in Appendix B.

Let $\bar{U} = \min(3m+2, 2n+1)$, which we will show is the upper bound of the size of Uncertain $\cup \{a''\}$.

The sample complexity result of our algorithm is given below.

**Theorem 1.** Denote $H_{\varepsilon}^{(P)} = \sum_{i=1}^{\bar{U}} \frac{1}{\max\{\Delta^{(i)}, \varepsilon/2\}^2} + \frac{m_{c, \Delta}}{\max\{\Delta_{\min, \varepsilon/2}\}^2}$. With probability $1 - \delta$, Causal-PE-parallel($G, A, \varepsilon/3, \delta/4$) returns an $\varepsilon$-optimal arm with sample complexity

$$T = O\left(H_{\varepsilon/3}^{(P)} \log \frac{nH_{\varepsilon/3}^{(P)}}{\delta}\right),$$

where $m$ and $m_{c, \Delta}$ are defined in Definitions 1 and 2.

There are two main terms in the sample complexity. The first term, $\sum_{i=1}^{\bar{U}} \frac{1}{\max\{\Delta^{(i)}, \varepsilon/2\}^2}$ is actually a fixed upper bound of $\sum_{a \in \text{Uncertain}} \frac{1}{\max\{\Delta_a, \varepsilon/2\}^2}$, which is the sample complexity of the LUCB algorithm used in Phase 2. The second term, $\frac{m_{c, \Delta}}{\max\{\Delta_{\min, \varepsilon/2}\}^2}$ is the upper bound of the sample complexity of Phase 1. We now compare our algorithm and the result with two main prior algorithms in the literature.

**Comparing with the algorithm and results in [12].** The parallel bandit algorithm in [12] is a two phase algorithm, and within each phase it uses uniform sampling. Therefore, their algorithm has very limited adaptivity, and the algorithm only removes some arms with larger $q_a$’s from the sampling in the second phase. In contrast, our Causal-PE-parallel algorithm is a fully adaptive algorithm, and we have several mechanisms for adaptive sampling, which result in savings in sample complexity. First, we use upper and lower confidence bounds to eliminate dominated arms for further consideration in Phase 1. Second, we divide Phase 1 into odd rounds and even rounds and use adaptive sampling of arms in even rounds to improve the effectiveness of arm eliminations. Third, we put aside arms with accurate estimates into the WellSampled set to reduce the second phase processing. Fourth, in the second phase we apply the adaptive LUCB algorithm for identifying the best among the remaining uncertain arms. These mechanisms together lead to better sample complexity results.

The original result in [12] is on the simple regret with a fixed budget $T$. To compare sample complexity, we first convert their result into the sample complexity result with a fixed confidence bound $\delta$, which is stated as $T = O\left(\frac{n}{\varepsilon^2} \log \frac{2n}{\delta}\right)$. Comparing to our result in Theorem 1, first we can see that when ignoring the logarithmic term, ours is always better than their result, because $m_{c, \Delta} \leq m$ and $\bar{U} = O(m)$. When the gaps $\Delta_{a}$’s have large variation and $\Delta_{\min}$ is significantly smaller than the other $\Delta_{a}$’s, we know that $m_{c, \Delta} \ll m$ and the first term $\sum_{i=1}^{\bar{U}} \frac{1}{\max\{\Delta^{(i)}, \varepsilon/2\}^2}$ is also much smaller than $\frac{m}{\varepsilon^2}$, and thus our algorithm would achieve much better sample complexity. Intuitively, this is because our algorithm fully adapts to the gaps of different arms while the parallel bandit algorithm in [12] is ignorant to these gaps.

**Compare with the LUCB algorithm [11].** LUCB is an adaptive algorithm that adapts to the gaps of arms, but it does not consider the special adaptivity in parallel causal bandits. If we apply LUCB directly to our parallel causal bandit problem, the sample complexity is $O(H_{\varepsilon}^{(P)})$, where $H_{\varepsilon}^{(P)} = \sum_{i=1}^{2n+1} \frac{1}{\max\{\Delta^{(i)}, \varepsilon/2\}^2}$. Comparing to Theorem 1, we see that our first term is always smaller than $H_{\varepsilon}^{(P)}$, while our second term is also smaller than $H_{\varepsilon}^{(P)}$ either when $m_{c, \Delta}$ is a constant or when the smallest $m_{c, \Delta}$ gaps are all close to $\Delta_{\min}$. For example, suppose for all $a \in A, q_a \geq c$ for some constant $c$, that is, a passive observation allow non-negligible probability to observe all actions. Then $m_{c, \Delta} \leq m \leq 1/c$, and $H_{\varepsilon/3}^{(P)} = O\left(\frac{3m+1}{\max\{\Delta^{(i)}, \varepsilon/2\}^2}\right)$, much smaller than $H_{\varepsilon}^{(P)}$. Intuitively, our algorithm integrates the passive observation with active selection in the first phase to eliminate many arms that have been sufficiently observed or dominated, leading to better sample complexity.
Overall, the innovation of our Causal-PE-parallel algorithm is a nontrivial integration of the adaptive sampling from confidence bound based algorithm LUCB with the parallel causal bandit algorithm on utilizing passive observations. As a result, we achieve improved sample complexity by making our algorithm fully adaptive to both reward gaps of arms and their observation probabilities.

For classical pure exploration, we can set $\varepsilon = 0$ and derive the following corollary from Theorem 1:

**Corollary 1.** Denote $H_0^{(P)} = \sum_{i=1}^{G} \frac{1}{(\Delta_i)^2} + \frac{m\Delta}{S_{\min}}$. With probability $1 - \delta$, Causal-PE-parallel$(G, A, 0, \delta/4)$ returns an optimal arm with sample complexity $T = O(H_0^{(P)} \log n H_0^{(P)})$.

5 Pure Exploration for General Graphs with Small Backdoor Parent Size

In this section we apply the similar idea for the parallel causal graphs to the general graph setting. The first issue is how to estimate the causal effect (or the do effect) $E[Y \mid do(X_i = x)]$ in general causal graphs from the observational data. Since in this paper we only consider $G$ without hidden variables, we are able to use the do-calculus [20] to achieve this estimation. For the causal graph $G$, a backdoor path from $X$ to $Y$ is a sequence of distinct nodes $(X, X_1, X_2, \ldots, X_k, Y)$ such that $X_i \in Pa(X_i)$, and $(X_i, X_{i+1}), i \in [k - 1]$ and $(X_k, Y)$ are connected by edges in either direction. We say that $X_1 \in Pa(X)$ is a backdoor parent of $X$ if $X_1$ is on a backdoor path from $X$ to $Y$. Let $Pa'(X)$ denote the set of backdoor parents of $X$. Define $D' = \max_{X \in X} |Pa'(X)|$ and denote $Z = 2^{D'}$ as the maximum number of possible value combinations for backdoor parents. Then the following formula is the result of the do-calculus [20] (also called the backdoor criterion):

$$E[Y \mid do(X_i = x)] = \sum_z P(Y = y \mid X = x, Pa'(X) = z) \cdot P(Pa'(X) = z).$$  

(8)

Using the above formula, we estimate each term of the right-hand side for every $z \in \{0, 1\}^[Pa'(X)]$ to obtain an estimate for $E[Y \mid do(X_i = x)]$. Note that to handle general causal graphs, we would encounter the term $Z = 2^{D'}$ in both sample complexity and computational complexity, and thus the results would be reasonable when $D'$ is small, e.g. $D' \leq \frac{1}{2} \log n$ (see Appendix A for graph examples satisfying such conditions).

Algorithm 3 provides the pseudocode of our algorithm. Compared to Algorithm 1, the main modification is the estimation for $\hat{\mu}_a$’s and the confidence bounds, as we explain below. Let $Pa'(X)$ be the random values of backdoor parents of $X$ at round $t$. For an arm $a = do(X = x)$ and a vector $z \in \{0, 1\}^[Pa'(X)]$, (a) let $d_{a,z} = P(X = x, Pa'(X) = z)$, $q_a = \min_z\{d_{a,z}\}$; (b) let $T_{a,z}(t)$ be the number of observations in the odd rounds of Phase 1 in which $X = x, Pa'(X) = z$, and $T_a(t) = \min_z\{T_{a,z}(t)\}$; and (c) let $r_{a,z}(t)$ be the estimator of $P[Y = y \mid X = x, Pa'(X) = z]$, and $p_{a,z}(t)$ be the estimator of $P[Pa'(X) = z]$, also in the odd rounds of Phase 1. The definitions of $m_{\varepsilon, \Delta}$ and $m$ are the same as in Definitions 1 and 2, once we fix $q_a$’s as defined above.

We use two confidence bounds, one from the estimate of the observational data in the odd rounds of Phase 1, and the other from the estimate of the interventional data in other rounds. Define $\hat{\mu}_O,a = \sum_{Pa(X)} r_{a,z}(t) p_{a,z}(t)$ and $\hat{\mu}_I,a = \frac{1}{D_{a}(t)} \sum_{j=1}^{n} \mathbb{I}\{a_j = a\} Y_j, D_a(t) = \sum_{j=1}^{n} \mathbb{I}\{a_j = a\}$ as the estimates to the true mean from the observational and the interventional rounds, respectively. Their confidence radius are given as:

$$\beta_O(T_a(t), t) = \sqrt{\frac{1}{2D_a(t)} \log \frac{16nZt^3}{\delta}}, \beta_I(D_a(t), t) = \sqrt{\frac{1}{2D_a(t)} \log \frac{16nZt^3}{\delta}}.$$  

The observational confidence interval and the interventional confidence interval are defined as $[L_{O,a}^t, U_{O,a}^t] = [\hat{\mu}_O,a - \beta_O(T_a(t), t), \hat{\mu}_O,a + \beta_O(T_a(t), t)], [L_{I,a}^t, U_{I,a}^t] = [\hat{\mu}_I,a - \beta_I(D_a(t), t), \hat{\mu}_I,a + \beta_I(D_a(t), t)].$ The final confidence interval becomes

$$[L_a^t, U_a^t] = [L_{O,a}^t, U_{O,a}^t] \cap [L_{I,a}^t, U_{I,a}^t].$$  

(9)

We then estimate the true mean $\hat{\mu}_a = (L_a^t + U_a^t)/2$. From above we can see $U_a^t - L_a^t \leq 2 \min\{\beta_O(T_a(t), t), \beta_I(D_a(t), t)\}$. Our sample complexity result is given below.
Algorithm 3: Causal-PE-general $(G, A, \varepsilon, \delta)$

**Input:** causal parallel graph $G$, action set $A$, parameter $\varepsilon, \delta$

1. Initialize $t = 1$, $Uncertain = A$, $WellSampled = \emptyset$, $Rejected = \emptyset$, $T_a(0) = 0$, $\tilde{\mu}_a = 0$ for all arms $a \in A$.
2. **phase 1:**
3. while $(|\{a \in Uncertain : \tilde{q}_a \geq \frac{2}{|Uncertain|-1}\}| > 1$ or $t < 36|Uncertain| \log \frac{6\varepsilon}{\delta})$ do

4. if $2 \mid t$ then

5. Perform $a_t = do()$ and observe $X_t$ and $Y_t$. For $a = do()$, $T_a(t) = T_a(t-1) + 1$, $\tilde{q}_a = 1$, $D_a(t) = D_a(t-1) + 1$,

6. $r_{a,0}(t) = \frac{1}{T_a(t)} \sum_{j=0}^{t-1} \mathbb{I}\{2 \mid j \}, Y_j, p_{a,0}(t) = 1.$

7. for $a = do(X = x) \in A, z \in \{0, 1\}$ do

8. Update $T_{a,z}(t) = \sum_{j=1}^{t} \mathbb{I}\{2 \mid j, X_j = x, Pa'(X)_j = z\}, T_a(t) = \min_z\{T_{a,z}(t)\},$

9. $\tilde{q}_a = \frac{2}{t+1} T_a(t), D_a(t) = D_a(t-1)$

10. Update $r_{a,z}(t) = \frac{1}{T_{a,z}(t)} \sum_{j=0}^{t} \mathbb{I}\{2 \mid j, X_j = x, Pa'(X)_j = z\} Y_j, p_{a,z}(t) = \frac{1}{|\{a \mid t\}} \sum_{j=1}^{t} \mathbb{I}\{a_j = a\} Y_j$.

11. Estimate $\hat{\mu}_{O,a} = \sum_z r_{a,z}(t)p_{a,z}(t), \hat{\mu}_{I,a} = \frac{1}{T_a(t)} \sum_{j=1}^{t} \mathbb{I}\{a_j = a\} Y_j.$

12. Update confidence bound $[L_a^t, U_a^t]$ using Eq. (9), $\hat{\mu}_a = (L_a^t + U_a^t)/2$, for each arm $a$.

end

if $2 \mid t$ then

14. Perform arm $a_t = \arg \max_{a \in A} 2 \cdot U_a^t - L_a^t.$

15. for $a \in A$ do

16. Update $D_a(t) = D_a(t-1) + \mathbb{I}\{a_t = a\}, T_a(z)(t) = T_a(z)(t-1), r_{a,z}(t),$ $p_{a,z}(t), \hat{\mu}_{O,a}, \hat{\mu}_{I,a}$. Calculate $L_a^t, U_a^t$ using Eq. (9), $\hat{\mu}_a = (L_a^t + U_a^t)/2$.

end

end

( Uncertain, WellSampled, Rejected) = Partition$(t, \{U_a^t, L_a^t \mid a \in A\}, \varepsilon)$

$t = t + 1$

end

**phase 2:** Same as Algorithm 1

---

**Theorem 2.** Define $H_{\varepsilon}^{(G)} = \sum_{i=1}^{U} \frac{1}{\max(\Delta_{ij},(\varepsilon/2)^2)} + \frac{m_{\Delta} + Z^2}{\max(\Delta_{ij},(\varepsilon/2)^2)}$. With probability $1 - \delta$, Causal-PE-general $(G, A, \varepsilon/3, \delta/5)$ returns an $\varepsilon$-optimal arm with sample complexity

$$T = O\left(\frac{H_{\varepsilon}^{(G)} \log nZH_{\varepsilon}^{(G)}}{\delta/3}\right),$$

(10)

where $m, m_{\varepsilon, \Delta}$ are defined in Definitions 1 and 2 in terms of $q_1$’s for the general graphs.

The above result contains $Z^2$ in the sample complexity. When $D' < \frac{1}{2} \log n$, we have $Z^2 < n$. For example, if $D' = \frac{1}{2} \log n$, $Z^2 = \sqrt{n} = o(n)$, and the sample is reasonably good. In Appendix A, we provide some graphs which satisfy this assumption. Prior study [16] also contains a similar term $K = 2D^2 \geq Z^2$, but they regard it as constant and ignore it in the final result. Similar to the case of parallel graphs, when comparing to prior studies on causal bandit algorithms, our algorithm wins when there are large variance in the reward gaps; and when comparing to prior studies on adaptive pure exploration algorithms, our algorithms win by estimating do effects using observational data and eliminate arms that do not need further interventions.

---

**6 Pure Exploration for Binary Generalized Linear Model**

In the previous section, we have a factor $Z = 2D'$ in the sample complexity, similar to prior study [16]. The intrinsic reason for this factor is because for general graphs, one needs about $Z$ parameters...
to specify the inference distribution from the parents of \( X \) to \( X \), and thus observing probabilities to estimate the causal effect cannot avoid the factor \( Z \). In this section, we discuss the pure exploration for a general class of causal graphs with a linear number of parameters, such that the factor \( Z \) can be avoided. This is the binary generalized linear model (BGLM) as described in Section 3.

Let \( \theta^* = (\theta_X)_{X \in \mathcal{X} \cup \{Y\}} \) be the vector of all weights. Since \( \theta^* \geq 0 \), do(\( X = 1 \)) is always better than do(\( X = 0 \)) and do(). Hence, we only need to consider the action set \( A = \{do(X_i = 1) \mid i = 1, 2, \cdots, n\} \). Let \( D = \max_{X \in \mathcal{X} \cup \{Y\}} |Pa(X)| \) be the maximum in-degree of nodes in \( \mathcal{X} \cup \{Y\} \).

Note that every node \( X \) has a parent \( X_1 = 1 \) in BGLM to model the natural probability that \( X = 1 \) when all other parents of \( X \) are 0. Following [13, 6], we have three assumptions:

**Assumption 1.** For any \( X \in \mathcal{X} \), \( f_X \) is twice differentiable. Its first and second order derivatives can be upper bounded by \( M^{(1)} \) and \( M^{(2)} \).

**Assumption 2.** \( \kappa := \inf_{X \in \mathcal{X} \cup \{Y\}, v \in [0,1]} |\theta - \theta_X^*| \) is a positive constant.

**Assumption 3.** There exists a constant \( \eta > 0 \) such that for any \( X \in \mathcal{X} \cup \{Y\} \) and \( X' \in Pa(X) \), for any \( v \in \{0,1\}^{Pa(X)-2} \) and \( x \in \{0,1\} \), we have

\[
Pr[X' = x \mid Pa(X) \setminus \{X', X_1\} = v] \geq \eta. 
\]

Assumptions 1 and 2 are the classical assumptions in generalized linear model [13]. Assumption 3 makes sure that each parent node of \( X \) has some freedom to become 0 and 1 with a non-zero probability, even when the values of all other parents of \( X \) are fixed. It is natural in the sense that the state of one parent is not fully determined by the states of other parents, and it is originally given in [6] with additional justifications.

Our algorithm Causal-PE-BGLM for BGLM patterns in the same way as our algorithms for the parallel graphs and general graphs: it contains three phases, and the second and the third phases are the same as the previous ones. In the first phase, we have the same odd and even round alternations between observations and interventions. The main difference is here we apply the maximum likelihood estimation method of [13, 6] to estimate parameters \( \theta^* \), and use a Lipschitz smoothness condition to derive the confidence bounds of all arms. Due to the space constraint, the full algorithm and explanations are included in Appendix C.

For node \( X \in \mathcal{X} \), let \( \ell_X \) be the number of nodes that lie in some paths from node \( X \) to reward variable \( Y \) excluding \( X \), and \( q_a^{(L)} = 1/\ell_X^2 \) for \( a = do(X = 1) \). Observation thresholds \( m^{(L)} \) and \( m^{(L)}_{\epsilon, \Delta} \) are defined as in Definitions 1 and 2, with respect to \( q_a^{(L)} \). For \( \bar{U}^{(L)} = \min\{3m^{(L)} + 1, n\} \), the algorithm provides the following result:

**Theorem 3.** Define \( H^{(L)}_\epsilon = \sum_{i=1}^{\bar{U}^{(L)}} \frac{1}{\max\{\Delta^{(U)}_{\epsilon, \Delta} / 2\}^2} + \frac{m^{(L)}_{\epsilon, \Delta}}{\eta \max\{\Delta^{(U)}_{\epsilon, \Delta} / 2\}^2} \). With probability \( 1 - \delta \), our Causal-PE-BGLM(\( G, A, \epsilon / 3, \delta / 4 \)) returns a \( \epsilon \)-optimal arm with sample complexity

\[
T = O \left( H^{(L)}_\epsilon \log \frac{nH^{(L)}_\epsilon / \delta}{\delta} \right),
\]

where \( m^{(L)} \) and \( m^{(L)}_{\epsilon, \Delta} \) are defined in Definitions 1 and 2 in terms of \( q_a^{(L)} \).

In Appendix A we provide a class of graphs in which \( m^{(L)}_{\epsilon, \Delta} D^3 \leq m^{(L)} D^3 \ll n \), where our algorithm outperforms existing algorithms such as LUCB and also beats the lower bound for the pure exploration of classical MAB problems.

### 7 Future Work

There are many interesting directions worth exploring in the future. First, for general graphs, one can try to eliminate the assumption on small number of backdoor parents (i.e. small \( D' \)) by possibly using some techniques other than the do-calculus. Second, it would be interesting to study the gap-dependent lower bounds for pure exploration of causal bandits. Furthermore, one can also consider developing efficient pure exploration algorithms for causal graph with hidden variables or even partially unknown graph structures.
References

[1] Raghavendra Addanki and Shiva Kasiviswanathan. Collaborative causal discovery with atomic interventions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 12761–12773. Curran Associates, Inc., 2021.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In Adam Tauman Kalai and Mehryar Mohri, editors, COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010, pages 41–53. Omnipress, 2010.

[3] Sébastien Bubeck, Nicolò 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.

[4] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In Jyrki Kivinen and Robert H. Sloan, editors, Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings, volume 2375 of Lecture Notes in Computer Science, pages 255–270. Springer, 2002. do10.1007/3-540-45435-7_18.

[5] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of Machine Learning Research, 7:1079–1105, 06 2006.

[6] Shi Feng and Wei Chen. Combinatorial Causal Bandits. arXiv e-prints, art. arXiv:2206.01995, June 2022.

[7] Torben Hagerup and Christine Rüb. A guided tour of chernoff bounds. Information Processing Letters, 33(6):305–308, 1990. ISSN 0020-0190. doihttps://doi.org/10.1016/0020-0190(90)90214-I.

[8] Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In 2014 48th Annual Conference on Information Sciences and Systems (CISS), pages 1–6, 2014. doi10.1109/CISS.2014.6814096.

[9] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil’ ucb : An optimal exploration algorithm for multi-armed bandits. Journal of Machine Learning Research, 35, 12 2013.

[10] Shivaram Kalyanakrishnan. Learning methods for sequential decision making with imperfect representations, 2011. PhD thesis, Department of Computer Science, The University of Texas at Austin.

[11] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. Proceedings of the 29th International Conference on Machine Learning, ICML 2012, 1, 01 2012.

[12] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: learning good interventions via causal inference. In Advances in Neural Information Processing Systems, pages 1189–1197, 2016.

[13] Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In International Conference on Machine Learning, pages 2071–2080. PMLR, 2017.

[14] Shuai Li, Fang Kong, Kejie Tang, Qizhi Li, and Wei Chen. Online influence maximization under linear threshold model. In Advances in Neural Information Processing Systems, 2020.

[15] 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.
[16] Aurghya Maiti, Vineet Nair, and Gaurav Sinha. Causal bandits on general graphs. *arXiv preprint arXiv:2107.02772*, 2021.

[17] Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, dec 2004. ISSN 1532-4435.

[18] 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.

[19] Zipei Nie. Matrix anti-concentration inequalities with applications. *arXiv preprint arXiv:2111.05553*, 2021.

[20] Judea Pearl. *Causality*. Cambridge University Press, 2009. 2nd Edition.

[21] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

[22] 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.

**Checklist**

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
Appendix

A General Classes of Graphs Supporting Theorem 2 and Theorem 3

A.1 Theorem 2

For Theorem 2, to achieve a reasonable result, we need to assume that \( Z^2 = 2^{2D'} = o(n) \), where
\[
D' = \max_{X \in \mathcal{X}} |Pa'(X)|.
\]
For all no-backdoor graphs defined in [18] such as parallel graphs and causal trees, we have \( D' = 0 \) and \( Z = 1 \), then \( Z^2 = o(n) \) holds. Now we provide some other graphs satisfying this assumption.

Two-layer graphs Consider \( X = A \cup B \), where \( A = \{X_1, \ldots, X_k\} \) is the set of key variables, \( B = \{X_{k+1}, \ldots, X_n\} \) are the rest of variables. Now we consider \( k \leq \frac{1}{7} \log_2 n \), and the edge set is in
\[
E \subseteq \{(X_i \rightarrow X_j) \mid X_i \in A, X_j \in A\} \cup \{(X_i \rightarrow Y) \mid X_i \in A, X_j \in B\} \cup \{(X_i \rightarrow Y) \mid X_i \in B\}.
\]

Collaborative graphs Consider \( X = X^1 \cup X^2 \cup \cdots \cup X^l \), where each \( X^i (1 \leq i \leq l) \) has at most \( k \leq \frac{1}{7} \log n \) nodes. The edge set is contained in
\[
E \subseteq \{X \rightarrow Y \mid X \in X^1\} \cup \{X_i \rightarrow X_j \mid X_i, X_j \in X^t, 1 \leq t \leq l\}. \]
We call this class of graphs collaborative graphs (see Figure 2), since it is motivated by [1] on collaborative causal discovery.
For these graphs, \( Z^2 = 2^{2D'} \leq 2^{2k} \leq \sqrt{n} = o(n) \), and thus this class of graphs also satisfies the assumption \( Z^2 = o(n) \). Collaborative graphs are useful in many real-world scenarios. For example, many companies want to cooperate and maximize their profits. Then each subgraph \( X_i (1 \leq i \leq l) \) represents a company, and they want to find the best intervention to generate the maximum profit.

A.2 Theorem 3

Multi-layer graphs Recall that for any node \( X \), \( \ell_X = |P_X,Y| \), where \( P_X,Y \) is set of nodes lying on a path from \( X \) to \( Y \) excluding \( X \). Then we consider a multi-layer graph \( T = T_1 \cup T_2 \cup \cdots \cup T_m \), where \( T_1 = \{ Y \} \), and all edges are from \( T_i \rightarrow T_j, i > j \). Assume nodes in \( T \) have in-degree at most \( d = o(n^{1/5}) \), out-degree at most a constant \( k \), and \( m = \frac{1}{\delta} \log_k n = O(\log n) \). For example, causal trees pointed toward \( Y \) with depth less than \( m \) and in-degree at most \( d \) satisfy these conditions.

Now we can see \( D \leq d = o(n^{1/5}) \). Consider \( m_{(L)}^{(L)} \) in this graph. For any nodes \( X \in T \), \( \ell_X = |P_{X,Y}| \leq k \text{depth}(T) \leq n^{1/5} \). Thus \( q_X = \frac{1}{\ell_X} \geq n^{-2/5} \). Now \( |\{ X \in X \mid q_X \leq n^{-2/5} \}| \leq k \leq n^{2/5} \). By definition of \( m_{(L)}^{(L)} \) and \( m^{(L)} \), we can get \( m_{(L)}^{(L)} \leq m^{(L)} \leq n^{2/5} \). Thus \( m_{(L)}^{(L)} D^3 = o(n) \).

This class of causal graphs is also useful in real-world scenario. For example, our goal is to maximize the probability that the viewer browses a particular website. Some relevant websites consists of a multi-layer network, while each websites contains at most \( k \) hyperlinks linking to the website closer to the target website than itself. We want to achieve our goal by advertising on a single website, which is modeled by causal bandits on this class of graphs.

B Pseudo-code of LUCB

For completeness, we provide the pseudo-code of LUCB in [11] here. Let \( T_u(t) \) be the number of times arm \( a \) has been sampled until round \( t \), \( \hat{\mu}_a(t) \) be the empirical mean for arm \( a \) until round \( t \), and \( \gamma(u,t) = \sqrt{\frac{2\log(5nt^2)}{du}} \).

Algorithm 4: LUCB(\( A, \varepsilon, \delta \))

Input: Action set \( A \), parameter \( \varepsilon, \delta \)
1. Pull each arm \( a \in A \) once. \( t \leftarrow |A| \). Initialize \( T_u(t) = 1 \) for all arms \( a \).
2. \( \hat{\mu}_a = \arg \max_{a \in A} \hat{\mu}_a \), \( \hat{a}_u = \arg \max_{a \in A} \{ a \} \hat{\mu}_a \).
3. while \( \hat{\mu}_{a^t} + \gamma(T_{a^t}(t), t) \geq \gamma(T_{a^t}(t), t) + \varepsilon \) do
   4. sample \( a^t \) and \( \hat{a}^t \).
   5. Update \( \hat{\mu}_{a^t+1}, T_{a^t+1} \) for all arm \( a \in A \).
6. Choose \( a_{t+1} = \arg \max_{a \in A} \hat{\mu}_{a}+1 \), \( a_{t+1} = \arg \max_{a \in A} \{ a \}_{a^t+1} \hat{\mu}_{a}+1 \).
7. \( t = t + 1 \)
8. end

C Pseudocode and Explanations For the BGLM Algorithm

In this appendix, we provide the pseudocode as well as explanations and related results for BGLM. The pseudocode for the main algorithm Causal-PE-BGLM is given in Algorithm 5, which causes the maximum likelihood procedure in Algorithm 6.

For BGLM, it is easy to calculate \( \sigma(\theta, X) = E[Y \mid do(X = 1), \theta] \) for every \( X \in X \), if we know \( \theta \) as the prior knowledge. At round \( t \), let our estimation for \( \theta \) be \( \hat{\theta}_t \). For every arm \( a = do(X = 1) \), we can calculate \( \hat{\mu}_{\theta,a} = \sigma(\hat{\theta}_t, X) \) by following the topological order of the graph \( G \) with \( \hat{\theta}_t \) as the parameters. To get the confidence bound for BGLM, we use the following lemma from [6]:
Input: causal parallel graph $G$, action set $A$, parameter $\varepsilon, \delta, M(1), M(2), \kappa, \eta, c$ in Assumptions 1, 2, 3 and Lemma 10.

1. Initialize $t = 1$, $Uncertain = A$, $WellSampled = \emptyset$, $Rejected = \emptyset$, $M_{-1,X} = I$, $b_{-1,X} = 0$

2. Phase 1:

3. while $\{a \in Uncertain : q_d^{(L)} \geq \frac{2}{\log(1 - \epsilon)}\} > 1$ or

   $t < \max\{36|Uncertain| \log \frac{n}{\delta} \log \frac{\eta^2}{\delta}, \frac{1284M(2)}{\kappa^2\eta} (D^2 + \log \frac{3n}{\delta})\}$ do

   4. if $2 \mid t$ then

      5. Perform do() and observe $X_t$ and $Y_t$. For $a = do()$, $D_a(t) = D_a(t - 1) + 1$.

      6. $\hat{\theta}_t = BGLM$-estimate($(X_1, Y_1), (X_3, Y_3), \cdots, (X_t, Y_t)$)

      7. for $a = do(X = x) \in A$ do

         8. Calculate $\hat{\mu}_{O,a}^t = \sigma(\hat{\theta}_t, X)$ and $\hat{\mu}_{I,a}^t = \frac{1}{D_a(t)} \sum_1^t \{a_j = a\} Y_j$, $D_a(t) = D_a(t - 1)$.

         9. update confidence bound $[L_a^t, U_a^t]$ for each arm $a$.

   10. end

   11. if $2 \nmid t$ then

       12. Perform $a$ such that $a_t = \text{arg max}_a 2 \cdot U_a^t - L_a^t$.

       13. for $a = do(X = x) \in A$ do

           14. $D_a(t) = D_a(t - 1) + \sum \{a_t = a\}$, $\hat{\theta}_t = \hat{\theta}_{t-1}$, $\hat{\mu}_{O,a} = \sigma(\hat{\theta}_t, X)$, $\hat{\mu}_{I,a} = \frac{1}{D_a(t)} \sum_1^t \{a_j = a\} Y_j$, $[L_a^t, U_a^t] = [L_{O,a}^t, U_{O,a}^t] \cap [L_{I,a}^t, U_{I,a}^t]$, $\hat{\mu}_a = (U_a^t + L_a^t)/2$

       15. end

   16. end

17. $(Uncertain, WellSampled, Rejected) = \text{Partition}(\{U_a^t, L_a^t \mid a \in A\}, \varepsilon)$

18. $t = t + 1$

19. end

20. Phase 2: Same as algorithm 1.

Algorithm 6: BGLM-estimate

Input: data pairs $((X_1, Y_1), (X_3, Y_3), \cdots, (X_t, Y_t))$

1. construct $(V_{t,X}, X_t)$ for each $X$, where $V_{t,X}$ is the value of parent of $X$ at round $t$, $X_t$ is the value of $X$ at round $t$.

2. for $X \in X \cup \{Y\}$ do

   3. $M_{t,X} = M_{t-2,X} + V_{t,X} V_{t,X}^T$, calculate $\hat{\theta}_{t,X}$ by solving the equation

      $\sum_{2i,1 \leq i \leq t} (X_i - f_X(V_{t,X} \hat{\theta}_{t,X})) V_{t,X} = 0$

4. end

5. return $\hat{\theta}_t$.

Lemma 1 ([6]). For any two weight vectors $\theta$ and $\theta'$, we have

$$|\sigma(\theta, X) - \sigma(\theta', X)| \leq \mathbb{E}_e \left[ \sum_{X' \in P_XY} |V_{X'}^T(\theta_{X'} - \theta'_{X'})| M(1) \right], \quad (13)$$

where $P_{X,Y}$ is the set of all nodes that lie on all possible paths from $X$ to $Y$ excluding $X$, $V_X$ is the value vector of a sample of the parents of $X$ according to parameter $\theta$, $M(1)$ is defined in Assumption 1, and the expectation is taken on the randomness of the noise term $e = (e_X)_{X \in X \cup \{Y\}}$ of causal model under parameter $\theta$.

For node $X \in X$, let $\ell_X$ be the number of nodes that lie on some paths from node $X$ to reward variable $Y$ excluding $X$, and $q_{d_a}^{(L)} = 1/\ell_X^2$ for $a = do(X = 1)$. Note that $\ell_X = |P_{X,Y}|$, so if $q_{d_a}^{(L)} = 1/\ell_X^2$ is small, the confidence bound would be large, which means that we need to re-sample.
a in Phase 2. Observation thresholds \( m^{(L)} \) and \( m_{c,\Delta}^{(L)} \) are defined as in Definitions 1 and 2, with respect to \( q_a^{(L)} \). Then we define the confidence bound for arm \( a \) as

\[
\beta_{Q}^2(t) = \frac{6M^{(1)}D^{1.5}}{\sqrt{t}} \sqrt{\frac{2}{q_a^{(L)} t} \log \frac{3n}{\delta}}, \tag{14}
\]

\[
\beta_I(u, t) = \sqrt{\frac{1}{2u} \log \frac{8nt^3}{\delta}}. \tag{15}
\]

\([L_{O,a}^t, U_{O,a}^t] = [\hat{\mu}_{O,a} - \beta_{Q}^2(t), \hat{\mu}_{O,a} + \beta_{Q}^2(t)]\). \( \hat{\mu}_{I,a} \) and \([L_{I,a}^t, U_{I,a}^t]\) are the same as general graph part. Same as general graph part, \([L_{a}^t, U_{a}^t]\) is defined by \([L_{O,a}^t, U_{O,a}^t] \cap [L_{I,a}^t, U_{I,a}^t]\).

Li et al. [13] first provide maximum likelihood estimation for generalized linear bandits, and the authors in [6] adapt it to the binary generalized linear model. However, their goals are to minimize the cumulative regret, while our goal is to minimize the sample complexity of algorithm in the pure exploration setting. It is non-trivial to integrate their results into our LUCB-based pure exploration framework. First, using Lemma 1, we can bound the difference between our estimate \( \hat{\sigma}(\theta^*, X) \) at round \( t \) and the true mean \( \sigma(\theta^*, X) \). Then, with the maximum likelihood estimation technique used in [13], we can use Lemma 9 to derive the interventional confidence bounds by bounding the right-hand side of Eq. (13), and integrate them into our algorithm framework. Since for \( a = do(X = 1) \), \( q_a^{(L)} = 1/P_X \) only depends on the graph structure and node \( X \), we can define corresponding \( m^{(L)} \) and \( m_{c,\Delta}^{(L)} \) with respect to \( q_a^{(L)} \), and integrate them into our previous algorithm framework. To our best knowledge, algorithm Causal-PE-BGLM is the first algorithm solving the pure exploration in BGLM utilizing the special features of causal bandits.

\section{Proof of Theorems}

\subsection{Proof of Theorem 1}

\textbf{Proof.} For convenience, we prove Causal-PE-parallel\((G, \varepsilon, \delta)\) outputs a \(3\varepsilon\)-optimal arm with probability \(1 - 3\delta\) and sample complexity

\[
T = O(H^{(P)}_{\varepsilon} \log \frac{nH^{(P)}_{\varepsilon}}{\delta}).
\]

By Chernoff bound [7], with probability at most \(\delta/3n\), at round \( t \) we have

\[
|\hat{q}_a - q_a| \geq \sqrt{\frac{6q_a}{t/2} \log \frac{6n}{\delta}}. \tag{16}
\]

By union bound,

\[
|\hat{q}_a - q_a| \leq \sqrt{\frac{6q_a}{t/2} \log \frac{6n}{\delta}} \tag{17}
\]

for all \( a \) with probability at least \( 1 - \frac{2n+1}{3n} \delta \geq 1 - \delta \). Let \( F_1 \) be the event that there exists an arm \( a \) such that \( |\hat{q}_a - q_a| > \sqrt{\frac{12n}{t} \log \frac{6n}{\delta}} \), then \( \Pr\{F_1\} < \delta \). Denote the optimal arm be \( a^* \) and \( \mu_{a^*} = \mu^* \).

The rest of proof is discussed based on the precondition that \( F_1 \) doesn’t happen. We provide a useful lemma.

\textbf{Lemma 2.} When \( q_{a^*} \geq \frac{\max\{\Delta_1, \varepsilon/2\}^2}{m_{c,\Delta} \max\{\Delta_1^2, (\varepsilon/2)^2\}} = \frac{1}{m_{c,\Delta}} \), during Phase 1, the algorithm takes at most \([T_1]\) rounds such that \( T_1 = 48 \frac{m_{c,\Delta}}{\max\{\Delta_1, \varepsilon/2\}^2} \log(\frac{8nT_1^3}{\delta}) \) with probability at least \( 1 - 2\delta\).

\textbf{Proof.} To prove this lemma, we first need to bound the probability for the event that some arms’ real mean are out of confidence bound. Similar to LUCB algorithm [11], we have
Lemma 3. $\sum_{i=1}^{\infty} \sum_{t=1}^{t} \exp (-2u\beta(u,t)^2) < \frac{\delta}{2n^4}$. 

Proof.

$$\sum_{i=1}^{\infty} \sum_{t=1}^{t} \exp (-2u\beta(u,t)^2) = \sum_{i=1}^{\infty} \sum_{u=1}^{u} \frac{\delta}{8nt^3}$$

$$= \sum_{i=1}^{\infty} \frac{\delta}{8nt^2}$$

$$< \frac{\delta}{8n \cdot (6/\pi^2)}$$

$$< \frac{\delta}{2n + 1}. \quad (18)$$

where the last inequality holds when $n > 3$. Thus define $F_2$ be the event that during the algorithm there’s an arm’s real mean are out of confidence bound. Then by Hoeffding’s inequality and union bound with inequality (18)

$$\Pr\{F_2\} = (2n + 1) \sum_{i=1}^{\infty} \sum_{t=1}^{t} \exp (-2u\beta(u,t)^2) < (2n + 1) \cdot \frac{\delta}{2n + 1} = \delta. \quad (19)$$

Now assume Phase I don’t terminate after $t = \lceil T_1 \rceil$ rounds, $F_1$ and $F_2$ don’t happen. Then $|\hat{q}_a - q_a| \leq \sqrt{\frac{6q_a}{t/2} \log \frac{6n}{\delta}}$. Thus for any arm $a$ such that $q_a \geq \frac{\max(\Delta_a, \epsilon/2)^2}{m_c, \Delta \max(\Delta_a, \epsilon/2)^2}$, we have

$$\sqrt{\frac{6q_a}{t/2} \log \frac{6n}{\delta}} \leq \sqrt{\frac{6q_a \max(\Delta, \epsilon/2)^2}{24 \cdot m_c, \Delta}}$$

$$\leq \frac{1}{2} \sqrt{q_a \cdot q_a} = \frac{q_a}{2}.$$ 

Thus $\hat{q}_a \geq q_a - \frac{q_a}{2} \geq \frac{q_a}{2} \geq \frac{\max(\Delta_a, \epsilon/2)^2}{2m_c, \Delta \max(\Delta_a, \epsilon/2)^2}$. Hence

$$T_a(t) \geq \hat{q}_a \cdot \frac{t + 1}{2} \geq \frac{\max(\Delta, \epsilon/2)^2}{2m_c, \Delta \max(\Delta_a, \epsilon/2)^2} \cdot 24 \cdot \frac{m_c, \Delta}{\max(\Delta_a, \epsilon/2)^2} \cdot \log(\frac{8nt^3}{\delta})$$

$$= \frac{12}{\max(\Delta_a, \epsilon/2)^2} \log(\frac{8nt^3}{\delta}), \quad (20)$$

where the inequality (20) is because we perform do() once for $2$ rounds. Thus

$$\beta(T_a(t), t) = \sqrt{\frac{1}{2T_a(t)} \log \frac{8nt^3}{\delta}}$$

$$\leq \frac{\max(\Delta_a, \epsilon/2)}{4}.$$ 

Since $a^*$ satisfies $q_{a^*} \geq \frac{\max(\Delta, \epsilon/2)^2}{m_c, \Delta \max(\Delta_a, \epsilon/2)^2}$, we also have $\beta(T_{a^*}(t), t) \leq \frac{\max(\Delta_a, \epsilon/2)}{4}$. Since $\beta(T_a(t), t) = \frac{T_a(t) - T_{a^*}(t)}{2}$, we prove a lemma:

Lemma 4. Suppose an optimal arm $a^*$ such that $\frac{T_{a^*} - T_a^*}{2} \leq \frac{\max(\Delta_a, \epsilon/2)}{4}$, if $F_2$ doesn’t happen, for any $a$ such that $\frac{T_a(t) - T_a^*}{2} \leq \frac{\max(\Delta_a, \epsilon/2)}{4}$, we have $a \preceq a^*$. 

17
Proof.

\[ U_a = \hat{\mu}_a + \beta(T_a(t), t) \leq \mu_a + 2\beta(T_a(t), t) \]
\[ \leq \mu_a + \frac{\max\{\Delta_a, \varepsilon/2\}}{2} \]
\[ = \mu_a - \Delta_a + \max\{\Delta_a, \varepsilon/2\}/2. \]

If \( \Delta_a \leq \varepsilon/2 \), we know \( \Delta_{a^*} \leq \Delta_a \leq \varepsilon/2 \). Thus
\[ \mu_{a^*} - \Delta_a + \max\{\Delta_a, \varepsilon/2\}/2 \leq \mu_{a^*} + \frac{\varepsilon}{4} \]
\[ = \mu_{a^*} - \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{2} + \frac{\varepsilon}{2} \]
\[ \leq \mu_{a^*} - 2\beta(T_{a^*}(t), t) + \frac{\varepsilon}{2} \leq L_{a^*} + \frac{\varepsilon}{2}. \]

If \( \Delta_a > \varepsilon/2 \),
\[ \mu_{a^*} - \Delta_a + \max\{\Delta_a, \varepsilon/2\}/2 = \mu_{a^*} - \frac{\Delta_a}{2} \]
\[ \leq \mu_{a^*} - \frac{\Delta_{a^*}}{2} - \frac{\varepsilon}{4} + \frac{\varepsilon}{4} \]
\[ \leq \mu_{a^*} - \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{2} + \frac{\varepsilon}{4} \]
\[ \leq L_{a^*} + \frac{\varepsilon}{4}. \]

Hence \( a^* \succeq a \), which means \( a \) is in \( \text{Rejected} \) at \( t \) rounds.

Now for all arms \( a \neq a^* \) such that \( q_a \geq \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}} \), they are in \( \text{Rejected} \) at \( t \) rounds. By the definition of \( m_{e, \Delta} \), there’re at most \( m_{e, \Delta} + 1 \) arms in set \( \text{Uncertain} \). (The extra +1 is because the optimal arm \( a^* \) is also contained in \( \text{Uncertain} \)). We have \( t \geq 48m_{e, \Delta} \log \frac{6n}{\delta} \geq 36|\text{Uncertain}| \log\left(\frac{6n}{\delta}\right) \) and the remaining arm \( a \) (except \( a^* \)) must satisfies:
\[ q_a < q_{a^*} + \sqrt{\frac{6q_{a^*}}{t/2} \log\left(\frac{6n}{\delta}\right)} < \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}} + \sqrt{\frac{1}{m_{e, \Delta}} < \frac{2}{|\text{Uncertain}| - 1}}, \]
which satisfies the stopping rule. Hence with probability at least \( 1 - \Pr\{F_1\} - \Pr\{F_2\} \geq 1 - 2\delta\), Phase 1 must be terminated after \( t \) rounds. \( \square \)

Lemma 5. When \( q_{a^*} < \frac{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}}{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}} \), during Phase 1, the algorithm takes at most \( |T_2| \) rounds such that \( T_2 = \frac{4 \log^{\frac{3}{\delta}}}{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}} \) with probability at least \( 1 - 2\delta \).

Proof. In the proof we assume \( F_1 \) and \( F_2 \) don’t happen. The probability is at least \( 1 - 2\delta \). Assume when \( t = |T_2| \) the Phase 1 don’t terminate at \( t \) rounds.

Since \( f(x) = \log^{\frac{x}{(8x^3/\delta)}} \) is a increasing function, \( t \geq 48 \frac{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}}{\max\{\Delta_{a^*}, \varepsilon/2\}} \log \frac{8nT_2^2}{\delta} \) for any \( t \in [T_1, T_2] \).

From above, for arm \( a \) such that \( q_a \geq \frac{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}}{m_{e, \Delta} \max\{\Delta_{a^*}, \varepsilon/2\}} \), \( \beta(T_a(t), t) \leq \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{4} \). Then during \( T \in [T_1, T_2] \),
\[ \hat{\mu}_a + 3\beta(T_a(t), t) \leq \mu_a + 4\beta(T_a(t), t) \leq \mu_{a^*} - \Delta_a + \max\{\Delta_a, \varepsilon/2\} \]
\[ \leq \mu_{a^*} + \frac{\varepsilon}{2}. \] (21)

If \( \beta(T_{a^*}(t), t) < \varepsilon/4 \) for some \( t \leq T_2 \), at \( t \) rounds,
\[ U_a \leq \mu_{a^*} + \frac{\varepsilon}{4} \leq \mu_{a^*} - 2\beta(T_{a^*}(t), t) + \frac{3\varepsilon}{4} \leq L_{a^*} + \varepsilon. \]
Thus the stopping rule is satisfied.

Now assume $\beta(T_{a^*}(t), t) \geq \varepsilon/4$ for any $t \leq T_2$, then

$$\hat{\mu}_a + 3\beta(T_a(t), t) \geq \mu_{a^*} + 2\beta(T_{a^*}(t), t) \geq \frac{\varepsilon}{2} \geq \hat{\mu}_a + 3\beta(T_a(t), t),$$

where the last inequality is from inequality (21). Thus for during $T \in [T_1, T_2]$, the interventions will only act on all arm $a$ such that $q_a < \frac{\max(\Delta_1, \varepsilon/2)^2}{m_{c, \Delta}}$. Denote these arms by $S$. By definition of $s$, $|S| \leq m_{c, \Delta}$ and

$$\sum_{a \in S} T_a(t) \geq \frac{(64 - 16) m_{c, \Delta}}{2 \max(\Delta_1, \varepsilon/2)^2} \log(8nT_2^3/\delta) = \frac{8m_{c, \Delta}}{\max(\Delta_1, \varepsilon/2)^2} \log(8nT_2^3/\delta).$$

Note that $a^*$ is optimal arm, then if for some non-optimal arm $a'$, $T_a(t) \geq T_{a^*}(t)$, then $\hat{\mu}_a + 3\beta(T_a(t), t) < \mu_{a^*} + 3\beta(T_{a^*}(t), t)$. Hence there’s at least one optimal arm $a^*$ such that

$$T_{a^*}(t) \geq \frac{1}{m_{c, \Delta}} \cdot \frac{8m_{c, \Delta}}{\max(\Delta_1, \varepsilon/2)^2} \log(8nT_2^3/\delta) = \frac{8}{\max(\Delta_1, \varepsilon/2)^2} \log(8nT_2^3/\delta).$$

Thus by inequality (22) and definition of $\beta(u, t)$,

$$\beta(T_{a^*}(t), t) \leq \frac{\max(\Delta_1, \varepsilon/2)}{4},$$

which satisfies the stopping rule Lemma 4. Thus the Phase 1 must terminate. \hfill \Box

From two Lemma 2 and Lemma 5 above, Phase 1 will always terminate when sample complexity $t = [T_2]$ satisfies $T_2 = 64 \frac{m_{c, \Delta}}{\max(\Delta_1, \varepsilon/2)^2} \log(8nT_2^3/\delta)$.

We then prove a lemma similar to [10] to change format of the bound.

**Lemma 6.** Suppose $t \leq 80Q \log(8n^3/\delta)$, then $t = O(Q \log(Qn/\delta))$.

**Proof.** Since $f(x) = \frac{x}{\log(8nx^3/\delta)}$ is a increasing function and $f(t) \leq 80Q$, we only need to show there exists a constant $C$ such that $f(CQ \log(Qn/\delta)) \geq 80Q$.

$$80Q \log\left(\frac{8n(CQ \log(Qn/\delta))^3}{\delta}\right)$$

$$\leq 80Q \log(8n) + 240Q \log\left(\frac{1}{\delta}\right) + 240Q \log(CQ \log(Qn/\delta))$$

$$< 80Q \log\left(\frac{8n}{\delta}\right) + 240Q \log C + 240Q \log Q + 240Q \log(Qn/\delta)$$

$$< (480 + 240 \log C)Q \log(Qn/\delta).$$

Hence we can choose $C = 2400$, since $C - 240 \log C > 480$. \hfill \Box

Now

$$T_2 = O\left(\frac{\max(\Delta_1, \varepsilon/2)^2}{2400^2} \log\left(\max\left\{\frac{m_{c, \Delta} \max(\Delta_1, \varepsilon/2)^2}{\delta}\right\}\right)^3\right).$$

Thus for $Q = \frac{m_{c, \Delta}}{\max(\Delta_1, \varepsilon/2)^2}$, by Lemma 6

$$T_2 = O\left(\sqrt{Q} \log\left(\frac{Qn}{\delta}\right)\right).$$

Now we consider the Phase 2. At most one arms in remaining set $\text{Uncertain}$ such that $\hat{q}_a \geq \frac{2}{|\text{Uncertain}|-1}$, thus at least $|\text{Uncertain}| - 1$ arms satisfy $q_a - \sqrt{\frac{\log 6}{t/2}} \leq \hat{q}_a \leq \frac{2}{|\text{Uncertain}|-1}$, since $g(x) = x - \sqrt{\frac{12x}{t}} \log \frac{6n}{\delta}$ is a increasing function when $g(x) \geq 0$, thus from $t \geq
36] Uncertain \[ \log \frac{6n}{\delta} \] we can deduce that \( g(\frac{3}{|\text{Uncertain}| - 1}) \geq \frac{2}{|\text{Uncertain}| - 1} \geq g(q_a), \) which means \( q_a \leq \frac{3}{|\text{Uncertain}| - 1}. \)

then we prove |Uncertain| \leq 3m + 1. Recall \( m = \min \{ \tau : |\{ q_a | q_a < \frac{1}{\tau} \}| \leq \tau \}. \) Since

\[ |\{ q_a | q_a < \frac{1}{(|\text{Uncertain}| - 1)/3} \}| \geq |\text{Uncertain}| - 1 > |\text{Uncertain}| - 1 \]

we have \( m \geq \frac{|\text{Uncertain}| - 1}{3}, |\text{Uncertain} \cup \{ a'' \}| \leq 3m + 2, \) where \( a'' \) is the arm with highest upper confidence bound in WellSampled.

Now we prove in Uncertain \( \cup \{ a'' \}, \) there is at least one \( 2\varepsilon \)-optimal arm. By the definition of Rejected, for any arm \( a \in \text{Rejected}, \) there’s another arm \( a' \) such that \( a' \geq a. \) For arm \( a_1, a_2, a_3 \in \text{Rejected} \) such that \( a_1 \leq a_2, a_2 \leq a_3, \) we have

\[ L_{a_3}^t = U_{a_2}^t - \varepsilon \geq L_{a_2}^t - \varepsilon \geq U_{a_1}^t - \varepsilon, \]

thus \( a_1 \leq a_3. \) Now for optimal arm \( a^*, \) we consider three cases:

**Case 1:** If \( a^* \in \text{Rejected}, \) then \( a^* \leq a_1 \leq a_2 \cdots \leq a_k, \) where \( a_1, \cdots, a_{k-1} \in \text{Rejected}, a_k \notin \text{Rejected}. \) Thus \( a^* \leq a_k, \) which means \( U_{a^*} - L_{a_k} \leq \varepsilon. \)

If \( a_k \in \text{WellSampled}, \) we know \( \mu_{a''} \geq U_{a''} - \varepsilon/2 \geq U_{a_k} - \varepsilon/2 \geq L_{a_k} - \varepsilon/2 \geq U_{a^*} - \frac{3}{2}\varepsilon \geq \mu_{a^*} - \frac{3}{2}\varepsilon, \)

thus \( a'' \) is \( \frac{3}{2} \varepsilon \)-optimal. (The first inequality is because \( a'' \in \text{WellSampled}. \))

If \( a_k \in \text{Uncertain}, \) since \( a' \) is a \( \varepsilon \)-optimal arm in Uncertain, we know \( \mu_{a_k} \geq \mu_{a^*} - \varepsilon. \) Thus \( a_k \) is \( \varepsilon \)-optimal.

**Case 2:** If \( a^* \in \text{Uncertain}, \) Uncertain has an optimal arm.

**Case 3:** If \( a^* \in \text{WellSampled}, \) then \( \mu_{a''} \geq U_{a''} - \varepsilon/2 \geq U_{a^*} - \varepsilon/2 \geq \mu_{a^*} - \varepsilon /2. \)

Thus at least one arm in Uncertain \( \cup \{ a'' \} \) is \( 2\varepsilon \)-optimal. Hence by classical LUCB algorithm [11], with probability \( 1 - \delta, \) Phase 2 has sample complexity

\[ T_3 = O(H_{\varepsilon/2,m} \log \frac{H_{\varepsilon/2,m}'}{\delta}), H_{\varepsilon/2,m}' = \sum_{i=1}^{3m+2} \frac{1}{\max\{\Delta_i, \varepsilon /2\}}, \]

\[ \leq 3 \sum_{i=1}^{3m+2} \frac{1}{\max\{\Delta_i, \varepsilon /2\}}, \quad (24) \]

where \( \Delta_i = \Delta_i - 2\varepsilon. \) We denote \( F_3 \) be the event that LUCB algorithm doesn’t terminate with a \( \varepsilon \)-optimal arm in Uncertain, then \( P(F_3) \leq \delta. \)

In total, by (23) and (24), with probability \( 1 - 3\delta \) \((F_1, F_2 \text{ and } F_3), \) for \( H_{\varepsilon} = Q + H_{\varepsilon/2,m}', \) the sample complexity is

\[ T = T_2 + T_3 = O(Q \log \frac{Qn}{\delta} + H_{\varepsilon/2,m}' \log \frac{H_{\varepsilon/2,m}'}{\delta}) \]

\[ = O(H_{\varepsilon} \log \frac{nH_{\varepsilon}}{\delta}). \]

Replace \( \delta \) to \( \delta/3 \) and \( \varepsilon \) to \( \varepsilon/3, \) we get the sample complexity in statement of Theorem 1.  

D.2 Proof of Theorem 2

**Proof.** For convenience, we prove Causal-PE-general \((G, \varepsilon, \delta)\) outputs a \( 3\varepsilon \)-optimal arm with probability \( 1 - 5\delta. \) In round \( t, T_{a,z}(t) = \sum_{j=1}^t \mathbb{I}\{ 2 \mid j, X_{j,i} = x, Pa(X)_j = z \}, \) \( \tilde{q}_{a,z} = \frac{T_{a,z}(t)}{t+1}. \) By Chernoff bound, at round \( t, \) with probability at most \( \delta/3nZ, \)

\[ |\tilde{q}_{a,z} - q_{a,z}| > \sqrt{\frac{6q_{a,z}}{t/2} \log \frac{6nZ}{\delta}}. \]

By union bound, with probability \( 1 - \delta \)

\[ |\tilde{q}_{a,z} - q_{a,z}| \leq \sqrt{\frac{6q_{a,z}}{(t + 1)/2} \log \frac{6nZ}{\delta}}. \]
Hence

\[ \hat{q}_a = \min_z \{q_{a,z} \} \leq \min_z \{q_{a,z} + \sqrt{\frac{6q_{a,z}}{t/2} \log \frac{6nZ}{\delta}} \} = q_a + \sqrt{\frac{12q_a}{t} \log \frac{6nZ}{\delta}}. \]  

(25)

When \( q_a \geq \frac{3}{7} \log \frac{6nZ}{\delta} \), \( f(x) = x - \sqrt{\frac{12x}{t} \log \frac{6nZ}{\delta}} \) is a increasing function.

\[ \hat{q}_a \geq \min_z \{q_{a,z} - \sqrt{\frac{12q_{a,z}}{t} \log \frac{6nZ}{\delta}} \} = q_a - \sqrt{\frac{12q_a}{t} \log \frac{6nZ}{\delta}}. \]  

(26)

Let \( F_1 \) be the event that at least one of above inequalities doesn’t hold, then \( \Pr \{ F_1 \} \leq \delta \). Now let \( F_2 \) and \( F_3 \) be the event that during some round \( t \), the true mean of an arm is out of range \([L_{1,a}, U_{1,a}]\) and \([L_{2,a}, U_{2,a}]\) respectively. Following lemma proves \( \Pr \{ F_3 \} \leq \delta \) with union bound.

**Lemma 7.** \( \sum_{t=1}^{\infty} \sum_{u=1}^{l_t} \exp(-2u\beta I(u, t)^2) < \frac{\delta}{2\alpha+1} \).

**Proof.** The proof is similar to Lemma 3. \( \square \)

To prove \( \Pr \{ F_2 \} \leq \delta \), we need the lemma in the paper.

**Lemma 8.** In round \( t \), with probability \( 1 - \frac{\delta}{8nT_3} \),

\[ |\hat{\mu}_{\text{obs}, a} - \mu_a| < \sqrt{\frac{Z^2}{t/2} \log \frac{16nZt^3}{\delta}} + \sqrt{\frac{1}{2T_a(t)} \log \frac{16nZt^3}{\delta}}. \]  

(27)

**Proof.** By Hoeffding’s inequality, for \( a = do(X = x) \),

\[ |r_{a,z}(t) - P(Y = 1 \mid X = x, z = z)| > \sqrt{\frac{1}{2T_{a,z}(t)} \log \frac{16nZt^3}{\delta}}, \]

\[ |p_{a,z}(t) - P(Pa(X) = z)| > \sqrt{\frac{1}{2nZt/2 \log \frac{16nZt^3}{\delta}}}, \]

at round \( 2t \) with probability \( 1 - 2Z \cdot \frac{\delta}{16nZt^3} = 1 - \frac{\delta}{8nT_3} \). Hence we get

\[ \hat{\mu}_{1,a} = \sum_z r_{a,z}(t) \cdot p_{a,z}(t) \]

\[ \leq \sum_z P(Y = 1 \mid X = x, Pa(X) = z) p_{a,z}(t) + \sum_z p_{a,z} \sqrt{\frac{1}{2T_{a,z}(t)} \log \frac{16nZt^3}{\delta}} \]

\[ \leq \sum_z P(Y = 1 \mid X = x, Pa(X) = z) p_{a,z}(t) + \sqrt{\frac{1}{2T_a(t)} \log \frac{16nZt^3}{\delta}} \]

\[ \leq \sum_z P(Y = 1 \mid X = x, Pa(X) = z) P(Pa(X) = z) + \sum_z \sqrt{\frac{1}{2T_a(t)} \log \frac{16nZt^3}{\delta}} \]

\[ \leq \mu_a + \sqrt{\frac{Z^2}{t} \log \frac{16nZt^3}{\delta}} + \sqrt{\frac{1}{2T_a(t)} \log \frac{16nZt^3}{\delta}}. \]

\( \square \)

Thus by union bound, \( \Pr \{ F_2 \} \leq \delta \). In later proof, we will always assume that \( F_1 \), \( F_2 \) and \( F_3 \) don’t happen. In this case, true mean \( \mu_a \) is \([L_{1,a}, U_{1,a}]\) for all rounds \( t \). Now we consider the optimal arm \( a^* \)
such that \( q_a' \geq \frac{\max(\Delta_1, \epsilon/2)^2}{m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \). In this condition, we prove Phase 1 will terminate after \( t = [T_1] \) rounds, where \( T_1 \geq 64((Z^2/\max(\Delta_1, \epsilon/2)^2) + \frac{m_{e, \Delta}}{\max(\Delta_2, \epsilon/2)^2}) \log(16nT^3/\delta) \). For all arm \( a \) such that \( q_a \geq \frac{\max(\Delta_1, \epsilon/2)^2}{m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \), we have

\[
q_a \geq \frac{\max(\Delta_1, \epsilon/2)^2}{m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \geq \frac{3}{t} \log \frac{6nZ}{\delta}.
\]

Since \( F_1 \) don’t happen, \( |\dot{q}_a - q_a| \leq \sqrt{\frac{12\epsilon}{t}} \log \frac{6nZ}{\delta} \). Then by (26)

\[
\sqrt{\frac{12\epsilon}{t}} \log \frac{6nZ}{\delta} \leq \frac{q_a}{2},
\]

we have \( \dot{q}_a \geq q_a - \sqrt{\frac{12\epsilon}{t}} \log \frac{6nZ}{\delta} \geq \frac{\max(\Delta_1, \epsilon/2)^2}{2m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \). Hence

\[
T_a(t) = q_a \left( 1 + \frac{t}{2} \right) \geq \frac{24}{\max(\Delta_a, \epsilon/2)^2} \log \frac{16nT^3}{\delta}.
\]

Thus

\[
\beta_O(T_a(t), t) = \sqrt{\frac{1}{2T_a(t)} \log \frac{16nT^3}{\delta}} + \sqrt{\frac{Z^2}{t} \log \frac{16nT^3}{\delta}} \leq \frac{\max(\Delta_a, \epsilon/2)}{\sqrt{64}} + \frac{\max(\Delta_1, \epsilon/2)}{\sqrt{64}} \leq \frac{\max(\Delta_a, \epsilon/2)}{4}.
\]

Then according to Lemma 4, we know \( a^* \geq a \), which means \( a \) is in Rejected at \([T_1]\) rounds. By definition of \( m_{e, \Delta} \), there’re at most \( m_{e, \Delta} + 1 \) arms in set Uncertain. Then \( t \geq 48m_{e, \Delta} \log(6n/\delta) > 36|\text{Uncertain}| \log(6n/\delta) \), the remaining arms (except \( a^* \)) must satisfies:

\[
\dot{q}_a \leq q_a + \sqrt{\frac{6q_a}{t/2}} \log \frac{6n}{\delta} < \frac{1}{m_{e, \Delta}} + \frac{1}{m_{e, \Delta}} < \frac{2}{m_{e, \Delta}} < \frac{2}{|\text{Uncertain}| - 1},
\]

which satisfies the stopping rule.

Now we consider \( q_a < \frac{\max(\Delta_1, \epsilon/2)^2}{m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \), assume \( T_2 \) satisfies \( T_2 = 80(\frac{m_{e, \Delta} + Z^2}{\max(\Delta_a, \epsilon/2)^2}) \log \frac{8nT^3}{\delta} \) and assume Phase 1 doesn’t terminate after \([T_2]\) rounds. From above, during \( t \in [T_1, T_2] \), for arm \( a \) such that \( q_a \geq \frac{\max(\Delta_1, \epsilon/2)^2}{m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \), \( \beta_O(T_a(t), t) \leq \frac{\max(\Delta_a, \epsilon/2)}{4} \).

\[
2 \cdot U_a^t - L_a^t \leq \mu_a + 2(U_a^t - L_a^t) = \mu_a + 4\beta_O(T_a(t), t) \leq \mu_a^* - \Delta_a + \max(\Delta_a, \epsilon/2) \leq \mu_a^* + \frac{\epsilon}{2}.
\]

If \( U_a^* - L_a^* < \epsilon/2 \) for some \( t \leq T_2 \), we know the confidence bound will not be updated. Thus

\[
U_a^t \leq \mu_a + [U_a^t - L_a^t] \leq \mu_a^* + \beta_O(T_a(t), t) \leq \mu_a^* + \frac{\epsilon}{4} \leq \mu_a^* - (U_a^t - L_a^t) + \frac{3}{4} \epsilon \leq L_a^t + \frac{3}{4} \epsilon
\]

which satisfies the stopping rule. If \((U_a^t - L_a^t) \geq \epsilon/2\), for any \( t \leq T_2 \),

\[
2 \cdot U_a^t - L_a^* \geq \mu_a^* + (U_a^t - L_a^t) \geq \mu_a^* + \frac{\epsilon}{2} \geq 2 \cdot U_a^t - L_a^t.
\]

Now for \( t \in [T_1, T_2] \), the interventions will only act on all arms \( a \) such that \( q_a < \frac{\max(\Delta_1, \epsilon/2)^2}{m_{e, \Delta} \max(\Delta_a, \epsilon/2)^2} \). Denoted these arms by \( S \). By definition of \( m_{e, \Delta} \), \( |S| \leq m_{e, \Delta} \) and
\[
\sum_{a \in S} D_a([T_2]) \geq \frac{(80 - 64)(m_{\epsilon, \Delta} + Z^2)}{2 \max\{\Delta_1, \epsilon/2\}^2} \log(16nZT_2^3) \geq \frac{8m_{\epsilon, \Delta}}{\max\{\Delta_1, \epsilon/2\}^2} \log(16nZT_2^3).
\]

Similar to the proof of Theorem 1,
\[
D_{a^*}([T_2]) \geq \frac{1}{m_{\epsilon, \Delta}} \cdot \frac{8m_{\epsilon, \Delta}}{\max\{\Delta_1, \epsilon/2\}^2} \log(16nZT_2^3) = \frac{8}{\max\{\Delta_1, \epsilon/2\}^2} \log(16nZT_2^3).
\]

Thus
\[
U_{a^*} - L_{a^*} \leq \beta_1(D_{a^*}(T_2), T_2) \leq \frac{\max\{\Delta_1, \epsilon/2\}}{4}.
\]

From Lemma 4, Phase 1 must terminate. Thus for \( T_2 = 64 \frac{m_{\epsilon, \Delta} + Z^2}{\max\{\Delta_1, \epsilon/2\}^2} \log(16nZT_2^3) \), Phase 1 must terminate after \([T_2]\) rounds.

Hence for \( Q = \frac{m_{\epsilon, \Delta} + Z^2}{\max\{\Delta_1, \epsilon/2\}^2} \), the sample complexity of Phase 1 is bounded by \( T_2 = O(Q \log QnZ) \).

For analysis of Phase 2, the proof is just a simple rewrite of proof for Theorem 1. The sample complexity of Phase 2 is \( T_3 = O(H_m^{\epsilon/2} \log(\frac{H_m^{\epsilon/2}}{\delta})) \).

In total, with probability \( 1 - 5\delta \) (\( F_1, F_2, F_3 \), Phase 2), for \( H_m^{\epsilon/2} = Q + H_m^{\epsilon/2} = O(Q + H_m^{\epsilon/2} + R) \), the sample complexity is
\[
T = T_2 + T_3 = O(H_m^{\epsilon/2} \log \frac{nZH_m^{\epsilon/2}}{\delta}),
\]
and the algorithm output a \( 3\epsilon \)-optimal arm (the proof is the same as the proof of Theorem 1). Substitute \( \delta \) to \( \delta/4 \) and \( \epsilon \) to \( \epsilon/3 \), we prove Theorem 2. \( \square \)

D.3 Proof of Theorem 3

Proof. We only need to prove that with probability \( 1 - 4\delta \), Causal-PE-BGLM(\( G, \epsilon, \delta \)) outputs a \( 3\epsilon \)-optimal arm with sample complexity
\[
O(H_m^{\epsilon/2} \log \frac{nH_m^{\epsilon/2}}{\delta}).
\]

We first provide a lemma in [14] to show the confidence for the linear regression.

Lemma 9 ([6]). For one node \( X \in X \cup \{Y\} \), assume Assumption 1 and 2 holds, and
\[
\lambda_{\text{min}}(M_{t,X}) \geq \frac{512 D(M^{2})^2}{\kappa^4} (D^2 + \ln \frac{3n}{\delta}),
\]
with probability \( 1 - \delta \), for any vector \( v \in \mathbb{R}^{P_\text{Pa}(X)} \), at all rounds \( t \) the estimator \( \hat{\theta}_{t,X} \) in Algorithm 6 satisfy
\[
|v^\top (\hat{\theta}_{t,X} - \theta_X^*)| \leq \frac{3}{\kappa} \sqrt{\log(3n/\delta)}|v|_{M_{t,X}^{-1}}.
\]

Since we need to estimate \( \theta_{t,X} \) for all nodes, let \( F_1 \) be the event that the above inequality doesn’t hold, then by union bound, \( \Pr \{ F_1 \} \leq \delta \). Now from [6], the true mean \( \sigma(\theta, X_i) \) and our estimation \( \sigma(\hat{\theta}, X_i) \) can be bounded by Lemma 1. We rewrite the Lemma 1 here.

Lemma 1 ([6]). For any two weight vectors \( \theta \) and \( \theta' \), we have
\[
|\sigma(\theta, X) - \sigma(\theta', X)| \leq \mathbb{E}_e \sum_{X' \in P_{X,Y}} |V_{X'}(\theta_X' - \theta'_X)| M^{(1)}(X'),
\]
where \( P_{X,Y} \) is the set of all nodes that lie on all possible paths from \( X \) to \( Y \) excluding \( X \), \( V_X \) is the value vector of a sample of the parents of \( X \) according to parameter \( \theta \), \( M^{(1)} \) is defined in Assumption 1, and the expectation is taken on the randomness of the noise term \( e = (e_X)_{X \in X \cup \{Y\}} \) of causal model under parameter \( \theta \).
Thus when $\lambda = d0(X = 1)$, $|P_XX| = \ell_X \in \{0, 1, \cdots, n\}$. We then introduce Lecué and Mendelson’s Inequality represented in [19].

**Lemma 10** ([19] Lecué and Mendelson’s Inequality). Let random column vector $v \in \mathbb{R}^D$, and $v_1, \cdots, v_n$ are $n$ independent copies of $v$. Assume $z \in \text{Sphere}(D)$ such that

$$\Pr[|v^Tz| > \alpha^{1/2}] \geq \beta,$$

then there exists a constant $c > 0$ such that when $n \geq \frac{\epsilon D}{\beta},$

$$\Pr \left[ \lambda_{\min} \left( \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T \right) \leq \frac{\alpha \beta}{2} \right] \leq e^{-n\beta^2/c}.$$

This lemma can help us to bound the minimum eigenvalue for $M_{t,X} = \sum_{t,i,t \leq i} V_{t,X} V_{t,X}^T$. To satisfy the condition for Lemma 10, we provide a similar lemma in [6]:

**Lemma 11.** Under Assumption 3, for any node $X \in \mathcal{X}$ and $v \in \text{Sphere}(|Pa(X)|),$

$$\Pr \left[ |Pa(X) \cdot z| > \frac{1}{\sqrt{4D^2-3}} \right] \geq \eta.$$

**Proof.** The proof is very similar to [6]. Let $|Pa(X)| = d \leq D$, $z = (z_1, z_2, \cdots, z_d)$. Let $Pa(X) = (X_{i_1} = X_1, X_{i_2}, \cdots, X_{i_d})$ and $pa(X) = (x_{i_1} = 1, x_{i_2}, \cdots, x_{i_d})$. We denote $d_0 = \sqrt{d-1} + \frac{1}{2\sqrt{d-1}}$. If $|z_1| \geq \frac{d_0}{\sqrt{d_0^2+1}}$, then by Cauchy-Schwarz inequality, we can deduce that

$$|pa(X) \cdot v| \geq \sum_{i=2}^{d} |z_i|$$

$$\geq \frac{d_0}{\sqrt{d_0^2+1}} - \sqrt{(d-1) \sum_{i=2}^{d} |z_i|^2}$$

$$\geq \frac{d_0}{\sqrt{d_0^2+1}} - \sqrt{(d-1)(1 - \frac{d_0^2}{d_0^2+1})}$$

$$= \frac{1}{2\sqrt{(d_0^2+1)(d-1)}}$$

$$= \frac{1}{4d^2 - 3}.$$

Thus when $|z_1| \geq \frac{d_0}{\sqrt{d_0^2+1}}$, $|Pa(X) \cdot z| > \frac{1}{4d^2 - 3} \geq \frac{1}{4d^2 - 3}$. If $|z_1| < \frac{d_0}{\sqrt{d_0^2+1}}$, assume $|z_2| = \max_{2 \leq i \leq d} |z_i|$, then

$$|z_2| \geq \frac{1}{\sqrt{d-1}} \sqrt{\sum_{i=2}^{d} |z_i|^2} \geq \sqrt{\frac{1-(d_0/\sqrt{d_0^2+1})^2}{d-1}} = \frac{1}{\sqrt{4d^2 - 3}}.$$

By Assumption 3

$$\Pr\{X_{i_1} = 1, X_{i_2} = x_{i_2}, \cdots, X_{i_d} = x_{i_d}\}$$

$$= \Pr\{X_{i_2} = x_{i_2} | X_{i_1} = 1, X_{i_3} = x_{i_3}, \cdots, X_{i_d} = x_{i_d}\} \cdot \Pr\{X_{i_1} = 1, X_{i_3} = x_{i_3}, \cdots, X_{i_d} = x_{i_d}\}$$

$$\geq \eta \cdot \Pr\{X_{i_1} = 1, X_{i_3} = x_{i_3}, \cdots, X_{i_d} = x_{i_d}\}.$$
where the last inequality is because

we have

\[ \Pr\left\{ |\mathcal{P}(X)| \cdot \zeta \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} \]

\[ = \sum_{x_{i_3}, \ldots, x_{i_d}} \Pr\{X_{i_1} = 1, X_{i_2} = 1, X_{i_3} = x_{i_3} \ldots, X_{i_d} = x_{i_d}\} \cdot \mathbb{I}\left\{ |(1, 1, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} \]

\[ + \sum_{x_{i_3}, \ldots, x_{i_d}} \Pr\{X_{i_1} = 1, X_{i_2} = 0, X_{i_3} = x_{i_3} \ldots, X_{i_d} = x_{i_d}\} \cdot \mathbb{I}\left\{ |(1, 0, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} \]

\[ \geq \eta \sum_{x_{i_3}, \ldots, x_{i_d}} \Pr\{X_{i_1} = 1, X_{i_3} = x_{i_3} \ldots, X_{i_d} = x_{i_d}\} \cdot \mathbb{I}\left\{ |(1, 1, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} \]

\[ + \eta \sum_{x_{i_3}, \ldots, x_{i_d}} \Pr\{X_{i_1} = 1, X_{i_3} = x_{i_3} \ldots, X_{i_d} = x_{i_d}\} \cdot \mathbb{I}\left\{ |(1, 0, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} \]

\[ \geq \eta. \]

where the last inequality is because

\[ \sum_{x_{i_3}, \ldots, x_{i_d}} \Pr\{X_{i_1} = 1, i_3 = x_{i_3} \ldots, X_{i_d} = x_{i_d}\} = 1, \]

and

\[ \left( \mathbb{I}\left\{ |(1, 1, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} + \mathbb{I}\left\{ |(1, 0, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| \geq \frac{1}{\sqrt{4D^2 - 3}} \right\} \right) \geq 1. \]

The above equation is because otherwise

\[ |z_2| = |(1, 1, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d) - (1, 0, x_{i_3}, \ldots, x_{i_d}) \cdot (z_1, \ldots, z_d)| < \frac{2}{\sqrt{4D^2 - 3}} \leq \frac{2}{4d^2 - 3}, \]

which leads to a contradiction of Eq. (28). We thus complete the proof of Lemma 11.

Now let \( F_2 \) be the event

\[ F_2 = \left\{ \exists X \in \mathcal{X} \cup \{Y\}, \lambda_{\min}\left( \frac{1}{n} \sum_{t=1}^{n} V_{t,X} V_{t,X}^\top \right) \leq \frac{\eta}{2(4D^2 - 3)}, t \geq \frac{cD}{\eta^2 \log nt^2 / \delta} \right\}. \]

Then

\[ \Pr\{F_2\} \leq n \sum_{t \geq (cD/\eta^2) \log(nt^2 / \delta)} e^{-\eta^2 / c} \]

\[ \leq n \sum_{t \geq (cD/\eta^2) \log(nt^2 / \delta)} \frac{\delta}{nt^2} \]

\[ \leq (\frac{\pi^2}{3} - 1) \delta \]

\[ \leq \delta. \]

Now from Lemmas 9, 1 and 10, for all \( a = do(X = 1) \), with probability \( 1 - 2\delta \), for all \( t \geq \max\left\{ \frac{cD}{\eta^2} \log \frac{t^2}{\delta}, \frac{1024(M^{(2)})^2(4D^2 - 3)D}{\kappa^4 \eta} \right\} \), we can deduce that

\[ \lambda_{\min}(M_t, X) \geq \frac{\eta t}{2(4D^2 - 3)}. \]

25
Then
\[|\sigma(\theta_t, X) - \mu_a| \leq \sum_{X' \in P_X} |V_{t,X}(\theta_t - \theta^*)|M^{(1)} \]
\[\leq \frac{3M^{(1)}}{\kappa} \sqrt{\log(3/\delta)} \sum_{X' \in P_X} ||V_{t,X}||_{M_{t,X}^{-1}} \]
\[\leq \frac{3M^{(1)}}{\kappa} \sqrt{\log(3/\delta)} \sum_{X' \in P_X} \sqrt{\lambda_{\min}(M_{t,X})} \]
\[\leq \frac{3M^{(1)}}{\kappa} \sqrt{D(4D^2 - 3) \log(3n/\delta)} \sum_{X' \in P_X} \frac{1}{\sqrt{\eta} t} \]
\[= \frac{6M^{(1)}D^{1.5}}{\kappa \sqrt{\eta} \epsilon} \sqrt{\frac{2L^2}{n} \log(3n/\delta)} \]
\[= \beta_O^a(t).\]

Now we prove that during Algorithm 5, the Phase 1 must terminate at \([T]\) rounds, where \(T = 1152 \frac{m^{(L)}_{\epsilon, \Delta} D^3(M^{(1)})^2}{\kappa \eta \max\{\Delta_1, \epsilon/2\}^2} \log \frac{3n}{\delta} + 16 \frac{m^{(L)}_{\epsilon, \Delta} D^3(M^{(1)})^2}{\kappa \eta \max\{\Delta_1, \epsilon/2\}^2} \log \frac{8nT^3}{\delta}.\) In the following proof, we assume \(F_1\) and \(F_2\) don’t happen.

If \(q_a \geq \frac{\max\{\Delta_1, \epsilon/2\}^2}{m^{(L)}_{\epsilon, \Delta} \max\{\Delta_1, \epsilon/2\}^2}\), we prove at \([T_1]\) rounds, the Phase 1 terminate, where \(T_1 = 1152 \frac{m^{(L)}_{\epsilon, \Delta} D^3(M^{(1)})^2}{\kappa \eta \max\{\Delta_1, \epsilon/2\}^2} \log \frac{3n}{\delta}.\) Assume it doesn’t terminate, for all \(q^{(L)}_a \geq \frac{\max\{\Delta_1, \epsilon/2\}^2}{m^{(L)}_{\epsilon, \Delta} \max\{\Delta_1, \epsilon/2\}^2},\) we have
\[\frac{U_t^a - L_t^a}{2} \leq \frac{6M^{(1)}D^{1.5}}{\kappa \sqrt{\eta} \epsilon} \sqrt{\frac{2L^2}{n} \log(3n/\delta)} \leq \frac{\max\{\Delta_1, \epsilon/2\}}{4}.\]

From Lemma 4, we know Phase 1 must terminate. If \(q^{(L)}_a < \frac{\max\{\Delta_1, \epsilon/2\}^2}{m^{(L)}_{\epsilon, \Delta} \max\{\Delta_1, \epsilon/2\}^2},\) assume \(T_2 = 1152 \frac{m^{(L)}_{\epsilon, \Delta} D^3(M^{(1)})^2}{\kappa \eta \max\{\Delta_1, \epsilon/2\}^2} \log \frac{3n}{\delta} + 16 \frac{m^{(L)}_{\epsilon, \Delta} D^3(M^{(1)})^2}{\kappa \eta \max\{\Delta_1, \epsilon/2\}^2} \log \frac{8nT^3}{\delta}.\) We assume Phase 1 doesn’t terminate at \([T_2]\) rounds, from above, for arm \(a\) such that \(q^{(L)}_a \geq \frac{\max\{\Delta_1, \epsilon/2\}^2}{m^{(L)}_{\epsilon, \Delta} \max\{\Delta_1, \epsilon/2\}^2},\) we have \(\frac{U_t^a - L_t^a}{2} \leq \frac{\max\{\Delta_1, \epsilon/2\}}{4}.\) Then during \([T_1, T_2],\) we have
\[2 \cdot U_{t_a} - L_{t_a} \leq \mu_a + 2(U_t^a - L_t^a) \leq \mu_a + 4\beta_O^a(t) \leq \mu_a - \Delta_a + \max\{\Delta_1, \epsilon/2\} \leq \mu_a + \frac{\epsilon}{2}.\]

If \(U_{t_a}^a - L_{t_a}^a < \epsilon/2\) for some \(t \leq T_2,\) we know
\[U_t^a \leq \mu_a + (U_t^a - L_t^a) \leq \mu_a + \beta_O^a(t) \leq \mu_a + \frac{\epsilon}{4} \leq \mu_a^* - (U_{t_a}^a - L_{t_a}^a) + \frac{3}{4} \epsilon \leq \mu_a^* + \epsilon,\]
which satisfies the stopping rule.

If \(U_{t_a}^a - L_{t_a}^a \geq \epsilon/2\) for any \(t \leq T_2,\)
\[2 \cdot U_{t_a}^a - L_{t_a}^a \geq \mu_a + (U_t^a - L_t^a) \geq \mu_a^* + \frac{\epsilon}{2} \geq 2 \cdot U_t^a - L_t^a. \quad (29)\]

Now for \(t \in [T_1, T_2],\) by (29), the interventions will only act on all arms \(a\) such that \(q^{(L)}_a < \frac{\max\{\Delta_1, \epsilon/2\}^2}{m^{(L)}_{\epsilon, \Delta} \max\{\Delta_1, \epsilon/2\}^2}\). Denoted these arms by \(S.\) By definition of \(m^{(L)}_{\epsilon, \Delta}, |S| \leq m_{\epsilon, \Delta}\) and
\[\sum_{a \in S} D_a ([T_2]) \geq \frac{16(m^{(L)}_{\epsilon, \Delta} D^3(M^{(1)})^2)}{2\kappa \eta \max\{\Delta_1, \epsilon/2\}^2} \log \frac{8nT^3}{\delta} \geq \frac{8m^{(L)}_{\epsilon, \Delta}}{\max\{\Delta_1, \epsilon/2\}^2} \log \frac{8nT^3}{\delta},\]

26
where the last inequality is because $M(1) \geq \kappa$ and $\eta \leq 1$. Similar to the proof of Theorem 1, $D_{a^*}(T_2) \geq \frac{8}{\max\{\Delta_1, \varepsilon/2\}} \log \frac{8nT_2^2}{\delta}$. Thus

$$U_{a^*} - L_{a^*} \leq \beta_1(D_{a^*}(T_2), T_2) \leq \frac{\max\{\Delta_1, \varepsilon/2\}}{4}.$$ 

Then by Lemma 4, Phase 1 should terminate.

Now we prove a sample complexity bound on Phase 1:

**Lemma 12.** If $T = 1152Q \log \frac{3n}{\delta} + 16Q \log \frac{5nT_3^3}{\delta}$, then $T = O(Q \log (Qn/\delta))$. 

**Proof.** Similar to proof in Lemma 6, we only need to show there’s a constant $C$ such that

$$1152Q \log \frac{3n}{\delta} + 16Q \log \frac{8n(CQ \log (Qn/\delta))^3}{\delta} \leq CQ \log \frac{Qn}{\delta}.$$ 

The left-hand side can be bounded by

$$1152Q \log \frac{3n}{\delta} + 48Q \log \frac{n}{\delta} + 48Q \log C + 48Q \log Q + 48Q \log (Qn/\delta)$$

$$\leq (1248 + 48 \log C) \log (Qn/\delta).$$

Thus we can choose $C = 1603$, then we complete the proof because $C > 1248 + 48 \log C$ for $C = 1603$. 

For analysis of Phase 2, the proof is just a simple rewrite. For phase 2, our algorithm has sample complexity at most $T_3 = O(H_{n(\varepsilon/2)}(L_\varepsilon/2, \delta)).$ Hence with probability $1 - 3\delta$ ($F_1, F_2, \text{Phase 2}$) the total sample complexity is

$$T = T_2 + T_3 = O(H^{(L)}_{\varepsilon/2} \log \frac{nH^{(L)}_{\varepsilon/2}}{\delta}),$$

and the algorithm output a $3\varepsilon$-optimal arm (the proof is the same as the proof of Theorem 1). Substitute $\delta$ to $\delta/4$, we complete the proof.