COMBINATORIAL PURE EXPLORATION OF CAUSAL BANDITS

Nuoya Xiong
Institute for Interdisciplinary Information Sciences, Tsinghua University
Beijing, China
xiongny20@mails.tsinghua.edu.cn

Wei Chen
Microsoft Research
Beijing, China
weic@microsoft.com

ABSTRACT

The combinatorial pure exploration of causal bandits is the following online learning task: given a causal graph with unknown causal inference distributions, in each round we choose a subset of variables to intervene 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 δ is a given confidence level. We provide the first gap-dependent and fully adaptive pure exploration algorithms on two types of causal models — the binary generalized linear model (BGLM) and general graphs. For BGLM, our algorithm is the first to be designed specifically for this setting and achieves polynomial sample complexity, while all existing algorithms for general graphs have either sample complexity exponential to the graph size or some unreasonable assumptions. For general graphs, our algorithm provides a significant improvement on sample complexity, and it nearly matches the lower bound we prove. Our algorithms achieve such improvement by a novel integration of prior causal bandit algorithms and prior adaptive pure exploration algorithms, the former of which utilize the rich observational feedback in causal bandits but are not adaptive to reward gaps, while the latter of which have the issue in reverse.

1 INTRODUCTION

Stochastic multi-armed bandits (MAB) is a classical framework in sequential decision making (Robbins, 1952). 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. Pure exploration is an important variant of the multi-armed bandit problem, where the goal is not to accumulate reward but to identify the best arm through possibly adaptive explorations of arms.

Causal bandits, first introduced by Lattimore et al. (2016), integrates causal inference (Pearl, 2009) with multi-armed bandits. In causal bandits, we have a causal graph structure $G = (\boldsymbol{X} \cup \{Y\} \cup \boldsymbol{U}, E)$, where $\boldsymbol{X} \cup \{Y\}$ are observable causal variables with Y being a special reward variable, $\boldsymbol{U}$ are unobserved hidden variables, 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 $\boldsymbol{S} \subseteq \boldsymbol{X}$ together with the choice of null intervention (natural observation), i.e. the action set is $A \subseteq \{a = do(\boldsymbol{S} = \boldsymbol{s}) \mid \boldsymbol{S} \subseteq \boldsymbol{X}, \boldsymbol{s} \in \{0,1\}^{|\boldsymbol{S}|}\}$ with $do() \in A$, where $do(\boldsymbol{S} = \boldsymbol{s})$ is the standard notation for intervening the causal graph by setting $\boldsymbol{S}$ to $\boldsymbol{s}$ (Pearl, 2009), and $do()$ means null intervention. The reward of an action a is the random outcome of Y, and thus the expected reward is $\mathbb{E}[Y \mid a = do(\boldsymbol{S} = \boldsymbol{s})]$. In each round, one action in A is played, and the random outcomes of all variables in $\boldsymbol{X} \cup \{Y\}$ are observed. Given the causal graph G but without knowing the distributions among nodes, the task of combinatorial pure exploration (CPE) of causal bandits is to select actions in each round, observe the feedback from all observable random variables, so that in the end the learner can identify the best or nearly best actions. Causal bandits are useful in many real scenarios. In drug testing, the physicians wants to adjust the dosage of some particular drugs to treat the patient. In policy design, the policy-makers select different actions to reduce the spread of disease.

Existing studies on CPE of causal bandits either requires the knowledge of $P(\boldsymbol{Pa}(Y) \mid a)$ for all action a or only consider causal graphs without hidden variables, and the algorithms proposed are not fully adaptive to reward gaps (Lattimore et al., 2016; Yabe et al., 2018). 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 ε, we aim at designing adaptive algorithms that output an action such that with probability at least $1 - \delta$, the expected reward difference between the output and the optimal action is at most ε. The algorithms should be fully adaptive in the follow two senses. First, it should adapt to the reward gaps between suboptimal and optimal actions similar to existing adaptive pure exploration bandit algorithms, such that actions with larger gaps should be explored less. Second, it should adapt to the observational data from causal bandit feedback, such that actions 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 interaction between the two aspects, we achieve better adaptivity than either of them alone.

First we introduce a particular term named gap-dependent observation threshold, which is a non-trivial gap-dependent extension for a similar term in Lattimore et al. (2016). Then we provide two algorithms, one for the binary generalized linear model (BGLM) and one for the general model with hidden variables. The sample complexity of both algorithms contains the gap-dependent observation threshold that we introduced, which shows significant improvement comparing to the prior work. In particular, our algorithm for BGLM achieves a sample complexity polynomial to the graph size, while all prior algorithms for general graphs have exponential sample complexity; and our algorithm for general graphs match a lower bound we prove in the paper. To our best knowledge, our paper is the first work considering a CPE algorithm specifically designed for BGLM, and the first work considering CPE on graphs with hidden variables, while all prior studies either assume no hidden variables or assume knowing distribution $P(\boldsymbol{Pa}(Y) \mid a)$ for the parents of reward variable $\boldsymbol{Pa}(Y)$ and all action a, which is not a reasonable assumption.

To summarize, our contribution is to propose the first set of CPE algorithms on causal graphs with hidden variables and fully adaptive to both the reward gaps and the observational causal data. The algorithm on BGLM is the first to achieve a gap-dependent sample complexity polynomial to the graph size, while the algorithm for general graphs improves significantly on sample complexity and matches a lower bound. Due to the space constraint, further materials including experimental results, an algorithm for the fixed-budget setting, and all proofs are moved to the appendix.

Related Work. Causal bandit is proposed by Lattimore et al. (2016), who discuss the simple regret for parallel graphs and general graphs with known probability distributions $P(\boldsymbol{Pa}(Y) \mid a)$. Nair et al. (2021) extend algorithms on parallel graphs to graphs without back-door paths, and Maiti et al. (2021) extend the results to the general graphs. All of them either regard $P(\boldsymbol{Pa}(Y) \mid a)$ as prior knowledge, or consider only atomic intervention. The study by Yabe et al. (2018) is the only one considering the general graphs with combinatorial action set, but their algorithm cannot work on causal graphs with hidden variables. All the above pure exploration studies consider simple regret criteria that is not gap-dependent. Cumulative regret is considered in (Lu et al., 2020; Nair et al., 2021; Maiti et al., 2021). To our best knowledge, Sen et al. (2017) 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 $P(X|\boldsymbol{Pa}(X))$) on *one single node*, which is different from causal bandits defined in Lattimore et al. (2016).

Pure exploration of multi-armed bandit has been extensively studied in the fixed-confidence or fixed-budget setting (Audibert et al., 2010; Kalyanakrishnan et al., 2012; Jamieson et al., 2013; Jamieson & Nowak, 2014). PAC pure exploration is a generalized setting aiming to find the ε-optimal arm instead of exactly optimal arm (Even-Dar et al., 2002; Mannor & Tsitsiklis, 2004). In this paper, we utilize the adaptive LUCB algorithm in (Kalyanakrishnan et al., 2012). CPE has also been studied for multi-armed bandits and linear bandits, etc.(Karnin et al. (2013b); Chen et al. (2014); Du et al. (2021)), but the feedback model in these studies either have feedback at the base arm level or have full or partial bandit feedback, which are all different from the causal bandit feedback.

The binary generalized linear model (BGLM) is studied in (Li et al., 2017; Feng & Chen, 2022) for cumulative regret MAB problems. We borrow the maximum likelihood estimation method and its result in (Li et al., 2017; Feng & Chen, 2022) for our BGLM part, but its integration with our adaptive sampling algorithm for the pure exploration setting is new.

2 Model and Preliminaries

Causal Models. From Pearl (2009), a causal graph $G = (\boldsymbol{X} \cup \{Y\} \cup \boldsymbol{U}, E)$ is a directed acyclic graph (DAG) with a set of observed random variables $\boldsymbol{X} \cup \{Y\}$ and a set of hidden random variables $\boldsymbol{U}$, where $\boldsymbol{X} = \{X_1, \cdots, X_n\}$, $\boldsymbol{U} = \{U_1, \cdots, U_k\}$ are two set of variables and Y is the special reward variable without outgoing edges. In this paper, for simplicity, we only consider that X_i's and Y are binary random variables with support $\{0, 1\}$. For any node V in G, we denote the set of its parents in G as $\boldsymbol{Pa}(V)$. The set of values for $\boldsymbol{Pa}(X)$ is denoted by $\boldsymbol{pa}(X)$. The causal influence is represented by $P(V \mid \boldsymbol{Pa}(V))$, modeling the fact that the probability distribution of a node V's value is determined by the value of its parents. Henceforth, when we refer to a causal graph, we mean both its graph structure $(\boldsymbol{X} \cup \{Y\} \cup \boldsymbol{U}, E)$ and its causal inference distributions $P(V \mid \boldsymbol{Pa}(V))$ for all $V \in \boldsymbol{X} \cup \{Y\} \cup \boldsymbol{U}$. A parallel graph $G = (\boldsymbol{X} \cup \{Y\}, E)$ is a special class of causal graphs with $\boldsymbol{X} = \{X_1, \cdots, X_n\}$, $\boldsymbol{U} = \emptyset$ and $E = \{X_1 \to Y, X_2 \to Y, \cdots, X_n \to Y\}$. An *intervention* $do(\boldsymbol{S} = \boldsymbol{s})$ in the causal graph G means that we set the values of a set of nodes $\boldsymbol{S} \subseteq \boldsymbol{X}$ to $\boldsymbol{s}$, while other nodes still follow the $P(V \mid \boldsymbol{Pa}(V))$ distributions. An *atomic intervention* means that $|\boldsymbol{S}| = 1$. When $\boldsymbol{S} = \emptyset$, $do(\boldsymbol{S} = \boldsymbol{s})$ is the null intervention denoted as $do()$, which means we do not set any node to any value and just observe all nodes' values.

In this paper, we also study a parameterized model with no hidden variables: *the binary generalized linear model* (BGLM). Specifically, in BGLM, we have $\boldsymbol{U} = \emptyset$, and $P(X = 1 \mid \boldsymbol{Pa}(X) = \boldsymbol{pa}(X)) = f_X(\boldsymbol{\theta}_X \cdot \boldsymbol{pa}(X)) + e_X$, where f_X is a strictly increasing function, $\boldsymbol{\theta}_X \in \mathbb{R}^{\boldsymbol{Pa}(X)}$ is the unknown parameter vector for X, e_X is a zero-mean bounded noise variable that guarantees the resulting probability to be within $[0, 1]$. To represent the intrinsic randomness of node X not caused by its parents, we denote $X_1 = 1$ as a global variable, which is a parent of all nodes.

Combinatorial Pure Exploration of Causal Bandits. Combinatorial pure exploration (CPE) of causal bandits describes the following setting and the online learning task. The causal graph structure is known but the distributions $P(V|\boldsymbol{Pa}(V))$'s are unknown. The action (arm) space $\boldsymbol{A}$ is a subset of possible interventions on combinatorial sets of variables, plus the null intervention, that is, $\boldsymbol{A} \subseteq \{do(\boldsymbol{S} = \boldsymbol{s}) \mid \boldsymbol{S} \subseteq \boldsymbol{X}, \boldsymbol{s} \in \{0, 1\}^{|\boldsymbol{S}|}\}$ and $\{do()\} \in \boldsymbol{A}$. For action $a = do(\boldsymbol{S} = \boldsymbol{s})$, define $\mu_a = \mathbb{E}[Y \mid do(\boldsymbol{S} = \boldsymbol{s})]$ to be the expected reward of action $do(\boldsymbol{S} = \boldsymbol{s})$. Let $\mu^* = \max_{a \in \boldsymbol{A}} \mu_a$.

In each round t, the learning agent plays one action $a \in \boldsymbol{A}$, observes the sample values $\boldsymbol{X}_t = (X_{t,1}, X_{t,2} \cdots, X_{t,n})$ and Y_t of all observed variables. The goal of the agent is to interact with the causal model with as small number of rounds as possible to find an action with the maximum expected reward μ^*. More precisely, we mainly focus on the following PAC pure exploration with the gap-dependent bound 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^o to guarantee that $\mu^* - \mu_{a^o} \le \varepsilon$ 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^o. Note that when $\varepsilon = 0$, the PAC setting is reduced to the classical pure exploration setting. We also consider the *fixed budget setting* in the appendix, where given an exploration round budget T and an error parameter $\varepsilon \in [0, 1)$, the agent is trying to adaptively play actions and output an action a^o at the end of round T, so that the error probability $\Pr\{\mu_{a^o} < \mu^* - \varepsilon\}$ is as small as possible.

We study the gap-dependent bounds, meaning that 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

$$\Delta_a = \begin{cases} \mu_{a^*} - \max_{a \in \boldsymbol{A} \setminus \{a^*\}}\{\mu_a\}, & a = a^*; \\ \mu_{a^*} - \mu_a, & a \neq a^*. \end{cases} \tag{1}$$

We further sort the gaps Δ_a's for all arms and assume $\Delta^{(1)} \le \Delta^{(2)} \cdots \le \Delta^{(|\boldsymbol{A}|)}$, where $\Delta^{(1)}$ is also denoted as $\Delta_{\min}$.

3 GAP-DEPENDENT OBSERVATION THRESHOLD

In this section, we introduce the key concept of gap-dependent observation threshold, which is instrumental to the fix-confidence algorithms in the next two sections. Intuitively, it describes for any action a whether we can derive its reward from pure observations of the causal model.

We assume that X_i's are binary random variables. First, we describe terms $q_a \in [0,1]$ for each action $a \in \boldsymbol{A}$, which can have different definitions in different settings. Intuitively, q_a represents how easily the action a is to be estimated by observation. For example, in Lattimore et al. (2016), for parallel graph with action set $\boldsymbol{A} = \{do(X_i = x) \mid 1 \le i \le n, x \in \{0,1\}\} \cup \{do()\}$, for action $a = do(X_i = x)$, $q_a = P(X_i = x)$ represents the probability for action $do(X_i = x)$ to be observed, since in parallel graph we have $P(Y \mid X_i = x) = P(Y \mid do(X_i = x))$. Thus, when $q_a = P(X_i = x)$ is larger, it is easier to estimate $P(Y \mid do(X_i = x))$ by observation. We will instantiate q_a's for BGLM and general graphs in later sections. For $a = do()$, we always set $q_a = 1$. Then, for set $q_a, a \in \boldsymbol{A}$ we define the *observation thershold* as follows:

Definition 1 (Observation threshold Lattimore et al. (2016))**.** *For a given causal graph G and its associated $\{q_a \mid a \in \boldsymbol{A}\}$, the* observation threshold m *is defined as:*

$$m = \min\{\tau \in [|\boldsymbol{A}|] : |\{a \in \boldsymbol{A} \mid q_a < 1/\tau\}| \le \tau\}. \tag{2}$$

The observation threshold can be equivalently defined as follows: When we sort $\{q_a \mid a \in \boldsymbol{A}\}$ as $q^{(1)} \le q^{(2)} \le \cdots \le q^{|\boldsymbol{A}|}$, $m = \min\{\tau : q^{(\tau+1)} \ge \frac{1}{\tau}\}$. Note that $m \le |\boldsymbol{A}|$ always holds since $q_{do()} = 1$. In some cases, $m \ll |\boldsymbol{A}|$. For example, in parallel graph, when $P(X_i = 1) = P(X_i = 0) = \frac{1}{2}$ for all $i \in [n]$, $q_{do(X_i=1)} = q_{do(X_i=0)} = \frac{1}{2}$, $q_{do()} = 1$.Then $m = 2 \ll 2n+1 = |\boldsymbol{A}|$. Intuitively, when we collect passive observation data without intervention, arms corresponding to $q^{(j)}$ with $j \le m$ are under observed while arms corresponding to $q^{(j)}$ with $j > m$ are sufficiently observed and can be estimated accurately. Thus, for convenience we name m as the observation threshold (the term is not given a name in Lattimore et al. (2016)).

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. Before introducing the definition, we first define the term H_r. Sort the arm set as $q_{a_1} \cdot \max\{\Delta_{a_1}, \varepsilon/2\}^2 \le q_{a_2} \cdot \max\{\Delta_{a_2}, \varepsilon/2\}^2 \le \cdot \le q_{a_{|\boldsymbol{A}|}} \cdot \max\{\Delta_{a_{|\boldsymbol{A}|}}, \varepsilon/2\}^2$, then H_r is defined by

$$H_r = \sum_{i=1}^{r} \frac{1}{\max\{\Delta_{a_i}, \varepsilon/2\}^2}. \tag{3}$$

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

$$m_{\varepsilon,\Delta} = \min\left\{\tau \in [|\boldsymbol{A}|] : \left|\left\{a \in \boldsymbol{A} \middle| q_a \cdot \max\left\{\Delta_a, \varepsilon/2\right\}^2 < \frac{1}{H_\tau}\right\}\right| \le \tau\right\}. \tag{4}$$

The Gap-dependent observation threshold can be regarded as a generalization of the observation threshold. Intuitively, when considering the gaps, $q_a \cdot \max\{\Delta_a, \varepsilon/2\}^2$ represents how easily the action a would to be distinguished from the optimal arm. To show the relationship between $m_{\varepsilon,\Delta}$ and m, we provide the following lemma. The proof of Lemma is in Appendix D.1.

Lemma 1. *$m_{\varepsilon,\Delta} \le 2m$.*

Lemma 1 shows that $m_{\varepsilon,\Delta} = O(m)$. In many real scenarios, $m_{\varepsilon,\Delta}$ might be much smaller than m. Consider some integer n with $4 < n < |\boldsymbol{A}|$, $\epsilon < 1/n$, $q_a = \frac{1}{n}$ for $a \in \boldsymbol{A} \setminus \{do()\}$ and $q_{do()} = 1$. Then $m = n$. Now we consider $\Delta_{a_1} = \Delta_{a_2} = \frac{1}{n}$, while other arms a have $\Delta_a = \frac{1}{2}$. Then $H_r \ge n^2$ for all $r \ge 1$. Then for $a \ne a_1, a_2$, we have $q_a \cdot \max\{\Delta_a, \varepsilon/2\}^2 \ge \frac{1}{4n} > \frac{1}{H_r}$, which implies that $m_{\varepsilon,\Delta} = 2$. This lemma will be used to show that our result improves previous causal bandit algorithm in Lattimore et al. (2016).

4 COMBINATORIAL PURE EXPLORATION FOR BGLM

In this section, we discuss the pure exploration for BGLM, a general class of causal graphs with a linear number of parameters, as defined in Section 2. In this section, we assume $\boldsymbol{U} = \emptyset$. Let

$\boldsymbol{\theta}^* = (\boldsymbol{\theta}^*_X)_{X\in\boldsymbol{X}\cup\{Y\}}$ be the vector of all weights. Since X_1 is a global variable, we only need to consider the action set $\boldsymbol{A} \subseteq \{do(\boldsymbol{S}=\boldsymbol{s}) \mid \boldsymbol{S} \subseteq \boldsymbol{X}\setminus\{X_1\}, \boldsymbol{s}\in\{0,1\}^{|\boldsymbol{S}|}\}$. Following Li et al. (2017); Feng & Chen (2022), we have three assumptions:

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

Assumption 2. $\kappa := \inf_{X\in\boldsymbol{X}\cup\{Y\},\boldsymbol{v}\in[0,1]^{Pa(X)},||\boldsymbol{\theta}-\boldsymbol{\theta}^*_X||\le 1} \dot{f}_X(\boldsymbol{v}\cdot\boldsymbol{\theta}) > 0$ *is a positive constant.*

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

$$Pr[X' = x \mid \boldsymbol{Pa}(X)\setminus\{X', X_1\} = \boldsymbol{v}] \ge \eta. \tag{5}$$

Assumptions 1 and 2 are the classical assumptions in generalized linear model Li et al. (2017). 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, and it is originally given in Feng & Chen (2022) with additional justifications. Henceforth, we use $\sigma(\boldsymbol{\theta}, a)$ to denote the reward μ_a on parameter $\boldsymbol{\theta}$.

Our main algorithm, Causal Combinatorial Pure Exploration-BGLM (CCPE-BGLM), is given in Algorithm 1. The algorithm follows the LUCB framework Kalyanakrishnan et al. (2012), but has several innovations. In each round t, we play three actions and thus it corresponds to three rounds in the general CPE model. In each round t, we maintain $\hat{\mu}^t_{O,a}$ and $\hat{\mu}^t_{I,a}$ as the estimates of μ_a from the observational data and the interventional data, respectively. For each estimate, we maintain its confidence interval, $[L^t_{O,a}, U^t_{O,a}]$ and $[L^t_{I,a}, U^t_{I,a}]$ respectively.

At the beginning of round t, similar to LUCB, we find two candidate actions, one with the highest empirical mean so far, a^{t-1}_h; and one with the highest UCB among the rest, a^{t-1}_l. If the LCB of a^{t-1}_h is higher than the UCB of a^{t-1}_l with an ε error, then the algorithm could stop and return a^{t-1}_h as the best action. However, the second stopping condition in line 5 is new, and it is used to guarantee that the observational estimates $\hat{\mu}^t_{O,a}$'s are from enough samples. If the stopping condition is not met, we will do three steps. The first step is the novel observation step comparing to LUCB. In this step, we do the null intervention $do()$, collect observational data, use maximum-likelihood estimate adapted from Li et al. (2017); Feng & Chen (2022) to obtain parameter estimate $\hat{\theta}_t$, and then use $\hat{\theta}_t$ to compute observational estimate $\hat{\mu}^t_{O,a} = \sigma(\hat{\theta}_t, a)$ for all action a, where $\sigma(\hat{\theta}_t, a)$ means the reward for action a on parameter $\hat{\theta}_t$. This can be efficiently done by following the topological order of nodes in G and parameter $\hat{\theta}_t$. From $\hat{\mu}^t_{O,a}$, we obtain the confidence interval $[L^t_{O,a}, U^t_{O,a}]$ using the bonus term defined later in Eq.(8). In the second step, we play the two candidate actions a^{t-1}_h and a^{t-1}_l and update their interventional estimates and confidence intervals, as in LUCB. In the third step, we merge the two estimates together and set the final estimate $\hat{\mu}^t_a$ to be the mid point of the intersection of two confidence intervals. While the second step follows the LUCB, the first and the third step are new, and they are crucial for utilizing the observational data to obtain quick estimates for many actions at once.

Utilizing observational data has been explored in past causal bandit studies, but they separate the exploration from observations and the interventions into two stages (Lattimore et al., 2016; Nair et al., 2021), and thus their algorithms are not adaptive and cannot provide gap-dependent sample complexity bounds. Our algorithm innovation is in that we interleave the observation step and the intervention step naturally into the adaptive LUCB framework, so that we can achieve an adaptive balance between observation and intervention, achieving the best of both worlds.

To get the confidence bound for BGLM, we use the following lemma from Feng & Chen (2022):

Lemma 2. *For an action $a = do(\boldsymbol{S}=\boldsymbol{s})$ and any two weight vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we have*

$$|\sigma(\boldsymbol{\theta}, a) - \sigma(\boldsymbol{\theta}', a)| \le \mathbb{E}_{\boldsymbol{e}}\left[\sum_{X\in N_{\boldsymbol{S},Y}} |\boldsymbol{V}_X^\top(\boldsymbol{\theta}_X - \boldsymbol{\theta}'_X)|M^{(1)}\right], \tag{6}$$

where $N_{\boldsymbol{S},Y}$ is the set of all nodes that lie on all possible paths from X_1 to Y excluding $\boldsymbol{S}$, $\boldsymbol{V}_X$ is the value vector of a sample of the parents of X according to parameter $\boldsymbol{\theta}$, $M^{(1)}$ is defined in Assumption 1, and the expectation is taken on the randomness of the noise term $\boldsymbol{e} = (e_X)_{X\in\boldsymbol{X}\cup\{Y\}}$ of causal model under parameter $\boldsymbol{\theta}$.

Algorithm 1 CCPE-BGLM$(G, \boldsymbol{A}, \varepsilon, \delta, M^{(1)}, M^{(2)}, \kappa, \eta, c)$

Input:causal general graph G, action set $\boldsymbol{A}$, parameter $\varepsilon, \delta, M^{(1)}, M^{(2)}, \kappa, \eta, c$ in Assumptions 1,2, 3 and Lemma 4 .
Initialize $M_{0,X} = I$ for all node X. $N_a = 0, \hat{\mu}_a^0 = 0, L_a^0 = -\infty, U_a^0 = \infty$ for arms $a \in \boldsymbol{A}$.
for $t = 1, 2, \cdots,$ **do**
  $a_h^{t-1} = \text{argmax}_{a\in\boldsymbol{A}} \hat{\mu}_a^{t-1}, a_l^{t-1} = \text{argmax}_{a\in\boldsymbol{A}\setminus\{a_h^{t-1}\}} U_a^{t-1}$.
  if $U_{a_l^{t-1}}^{t-1} \le L_{a_h^{t-1}}^{t-1} + \varepsilon$ and $t \ge \max\{\frac{cD}{\eta^2} \log \frac{nt^2}{\delta}, \frac{1024(M^{(2)})^2(4D^2-3)D}{\kappa^4\eta}(D^2 + \log \frac{3nt^2}{\delta})\}$ **then**
    return a_h^{t-1}.
  end if
  /* *Step 1. Conduct a passive observation and estimate from the observational data* */
  Perform action $do()$ and observe $\boldsymbol{X}_t$ and Y_t. For $a = do()$, $N_a = N_a + 1$.
  $\hat{\theta}_t = \text{BGLM-estimate}((\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_t, Y_t))$.
  For $a = do(\boldsymbol{S} = \boldsymbol{s}) \in \boldsymbol{A}$, calculate $\hat{\mu}_{O,a} = \sigma(\hat{\theta}_t, \boldsymbol{S})$, and $[L_{O,a}^t, U_{O,a}^t] = [\hat{\mu}_{O,a} - \beta_O^a(t), \hat{\mu}_{O,a} + \beta_O^a(t)]$. /* $\beta_O^a(t)$ is defined in Eq.(8) */
  /* *Step 2. Do two interventions and estimate from the interventional data* */
  Perform actions a_l^{t-1} and a_h^{t-1}, get the reward $Y_t^{(l)}$ and $Y_t^{(h)}$.
  $N_{a_l^{t-1}} = N_{a_l^{t-1}} + 1, N_{a_h^{t-1}} = N_{a_h^{t-1}} + 1$.
  For $a \in \{a_l^{t-1}, a_h^{t-1}, do()\}$, update the empirical mean
  $\hat{\mu}_{I,a} = \sum_{j=1}^t \frac{1}{N_a}(\mathbb{I}\{a = a_l^{j-1}\}Y_j^{(l)} + \mathbb{I}\{a = a_h^{j-1}\}Y_j^{(h)} + \mathbb{I}\{a = do()\}Y_j)$ and $[L_{I,a}^t, U_{I,a}^t] = [\hat{\mu}_{I,a} - \beta_I(N_a), \hat{\mu}_{I,a} + \beta_I(N_a)]$. /* $\beta_I(t)$ is defined in Eq.(8) */
  /* *Step 3. Merge the observational estimate and the interventional estimate* */
  For $a \in \boldsymbol{A}$, calculate $[L_a^t, U_a^t] = [L_{O,a}^t, U_{O,a}^t] \cap [L_{I,a}^t, U_{I,a}^t]$ and $\hat{\mu}_a^t = \frac{L_a^t + U_a^t}{2}$.
end for

The key idea in the design and analysis of the algorithm is to divide the actions into two sets — the easy actions and the hard actions. Intuitively, the easy actions are the ones that can be easily estimated by observational data, while the hard actions require direction playing these actions to get accurate estimates. The quantity q_a mentioned in Section 3 indicates how easy is action a, and it determines the gap-dependent observational threshold $m_{\varepsilon,\Delta}$ (Definition 2), which essentially gives the number of hard actions. In fact, the set of actions in Eq.(4) with $\tau = m_{\varepsilon,\Delta}$ is the set of hard actions and the rest are easy actions. We need to define q_a representing the hardness of estimation for each a.

Algorithm 2 BGLM-estimate

Input: data pairs $((\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \cdots, (\boldsymbol{X}_t, Y_t))$
Construct $(\boldsymbol{V}_{t,X}, X_t)$ for each X, where $\boldsymbol{V}_{t,X}$ is the value of parent of X at round t, X_t is the value of X at round t.
for $X \in \boldsymbol{X} \cup \{Y\}$ **do**
  $M_{t,X} = M_{t-1,X} + \boldsymbol{V}_{t,X}\boldsymbol{V}_{t,X}^\top$, calculate $\hat{\boldsymbol{\theta}}_{t,X}$ by solving $\sum_{i=1}^t (X_i - f_X(\boldsymbol{V}_{i,X}^T \hat{\boldsymbol{\theta}}_{t,X}))\boldsymbol{V}_{i,X} = 0$.
end for
return $\hat{\boldsymbol{\theta}}_t$.

For CCPE-BGLM, we define its $q_a^{(L)}$ as follows. Let $D = \max_{X\in\boldsymbol{X}\cup\{Y\}} |\boldsymbol{Pa}(X)|$. For node $\boldsymbol{S} \subseteq \boldsymbol{X}$, let $\ell_{\boldsymbol{S}} = |N_{\boldsymbol{S},Y}|$. Then for $a = do(\boldsymbol{S} = \boldsymbol{s})$, we define

$$q_a^{(L)} = \frac{1}{\ell_{\boldsymbol{S}}^2 D^3}. \tag{7}$$

Intuitively, based on Lemma 2 and $\ell_{\boldsymbol{S}} = |N_{\boldsymbol{S},Y}|$, a large $\ell_{\boldsymbol{S}}$ means that the right-hand side of Inequality (6) could be large, and thus it is difficult to estimate μ_a accurately. Hence the term $q_a^{(L)}$ represents how easy it is to estimate for action a. Note that $q_a^{(L)}$ only depends on the graph structure and set $\boldsymbol{S}$. We can define $m^{(L)}$ and $m_{\varepsilon,\Delta}^{(L)}$ with respect to $q_a^{(L)}$'s by Definitions 1 and 2. We use two confidence radius terms as follows, one from the estimate of the observational data, and the other from the estimate of the interventional data.

$$\beta_O^a(t) = \frac{\alpha_O M^{(1)} D^{1.5}}{\kappa\sqrt{\eta}} \sqrt{\frac{1}{q_a^{(L)} t} \log \frac{3nt^2}{\delta}}, \beta_I(t) = \alpha_I \sqrt{\frac{1}{t} \log \frac{|\boldsymbol{A}| \log(2t)}{\delta}}. \tag{8}$$

Parameters α_O and α_I are exploration parameters for our algorithm. For a theoretical guarantee, we choose $\alpha_O = 6\sqrt{2}$ and $\alpha_I = 2$, but more aggressive α_O and α_I could be used in experiments. (e.g. Mason et al. (2020), Kaufmann et al. (2016a), Jamieson et al. (2013)) The sample complexity of CCPE-BGLM is summarized in the following theorem.

Theorem 1. *With probability $1-\delta$, our* CCPE-BGLM$(G, \boldsymbol{A}, \varepsilon, \delta/2)$ *returns an ε-optimal arm with sample complexity*

$$T = O\left(H_{m^{(L)}_{\varepsilon,\Delta}} \log \frac{|\boldsymbol{A}| H_{m^{(L)}_{\varepsilon,\Delta}}}{\delta}\right), \tag{9}$$

where $m^{(L)}_{\varepsilon,\Delta}, H_{m^{(L)}_{\varepsilon,\Delta}}$ are defined in Definition 2 and Eq.(3) in terms of $q_a^{(L)}$'s for $a \in \boldsymbol{A} \setminus \{do()\}$ defined in Eq.(7).

If we treat the problem as a naive $|\boldsymbol{A}|$-arms bandit, the sample complexity of LUCB1 is $\widetilde{O}(H) = \widetilde{O}(\sum_{a\in\boldsymbol{A}} \frac{1}{\max\{\Delta_a, \varepsilon/2\}^2})$, which may contain an exponential number of terms. Now note that $q_a^{(L)} \geq \frac{1}{n^5}$, it is easy to show that $m^{(L)}_{\varepsilon,\Delta} \leq 2n^5$. Hence $H_{m^{(L)}_{\varepsilon,\Delta}}$ contains only a polynomial number of terms. Other causal bandit algorithms also suffer an exponential term, unless they rely on a strong and unreasonable assumption as describe in the related work. We achieve an exponential speedup by (a) a specifically designed algorithm for the BGLM model, and (b) interleaving observation and intervention and making the algorithm fully adaptive.

The idea of the analysis is as follows. First, for the $m_{\varepsilon,\Delta}$ hard actions, we rely on the adaptive LUCB to identify the best, and its sample complexity according to LUCB is $O(H_{m^{(L)}_{\varepsilon,\Delta}} \log(|\boldsymbol{A}| H_{m^{(L)}_{\varepsilon,\Delta}}/\delta))$. Next, for easy actions, we rely on the observational data to provide accurate estimates. According to Eq.(4), every easy action a has the property that $q_a \cdot \max\{\Delta_a, \varepsilon/2\}^2 \geq 1/H_{m_{\varepsilon,\Delta}}$. Using this property together with Lemma 2, we would show that the sample complexity for estimating easy action rewards is also $O(H_{m^{(L)}_{\varepsilon,\Delta}} \log(|\boldsymbol{A}| H_{m^{(L)}_{\varepsilon,\Delta}}/\delta))$. Finally, the interleaving of observations and interventions keep the samply complexity in the same order.

5 Combinatorial Pure Exploration for General Graphs

5.1 CPE Algorithm for General Graphs

In this section, we apply a similar idea to the general graph setting, which further allows the existence of hidden variables. The first issue is how to estimate the causal effect (or the do effect) $\mathbb{E}[Y \mid do(\boldsymbol{S} = \boldsymbol{s})]$ in general causal graphs from the observational data. The general concept of identifiability (Pearl, 2009) is difficult for sample complexity analysis. Here we use the concept of *admissible sequence* (Pearl, 2009) to achieve this estimation.

Definition 3 (Admissible sequence)**.** *An admissible sequence for general graph G with respect to Y and $\boldsymbol{S} = \{X_1, \cdots, X_k\} \subseteq \boldsymbol{X}$ is a sequence of sets of variables $\boldsymbol{Z}_1, \cdots \boldsymbol{Z}_k \subseteq \boldsymbol{X}$ such that*

(1) $\boldsymbol{Z}_i$ consists of nondescendants of $\{X_i, X_{i+1}, \cdots, X_k\}$,

(2) $(Y \perp\!\!\!\perp X_i \mid X_1, \cdots, X_{i-1}, \boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_i)_{G_{\underline{X_i}, \overline{X}_{i+1}, \cdots, \overline{X}_k}}$, where $G_{\underline{X}}$ means graph G without out-edges of X, and $G_{\overline{X}}$ means graph G without in-edges of X.

Then, for $\boldsymbol{S} = \{X_1, \cdots, X_k\}$, $\boldsymbol{s} = \{x_1, \cdots, x_k\}$, we can calculate $\mathbb{E}[Y \mid do(\boldsymbol{S} = \boldsymbol{s})]$ by

$$\mathbb{E}[Y \mid do(\boldsymbol{S} = \boldsymbol{s})] = \sum_{\boldsymbol{z}} P(Y = 1 \mid \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{Z}_i = \boldsymbol{z}_i, i \leq k)$$
$$\cdot P(\boldsymbol{Z}_1 = \boldsymbol{z}_1) \cdots P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1), \tag{10}$$

where $\boldsymbol{z}$ means the value of $\cup_{i=1}^k \boldsymbol{Z}_i$, and $\boldsymbol{z}_i$ means the projection of $\boldsymbol{z}$ on $\boldsymbol{Z}_i$. For $a = do(\boldsymbol{S} = \boldsymbol{s})$ with $|\boldsymbol{S}| = k$, we use $\{\boldsymbol{Z}_{a,i}\}_{i=1}^k$ to denote the admissible sequence with respect to Y and $\boldsymbol{S}$, and $\boldsymbol{Z}_a = \cup_{i=1}^k \boldsymbol{Z}_{a,i}$. $Z_a = |\boldsymbol{Z}_a|$ and $Z = \max_a Z_a$. In this paper, we simplify $\boldsymbol{Z}_{a,i}$ to $\boldsymbol{Z}_i$ if there is no ambiguity.

For any $\boldsymbol{P} \subseteq \boldsymbol{X}$, denote $\boldsymbol{P}_t = \boldsymbol{X}_t|_{\boldsymbol{P}}$ as the projection of $\boldsymbol{X}_t$ on $\boldsymbol{P}$. We define

Algorithm 3 CCPE-General$(G, \boldsymbol{A}, \varepsilon, \delta)$

Input:causal graph G, action set $\boldsymbol{A}$, parameter ε, δ, admissible sequence $\{(Z_a)_i\}$ for each action $a \in \boldsymbol{A}$
Initialize $t = 1, T_a = 0, T_{a,\boldsymbol{z}} = 0, N_a = 0, \hat{\mu}_a = 0$ for all arms $a \in \boldsymbol{A}$, $\boldsymbol{z} \in \{0,1\}^z$, $z \in [|X|]$.
for $t = 1, 2, \cdots,$ **do**
  $a_h^{t-1} = \text{argmax}_{a \in \boldsymbol{A}} \hat{\mu}_a^{t-1}, a_l^{t-1} = \text{argmax}_{a \in \boldsymbol{A} \setminus a_h^{t-1}} (U_a^{t-1})$
  if $U_{a_l^{t-1}} \le L_{a_h^{t-1}} + \varepsilon$ **then**
    return a_h^{t-1}
  end if
  / Step 1. Conduct a passive observation and estimate from the observational data */*
  Perform $do()$ operation and observe $\boldsymbol{X}_t$ and Y_t. For $a = do()$, $N_a = N_a + 1$.
  for $a = do(\boldsymbol{S} = \boldsymbol{s}) \in \boldsymbol{A} \setminus \{do()\}$ with an admissible sequence and $\boldsymbol{S} = \{X_1, \cdots, X_k\}, \boldsymbol{s} = \{x_1, \cdots, x_k\}$ **do**
    Estimate $\hat{\mu}_{O,a}$ using (14) and $[L_{O,a}^t, U_{O,a}^t] = [\hat{\mu}_{O,a} - \beta_O^a(T_a, t), \hat{\mu}_{O,a} + \beta_O^a(T_a)]$. */* $\beta_O^a(t)$ is defined in Eq.(16), $T_{a,\boldsymbol{z}}$ is defined in Eq.(11) and $T_a = \min_{\boldsymbol{z}} T_{a,\boldsymbol{z}}$. */*
  end for
  / Step 2. Do two interventions and estimate from the interventional data */*
  Perform actions a_l^{t-1} and a_h^{t-1}, get the reward $Y_t^{(l)}$ and $Y_t^{(h)}$.
  $N_{a_l^{t-1}} = N_{a_l^{t-1}} + 1, N_{a_h^{t-1}} = N_{a_h^{t-1}} + 1$.
  For $a \in \{a_l^{t-1}, a_h^{t-1}, do()\}$, update the empirical mean $\hat{\mu}_{I,a}$ as Line 16 in Algorithm 1.
  Update $[L_{I,a}^t, U_{I,a}^t] = [\hat{\mu}_{I,a} - \beta_I(N_a), \hat{\mu}_{I,a} + \beta_I(N_a)]$. */* $\beta_I(t)$ is defined in Eq.(16) */*
  / Step 3. Merge the observational estimate and the interventional estimate */*
  For $a \in \boldsymbol{A}$, calculate $[L_a^t, U_a^t] = [L_{O,a}^t, U_{O,a}^t] \cap [L_{I,a}^t, U_{I,a}^t]$ and $\hat{\mu}_a^t = \frac{L_a^t + U_a^t}{2}$.
end for

$$T_{a,\boldsymbol{z}} = \sum_{j=1}^{t} \mathbb{I}\{\boldsymbol{S}_j = \boldsymbol{s}, (\boldsymbol{Z}_a)_j = \boldsymbol{z}\}, r_{a,\boldsymbol{z}}(t) = \frac{1}{T_{a,\boldsymbol{z}}} \sum_{j=1}^{t} \mathbb{I}\{\boldsymbol{S}_j = \boldsymbol{s}, (\boldsymbol{Z}_a)_j = \boldsymbol{z}\} Y_j \tag{11}$$

$$n_{a,\boldsymbol{z},l}(t) = \sum_{j=1}^{t} \mathbb{I}\{(\boldsymbol{Z}_i)_j = \boldsymbol{z}_i, (X_i)_j = x_i, i \le l-1\} \tag{12}$$

$$p_{a,\boldsymbol{z},l}(t) = \frac{1}{n_{a,\boldsymbol{z},l}(t)} \sum_{j=1}^{t} \mathbb{I}\{(\boldsymbol{Z}_l)_j = \boldsymbol{z}_l, (\boldsymbol{Z}_i)_j = \boldsymbol{z}_i, (X_i)_j = x_i, i \le l-1\} \tag{13}$$

where the $r_{a,\boldsymbol{z}}(t)$ and $p_{a,\boldsymbol{z},l}(t)$ are the empirical mean of $P(Y \mid \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{Z}_a = \boldsymbol{z})$ and $P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le l-1)$. Also, we denote $T_a = \min_{\boldsymbol{z}} T_{a,\boldsymbol{z}}$. Using the above Eq.(10), we estimate each term of the right-hand side for every $\boldsymbol{z} \in \{0,1\}^{Z_a}$ to obtain an estimate for $\mathbb{E}[Y \mid a]$ as follows:

$$\hat{\mu}_{O,a} = \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \prod_{l=1}^{k} p_{a,\boldsymbol{z},l}(t). \tag{14}$$

For general graphs, there is no efficient algorithm to determine the existence of the admissible sequence and extract it when it exists. But we could rely on several methods to find admissible sequences in some special cases. First, we can find the *adjustment set*, a special case of admissible sequences. For a causal graph G, $\boldsymbol{Z}$ is an adjustment set for variable Y and set $\boldsymbol{S}$ if and only if $P(Y = 1 \mid do(\boldsymbol{S} = \boldsymbol{s})) = \sum_{\boldsymbol{z}} P(Y = 1 \mid \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{Z} = \boldsymbol{z}) P(\boldsymbol{Z} = \boldsymbol{z})$. There is an efficient algorithm for deciding the existence of a minimal adjustment set with respect to any set $\boldsymbol{S}$ and Y and finding it (van der Zander et al., 2019). Second, for general graphs without hidden variables, the admissible sequence can be easily found by $\boldsymbol{Z}_j = \boldsymbol{Pa}(X_j) \setminus (\boldsymbol{Z}_1 \cup \cdots \boldsymbol{Z}_{j-1} \cup X_1 \cdots \cup X_{j-1})$ (See Theorem 4 in Appendix D.2). Finally, when the causal graph satisfies certain properties, there exist algorithms to decide and construct admissible sequences Dawid & Didelez (2010).

Algorithm 3 provides the pseudocode of our algorithm CCPE-General, which has the same framework as Algorithm 1. The main difference is in the first step of updating observational estimates, in which we rely on the do-calculus formula Eq.(10).

For an action $a = do(\boldsymbol{S} = \boldsymbol{s})$ without an admissible sequence, define $q_a^{(G)} = 0$, meaning that it is hard to be estimated through observation. Otherwise, define q_a as:

$$q_a^{(G)} = \min_{\boldsymbol{z}}\{q_{a,\boldsymbol{z}}\}, \text{where } q_{a,\boldsymbol{z}} = P(\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{Z}_a = \boldsymbol{z}), \forall \boldsymbol{z} \in \{0,1\}^{Z_a}. \tag{15}$$

Similar to CCPE-BGLM, for $a = do(\boldsymbol{S} = \boldsymbol{s})$ with $|\boldsymbol{S}| = k$, we use observational and interventional confidence radius as:

$$\beta_O^a(n,t) = \alpha_O\sqrt{\frac{1}{n}\log\frac{20k|\boldsymbol{A}|Z_a I_a \log(2t)}{\delta}}; \beta_I(t) = \alpha_I\sqrt{\frac{1}{n}\log\frac{|\boldsymbol{A}|\log(2t)}{\delta}}, \tag{16}$$

where α_O and α_I are exploration parameters, and $I_a = 2^{Z_a}$. For a theoretical guarantee, we will choose α_O = 8 and α_I = 2. Our sample complexity result is given below.

Theorem 2. *With probability $1-\delta$,* CCPE-General$(G, \boldsymbol{A}, \varepsilon, \delta/5)$ *returns an ε-optimal arm with sample complexity*

$$T = O\left(H_{m_{\varepsilon,\Delta}^{(G)}} \log\frac{|\boldsymbol{A}|H_{m_{\varepsilon,\Delta}^{(G)}}}{\delta}\right), \tag{17}$$

where $m_{\varepsilon,\Delta}^{(G)}, H_{m_{\varepsilon,\Delta}^{(G)}}$ are defined in Definitions 2 and 3 in terms of $q_a^{(G)}$'s defined in Eq.(15).

Comparing to LUCB1, since $m_{\varepsilon,\Delta}^{(G)} \le |\boldsymbol{A}|$, our algorithm is always as good as LUCB1. It is easy to construct cases where our algorithm would perform significantly better than LUCB1. Comparing to other causal bandit algorithms, our algorithm also performs significantly better, especially when $m_{\varepsilon,\Delta}^{(G)} \ll m^{(G)}$ or the gap Δ_a is large relative to ε. Some causal graphs with candidate action sets and valid admissible sequence are provided in Appendix A, and more discussion is in Appendix B.

5.2 Lower Bound for the General Graph Case

To show that our CCPE-General algorithm is nearly minimax optimal, we provide the following lower bound, which is based on parallel graphs. We consider the following class of parallel bandit instance ξ with causal graph $G = (\{X_1, \cdots, X_n, Y\}, E)$: the action set is $\boldsymbol{A} = \{do(X_i = x) \mid x \in \{0,1\}, 1 \le i \le n\} \cup \{do()\}$. The $q_a^{(G)}$ in this case is reduced to $q_{do(X_i=x)}^{(G)} = P(X_i = x)$ and $q_{do()} = 1$. Sort the action set as $q_{a_1}^{(G)} \cdot \max\{\Delta_{a_1}, \varepsilon/2\}^2 \le q_{a_2}^{(G)} \cdot \max\{\Delta_{a_2}, \varepsilon/2\}^2 \le \cdots \le q_{a_{2n+1}}^{(G)} \cdot \max\{\Delta_{a_{2n+1}}, \varepsilon/2\}^2$. Let $p_{\min} = \min_{\boldsymbol{x}\in\{0,1\}^n} P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}), p_{\max} = \max_{\boldsymbol{x}\in\{0,1\}^n} P(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$. Let $p_{\max} + 2\Delta_{2n+1} + 2\varepsilon \le 0.9, p_{\min} + \Delta_{\min} \ge 0.1$.

Theorem 3. *For the parallel bandit instance class ξ defined above, any (ε, δ)-PAC algorithm has expected sample complexity at least*

$$\Omega\left(\left(H_{m_{\varepsilon,\Delta}^{(G)}-1} - \frac{1}{\min_{i<m_{\varepsilon,\Delta}^{(G)}}\max\{\Delta_{a_i}, \varepsilon/2\}^2} - \frac{1}{\max\{\Delta_{do(),\varepsilon/2}\}^2}\right)\log\left(\frac{1}{\delta}\right)\right). \tag{18}$$

Theorem 3 is the first gap-dependent lower bound for causal bandits, which needs brand-new construction and technique. Comparing to the upper bound in Theorem 2, the main factor $H_{m_{\varepsilon,\Delta}^{(G)}}$ is the same, except that the lower bound subtracts several additive terms. The first term $H_{m_{\varepsilon,\Delta}-1}$ is almost equal to $H_{m_{\varepsilon,\Delta}}$ appearing in Eq.(17), except the it omits the last and the smallest additive term in $H_{m_{\varepsilon,\Delta}}$. The second term is to eliminate one term with minimal Δ_{a_i}, which is common in multi-armed bandit. (Lattimore (2018),Karnin et al. (2013a)) The last term is because $do()$'s reward must be in-between $\mu_{do(X_i=0)}$ and $\mu_{do(X_i=1)}$ and thus cannot be the optimal arm.

6 Future Work

There are many interesting directions worth exploring in the future. First, how to improve the computational complexity for CPE of causal bandits is an important direction. Second, one can consider developing efficient pure exploration algorithms for causal graphs with partially unknown graph structures. Lastly, identifying the best intervention may be connected with the markov decision process and studying their interactions is also an interesting direction.

REFERENCES

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12761–12773. Curran Associates, Inc., 2021.

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

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pp. 379–387, Cambridge, MA, USA, 2014. MIT Press.

Alexander Dawid and Vanessa Didelez. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4, 10 2010. *do*10.1214/10-SS081.

Yihan Du, Yuko Kuroki, and Wei Chen. Combinatorial pure exploration with full-bandit or partial linear feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:7262–7270, 05 2021. *do*10.1609/aaai.v35i8.16892.

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 (eds.), *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*, pp. 255–270. Springer, 2002. *do*10.1007/3-540-45435-7_18.

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

Kristjan Greenewald, Dmitriy Katz, Karthikeyan Shanmugam, Sara Magliacane, Murat Kocaoglu, Enric Boix Adsera, and Guy Bresler. Sample efficient active learning of causal trees. *Advances in Neural Information Processing Systems*, 32, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

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)*, pp. 1–6, 2014. *do*10.1109/CISS.2014.6814096.

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.

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.

Zohar Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. *30th International Conference on Machine Learning, ICML 2013*, pp. 2275–2283, 01 2013a.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. III–1238–III–1246. JMLR.org, 2013b.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016a.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016b.

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

T. Lattimore. Refining the confidence level for optimistic bandit strategies. *Journal of Machine Learning Research*, 19:1–32, 07 2018.

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*, pp. 661–670, 2010.

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

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.

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*, pp. 141–150. PMLR, 2020.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24817–24828. Curran Associates, Inc., 2021.

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

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.

Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ -good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33:20707–20718, 2020.

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

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

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

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

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

Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence*, 270:1–40, 2019.

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*, pp. 5512–5520. PMLR, 2018.

Chang Ming Yin and Lincheng Zhao. Strong consistency of maximum quasi-likelihood estimates in generalized linear models. *Science in China Series A: Mathematics*, 48:1009–1014, 2005.

Zhijie Zhang, Wei Chen, Xiaoming Sun, and Jialin Zhang. Online influence maximization with node-level feedback using standard offline oracles. In *AAAI*, 2022.

APPENDIX

A GENERAL CLASSES OF GRAPHS SUPPORTING THEOREM 2

For Theorem 2, in this section we show some graphs with small size of admissible sequence for all arms, which makes our result much better than previous algorithms. By comparison in the Appendix B, we show that if $2^{Z+l} \leq |\boldsymbol{A}|$, where $Z = \max_a Z_a, l = \max\{|\boldsymbol{S}| \mid do(\boldsymbol{S} = \boldsymbol{s}) \in A\}$, our algorithm can perform better than previous classical bandit algorithms.

Two-layer graphs Consider $\boldsymbol{X} = A \cup B$, where $A = \{X_1, \cdots, X_k\}$ is the set of key variables, $B = \{X_{k+1} \cdots, X_n\}$ are the rest of variables. Now we consider $k \leq \frac{1}{2}\log_2 n$, and the edge set is in $E \subseteq \{(X_i \to X_j) \mid X_i \in A, X_j \in A\} \cup \{(X_i \to X_j) \mid X_i \in A, X_j \in B\} \cup \{(X_i \to Y) \mid X_i \in B\}$. There can also exist some hidden confounders between two variables in A, namely, $A_1 \leftarrow U \to A_2$ for unobserved variables U and $A_1, A_2 \in A$.

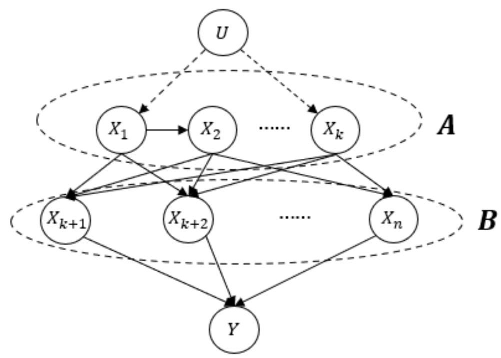


Figure 1: An Example of Two-layer Graphs

We define the action set as $\{do(\boldsymbol{S} = \boldsymbol{s}) \mid \boldsymbol{S} \subset B, |\boldsymbol{S}| \leq l, \boldsymbol{s} \in \{0,1\}^{|\boldsymbol{S}|}\}$ for some l. Then, since for arm $do(\boldsymbol{S} = \boldsymbol{s})$, A is the adjustment set for it, we know $Z_a \leq k \leq \frac{1}{2}\log_2 n$ for all action a. Then $2^{Z+l} \leq \sqrt{n} \cdot 2^l < \binom{n}{l} \cdot 2^l < |\boldsymbol{A}|$.

Consider the scenario in which a farmer wants to optimize the crop yield Lattimore et al. (2016). $A = \{X_1, X_2, \cdots, X_k\}$ are key elements influencing crop yields, such as temperature, humidity, and soil nutrient. $B = \{X_{k+1}, \cdots, X_n\}$ are different kinds of crops, and Y is the final total reward collected from all crops. Each kind of crop may be influenced by key elements in A in different ways. Moreover, the elements in A may have some causal relationships: higher humidity will lead to lower temperature. The above causal graph represents this problem very well.

Collaborative graphs Consider $\boldsymbol{X} = \boldsymbol{X}^1 \cup \boldsymbol{X}^2 \cup \cdots \cup \boldsymbol{X}^l$, where each $\boldsymbol{X}^i (1 \leq i \leq l)$ has at most $k \leq \frac{1}{2}\log n$ nodes. The edge set is contained in $E = \{X \to Y \mid X \in \boldsymbol{X}\} \cup \{X_i \to X_j \mid X_i, X_j \in \boldsymbol{X}^t, 1 \leq t \leq l\}$. In each subgraph $\boldsymbol{X}^i$, we allow the existence of unobserved confounders between two variables in $\boldsymbol{X}^i$. (We use dashed arrows to represent the confounders.) We call this class of graphs collaborative graphs (see Figure 2), since it is modified by Addanki & Kasiviswanathan (2021) on collaborative causal discovery.

For simplicity, the action set is defined by $\{do(\boldsymbol{S} = \boldsymbol{s}) \mid |\{\boldsymbol{S} \cap \boldsymbol{X}^i\}| \leq 1, |\boldsymbol{S}| \leq d\}$. Then for a particular $\boldsymbol{S} = \{X_{i_1}, X_{i_2}, \cdots, X_{i_d}\}$ and $i_1, \cdots, i_d$ such that $X_{i_j} \in \boldsymbol{X}^{i_j}, 1 \leq j \leq d$ for some $d \in [0, l]$ For these graphs, we know $T = \cup_{j=1}^d \boldsymbol{X}^{i_j} \setminus \boldsymbol{S}$ is a adjustment set (then also a admissible sequence) for $\boldsymbol{S}$ and Y with $|T| \leq \frac{1}{2} d \log n$. Then $2^Z < n^{d/2} \cdot 2^d < \binom{n/k}{d} \cdot 2^d < |\boldsymbol{A}|$ when n is large. Collaborative graphs are useful in many real-world scenarios. For example, many companies want to cooperate and maximize their profits. Then each subgraph $\boldsymbol{X}^i (1 \leq i \leq l)$ represents a company, and they want to find the best intervention to generate the maximum profit.

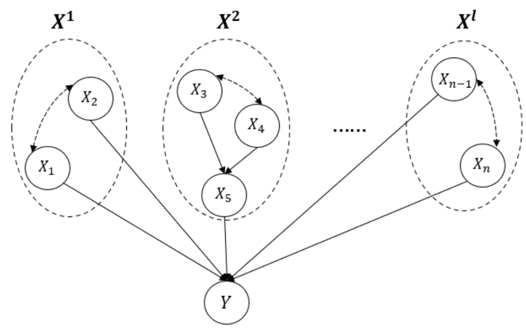


Figure 2: An Example of Collaborative Graphs

Causal Tree Causal tree is a useful structure in real scenario, which is consider in Lu et al. (2021) and Greenewald et al. (2019). In this class of graph, the underlying causal graph of causal model is a directed tree, in which all its edges point away from the root. Denote the root as layer 0, and layer i, $\boldsymbol{L}_i$ contains all the nodes with distance i to the root. For simplicity, we assume all unobserved confounders point to two nodes in same layer. For a set T, its c-component C_T means all the nodes connected to T by only bi-directed edges (confounders) .

For each action set $\{do(\boldsymbol{S} = \boldsymbol{s}) \mid \boldsymbol{S} \subseteq \boldsymbol{X}\}$, we consider $\boldsymbol{S} \cap \boldsymbol{L}_i = \boldsymbol{S}_i$. Then the sequence $\boldsymbol{Z}_i = C_{\boldsymbol{S}_i} \cup \boldsymbol{Pa}(C_{\boldsymbol{S}_i}) \setminus \{Z_0, \cdots, Z_{i-1}, S_0, \cdots, S_{i-1}\}$ is the admissible sequence. We give an example in Figure 3. For example, if we consider action $do(\{X_3, X_4, X_8\} = \boldsymbol{s}, \boldsymbol{s} \in \{0,1\}^3)$, then the admissible sequence is $\boldsymbol{Z}_1 = \{X_1, X_2\}, \boldsymbol{Z}_2 = \emptyset, \boldsymbol{Z}_3 = \{X_7\}$, and we can write

$$P(Y \mid do(X_3, X_4, X_8)) = \sum_{X_1, X_2, X_7} P(Y \mid X_1, X_2, X_3, X_4, X_7, X_8)$$
$$P(X_1, X_2) P(X_7 \mid X_1, X_2, X_3, X_4).$$

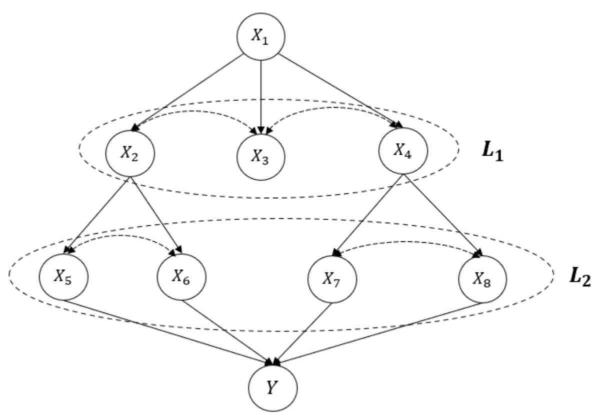


Figure 3: An Example of Causal Tree with Confounders

B More Detailed Comparison of Sample Complexity

Here we provide a bit more detailed sample complexity comparison between our Theorem 2 on general graphs with hidden variables and prior studies.

Compare with LUCB1 algorithm Comparing to LUCB1, since $m^{(G)}_{\varepsilon,\Delta} \leq |\boldsymbol{A}|$, our algorithm will not perform worse than LUCB1. Our algorithm can also perform much better than LUCB1 algorithm in some cases. For example, when we consider $\boldsymbol{A} = \{do(\boldsymbol{S} = \boldsymbol{s}) \mid |\boldsymbol{S}| = k, \boldsymbol{s} \in \{0,1\}^{|\boldsymbol{S}|}\}$ for some constant k, we have $|\boldsymbol{A}| = \binom{n}{k} \cdot 2^k$. Assume $Z = \max_a Z_a \le c \log n$ for a constant c, and q_a can be approximately $\Theta(\frac{1}{2^{Z_a+k}})$ ($q_a = \min_{\boldsymbol{z}} P(\boldsymbol{S} = \boldsymbol{s}, \boldsymbol{Z}_a = \boldsymbol{z})$) for all action a. Then we can get $m^{(G)}_{\varepsilon,\Delta} \approx 2^{Z+k} \le n^c \cdot 2^k = o(|\boldsymbol{A}|)$. Thus our algorithm performs much better than LUCB1.

Compare with previous causal bandit algorithms Since there is no previous causal bandit algorithm working on combinatorial action set with hidden variables, we compare two previous causal bandit algorithms in some special cases. First, compare to Lattimore et al. (2016) with parallel graph and atomic intervention, we first transfer the simple regret result in (Lattimore et al., 2016) to sample complexity $\widetilde{O}(\frac{m^{(G)}}{\varepsilon^2}\log(\frac{1}{\delta}))$. For parallel graph and $a = do(X_i = x)$, we know $q_a = P(X_i = x)$ since there is no parent for X_i, and our algorithm result is $\widetilde{O}(H_{m_{\varepsilon,\Delta}})$. Then since $m^{(G)}_{\varepsilon,\Delta} = O(m^{(G)})$ and $\max\{\Delta_a, \varepsilon/2\} \ge \varepsilon/2$, our algorithm always perform better. When the gap Δ_a is large relative to ε, our algorithm perform much better because of our gap-dependent sample complexity. Yabe et al. (2018) consider combinatorial intervention on graphs without hidden variables, so we can compare our algorithm's result with theirs in this setting. We also transfer their simple regret result to sample complexity $\widetilde{O}(\frac{\max\{nC, n|\boldsymbol{A}|\}}{\varepsilon^2}\log(1/\delta))$, where $C = \sum_{X\in\boldsymbol{X}\cup\{Y\}} 2^{|\boldsymbol{Pa}(X)|}$. Note that when $|\boldsymbol{Pa}(Y)|$ is large, $C \ge 2^{|\boldsymbol{Pa}(Y)|}$ can be really large. However, our algorithm even does not need the knowledge of $\boldsymbol{Pa}(Y)$. Indeed, considering $\max_{a=do(\boldsymbol{S}=\boldsymbol{s})} |\boldsymbol{S}| = k$ is a constant, and assume $Z \le \log C - k$ and $q_a = \Theta(\frac{1}{2^{Z+k}})$, we have $m^{(G)}_{\varepsilon,\Delta} \le \Theta(C)$, then our dominating term $H^{(G)}_{m_{\varepsilon,\Delta}}$ is smaller than $\frac{nC}{\varepsilon^2}$ because both $\max\{\Delta_a, \varepsilon/2\} \ge \varepsilon/2$ and $|M^{(G)}| = m^{(G)}_{\varepsilon,\Delta} \le nC$. Also, at the worst case our algorithm's sample complexity is not more than $\widetilde{O}(\frac{|\boldsymbol{A}|}{\varepsilon^2}\log(\frac{1}{\delta}))$, while the algorithm in Yabe et al. (2018) may result in $\widetilde{O}(\frac{n|\boldsymbol{A}|}{\varepsilon^2}\log(1/\delta))$. The experiments are provided in Appendix E.

In summary, when compared to prior studies on causal bandit algorithms, our algorithm wins when the reward gaps are relatively large or the size of the admissible sequence is small; and when compared to prior studies on adaptive pure exploration algorithms, our algorithm wins by estimating do effects using observational data and saving estimates on those easy actions.

C Proof of Theorems

C.1 Proof of Theorem 1

Proof. We first provide a lemma in Li et al. (2020) to show the confidence for the maximum likelihood estimation.

Lemma 3. *For one node $X \in \boldsymbol{X} \cup \{Y\}$, assume Assumption 1 and 2 holds, and*

$$\lambda_{min}(M_{t,X}) \ge \frac{512D(M^{(2)})^2}{\kappa^4}(D^2 + \ln\frac{3nt^2}{\delta}),$$

with probability $1 - \delta/nt^2$, for any vector $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$, at all rounds t the estimator $\hat{\boldsymbol{\theta}}_{t,X}$ in Algorithm 2 satisfy

$$|\boldsymbol{v}^\top(\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*_X)| \le \frac{3}{\kappa}\sqrt{\log(3nt^2/\delta)}||v||_{M^{-1}_{t,X}}.$$

Since we need to estimate $\boldsymbol{\theta}_{t,X}$ for all nodes, let F_1 be the event that the above inequality doesn't hold, then by union bound, $\Pr\{F_1\} \le n\sum_{t>1}\frac{\delta}{nt^2} \le \delta$.(We can consider $t > 1$) Now from Feng & Chen (2022), the true mean $\sigma(\hat{\boldsymbol{\theta}}_t, X_i)$ and our estimation $\sigma(\boldsymbol{\theta}^*, X_i)$ can be bounded by Lemma 2. We rewrite the Lemma 2 here, and give proof in Appendix D.3.

Lemma 2. *For an action $a = do(\boldsymbol{S} = \boldsymbol{s})$ and any two weight vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we have*

$$|\sigma(\boldsymbol{\theta}, a) - \sigma(\boldsymbol{\theta}', a)| \le \mathbb{E}_{\boldsymbol{e}}\left[\sum_{X\in N_{\boldsymbol{S},Y}} |\boldsymbol{V}_X^\top(\boldsymbol{\theta}_X - \boldsymbol{\theta}'_X)|M^{(1)}\right], \tag{6}$$

where $N_{\boldsymbol{S},Y}$ is the set of all nodes that lie on all possible paths from X_1 to Y excluding $\boldsymbol{S}$, $\boldsymbol{V}_X$ is the value vector of a sample of the parents of X according to parameter $\boldsymbol{\theta}$, $M^{(1)}$ is defined in Assumption 1, and the expectation is taken on the randomness of the noise term $\boldsymbol{e} = (e_X)_{X \in \boldsymbol{X} \cup \{Y\}}$ of causal model under parameter $\boldsymbol{\theta}$.

By definition, for any action $a = do(\boldsymbol{M} = s)$, $|P_{\boldsymbol{S},Y}| = \ell_a \in \{1, \cdots, n\}$. We then introduce Lecué and Mendelson's Inequality represented in Nie (2021).

Lemma 4 (Nie (2021) Lecué and Mendelson's Inequality)**.** *Let random column vector $\boldsymbol{v} \in \mathbb{R}^D$, and $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ are n independent copies of $\boldsymbol{v}$. Assume $\boldsymbol{z} \in Sphere(D)$ such that*

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

then there exists a constant $c > 0$ such that when $n \geq \frac{cD}{\beta^2}$

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

This lemma can help us to bound the minimum eigenvalue for $M_{t,X} = \sum_{1\leq i\leq t} \boldsymbol{V}_{t,X} \boldsymbol{V}_{t,X}^\top$. To satisfy the condition for Lemma 4, we provide a similar lemma in Feng & Chen (2022):

Lemma 5. *Under Assumption 3, for any node $X \in \boldsymbol{X}$ and $\boldsymbol{v} \in Sphere(|\boldsymbol{Pa}(X)|)$,*

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

Proof. The proof is similar to Feng & Chen (2022) with a modification. For completeness, we provide the full proof below. Let $|\boldsymbol{Pa}(X)| = d \leq D$, $\boldsymbol{z} = (z_1, z_2, \cdots, z_d)$. Let $\boldsymbol{Pa}(X) = (X_{i_1} = X_1, X_{i_2}, \cdots, X_{i_d})$ and $\boldsymbol{pa}(X) = (x_{i_1} = 1, x_{i_2}, \cdots, x_{i_d})$. 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

$$\begin{aligned}
|\boldsymbol{pa}(X) \cdot \boldsymbol{v}| &\geq |z_1| - \sum_{i=2}^{d} |z_i| \\
&\geq \frac{d_0}{\sqrt{d_0^2+1}} - \sqrt{(d-1)\sum_{i=2}^{d}|z_i|^2} \\
&\geq \frac{d_0}{\sqrt{d_0^2+1}} - \sqrt{(d-1)(1 - \frac{d_0^2}{d_0^2+1})} \\
&= \frac{1}{2\sqrt{(d_0^2+1)(d-1)}} \\
&= \frac{1}{4d^2 - 3}.
\end{aligned}$$

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

$$|z_2| \geq \frac{1}{\sqrt{d-1}}\sqrt{\sum_{i=2}^{d}|z_i|^2} \geq \frac{\sqrt{1 - (d_0/\sqrt{d_0^2+1})^2}}{\sqrt{d-1}} = \frac{1}{\sqrt{4d^2-3}}. \tag{19}$$

By Assumption 3

$$\begin{aligned}
&\Pr\{X_{i_1} = 1, X_{i_2} = x_{i_2}, \cdots, X_{i_d} = x_{i_d}\} \\
&= \Pr\{X_{i_2} = x_{i_2} \mid X_{i_1} = 1, X_{i_3} = x_{i_3}, \cdots, X_{i_d} = x_{i_d}\} \cdot \Pr\{X_{i_1} = 1, X_{i_3} = x_{i_3}, \cdots, X_{i_d} = x_{i_d}\} \\
&\geq \eta \cdot \Pr\{X_{i_1} = 1, X_{i_3} = x_{i_3}, \cdots, X_{i_d} = x_{i_d}\},
\end{aligned}$$

we have

$$
\begin{aligned}
&\Pr\left\{|\boldsymbol{Pa}(X)\cdot\boldsymbol{z}| \geq \frac{1}{\sqrt{4D^2-3}}\right\} \\
&= \sum_{x_{i_3},\cdots,x_{i_d}} \Pr\{X_{i_1}=1, X_{i_2}=1, X_{i_3}=x_{i_3}\cdots, X_{i_d}=x_{i_d}\}\cdot\mathbb{I}\left\{|(1,1,x_{i_3},\cdots,x_{i_d})\cdot(z_1,\cdots,z_d)| \geq \frac{1}{\sqrt{4D^2-3}}\right\} \\
&+ \sum_{x_{i_3},\cdots,x_{i_d}} \Pr\{X_{i_1}=1, X_{i_2}=0, X_{i_3}=x_{i_3}\cdots, X_{i_d}=x_{i_d}\}\cdot\mathbb{I}\left\{|(1,0,x_{i_3},\cdots,x_{i_d})\cdot(z_1,\cdots,z_d)| \geq \frac{1}{\sqrt{4D^2-3}}\right\} \\
&\geq \eta \sum_{x_{i_3},\cdots,x_{i_d}} \Pr\{X_{i_1}=1, X_{i_3}=x_{i_3}\cdots, X_{i_d}=x_{i_d}\}\cdot\mathbb{I}\left\{|(1,1,x_{i_3},\cdots,x_{i_d})\cdot(z_1,\cdots,z_d)| \geq \frac{1}{\sqrt{4D^2-3}}\right\} \\
&+ \eta \sum_{x_{i_3},\cdots,x_{i_d}} \Pr\{X_{i_1}=1, X_{i_3}=x_{i_3}\cdots, X_{i_d}=x_{i_d}\}\cdot\mathbb{I}\left\{|(1,0,x_{i_3},\cdots,x_{i_d})\cdot(z_1,\cdots,z_d)| \geq \frac{1}{\sqrt{4D^2-3}}\right\} \\
&\geq \eta \sum_{x_{i_3},\cdots,x_{i_d}} \Pr\{X_{i_1}=1, i_3=x_{i_3}\cdots, X_{i_d}=x_{i_d}\}\cdot \\
&\left(\mathbb{I}\left\{|(1,1,x_{i_3},\cdots,x_{i_d})\cdot(z_1,\cdots,z_d)| \geq \frac{1}{\sqrt{4D^2-3}}\right\} + \mathbb{I}\left\{|(1,0,x_{i_3},\cdots,x_{i_d})\cdot(z_1,\cdots,z_d)| \geq \frac{1}{\sqrt{4D^2-3}}\right\}\right) \\
&\geq \eta.
\end{aligned}
$$

where the last inequality is because

$$\sum_{x_{i_3},\cdots,x_{i_d}} \Pr\{X_{i_1}=1, i_3=x_{i_3}\cdots, X_{i_d}=x_{i_d}\} = 1,$$

and

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

The above equation is because otherwise

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

which leads to a contradiction of Eq. (19). We thus complete the proof of Lemma 5. □

Now let F_2 be the event

$$F_2 = \left\{\exists X \in \boldsymbol{X}\cup\{Y\},\ \lambda_{min}\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{V}_{i,X}\boldsymbol{V}_{i,X}^{\top}\right) \leq \frac{\eta}{2(4D^2-3)},\ \forall t \geq \frac{cD}{\eta^2}\log\frac{nt^2}{\delta}\right\}.$$

Then

$$
\begin{aligned}
\Pr\{F_2\} &\leq n \sum_{t\geq(cD/\eta^2)\log(nt^2/\delta)} e^{-t\eta^2/c} \\
&\leq n \sum_{t\geq(cD/\eta^2)\log(nt^2/\delta)} \frac{\delta}{nt^2} \\
&\leq (\frac{\pi^2}{3}-1)\delta \\
&\leq \delta.
\end{aligned}
$$

Now from Lemmas 2, 3 and 4, for all $a = do(\boldsymbol{S}=1)$, with probability $1-2\delta$, for all $t \geq \max\{\frac{cD}{\eta^2}\log\frac{nt^2}{\delta}, \frac{1024(M^{(2)})^2(4D^2-3)D}{\kappa^4\eta}(D^2+\ln\frac{1}{\delta})\}$, we can deduce that

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

Then

$$\begin{aligned}
|\sigma(\hat{\boldsymbol{\theta}}_t, \boldsymbol{S}) - \mu_a| &\leq \sum_{X \in P_{\boldsymbol{S},Y}} |\boldsymbol{V}_{t,X}^\top (\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)| M^{(1)} \\
&\leq \frac{3M^{(1)}}{\kappa} \sqrt{\log(3nt^2/\delta)} \sum_{X \in P_{\boldsymbol{S},Y}} ||\boldsymbol{V}_{t,X}||_{M_{t,X}^{-1}} \\
&\leq \frac{3M^{(1)}}{\kappa} \sqrt{\log(3nt^2/\delta)} \sum_{X \in P_{\boldsymbol{S},Y}} \frac{\sqrt{D}}{\sqrt{\lambda_{min}(M_{t,X})}} \\
&\leq \frac{3\sqrt{2}M^{(1)}}{\kappa} \sqrt{D(4D^2-3)} \sqrt{\log(3nt^2/\delta)} \sum_{X' \in P_{\boldsymbol{S},Y}} \frac{1}{\sqrt{\eta t}} \\
&\leq \frac{6\sqrt{2}M^{(1)}}{\kappa\sqrt{\eta}} \sqrt{\frac{\ell_a^2 D^3}{t} \log(3nt^2/\delta)} \\
&= \frac{6\sqrt{2}M^{(1)}}{\kappa\sqrt{\eta}} \sqrt{\frac{1}{q_a t} \log(3nt^2/\delta)} \\
&= \beta_O^a(t).
\end{aligned}$$

Now we prove that Algorithm 1 must terminate after $\lceil T \rceil$ rounds, where $T = \frac{1152(M^{(1)})^2}{\kappa^2\eta} H_{m_{\varepsilon,\Delta}^{(L)}} \log \frac{3nT^2}{\delta} + 16 H_{m_{\varepsilon,\Delta}^{(L)}} \log \frac{4|\boldsymbol{A}|\log(2T)}{\delta}$. In the following proof, we assume F_1 and F_2 do not happen. Then the true mean will not out of observational confidence bound and interventional confidence bound.

When $t \geq T_1$ such that $T_1 = \frac{1152(M^{(1)})^2}{\kappa^2\eta} H_{m_{\varepsilon,\Delta}^{(L)}} \log \frac{3nT_1^2}{\delta}$, for all $a \neq do()$ such that $q_a^{(L)} \geq \frac{1}{H_{m_{\varepsilon,\Delta}^{(L)}} \cdot \max\{\Delta_a, \varepsilon/2\}^2}$, let $\beta_a(t) = \frac{U_a^t - L_a^t}{2}$, we have

$$\beta_a(t) := \frac{U_a^t - L_a^t}{2} \leq \beta_O^a(\lceil T_1 \rceil) \leq \frac{6\sqrt{2}M^{(1)}}{\kappa\sqrt{\eta}} \sqrt{\frac{1}{q_a^{(L)} \lceil T_1 \rceil} \log(3nt^2/\delta)} \leq \frac{\max\{\Delta_a, \varepsilon/2\}}{4}.$$

Then we provide the following lemma:

Lemma 6. *If at round t, we have*

$$\beta_{a_h^t}(t) \leq \frac{\max\{\Delta_{a_h^t}, \varepsilon/2\}}{4}, \beta_{a_l^t}(t) \leq \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{4},$$

where a_h^t, a_l^t are the actions performed by algorithm at round t. then the algorithm will stop at round $t+1$.

Proof. From above, if the optimal arm $a^* = a_h^t$,

$$\begin{aligned}
\hat{\mu}_{a_l^t} + \beta_{a_l^t}(t) &\leq \mu_{a_l^t} + 2\beta_{a_l^t}(t) \\
&\leq \mu_{a_l^t} + \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{2} \\
&\leq \mu_{a_h^t} - \Delta_{a_l^t} + \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{2} \\
&\leq \hat{\mu}_{a_h^t} + \beta_{a^*}(T_{a^*}(t)) - \Delta_{a_l^t} + \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{2} \\
&\leq \hat{\mu}_{a_h^t} - \beta_{a^*}(T_{a^*}(t)) + \frac{\max\{\Delta_{a^*}, \varepsilon/2\} + \max\{\Delta_{a_l^t}, \varepsilon/2\}}{2} - \Delta_{a_l^t} \\
&\leq \hat{\mu}_{a_h^t} - \beta_{a^*}(T_{a^*}(t)) + \frac{\Delta_{a^*} + \varepsilon/2 + \Delta_{a_l^t} + \varepsilon/2}{2} - \Delta_{a_l^t}
\end{aligned}$$

$$\leq \hat{\mu}_{a_h^t} - \beta_{a^*}(T_{a^*}(t)) + \varepsilon.$$

If optimal arm $a^* \neq a_h^t$, and the algorithm doesn't stop at round $t+1$, then we prove $a^* \neq a_l^t$. Otherwise, assume $a^* = a_l^t$

$$\hat{\mu}_{a_h^t}^t \leq \mu_{a_h^t}^t + \frac{\max\{\Delta_{a_h^t}, \varepsilon/2\}}{4} \tag{20}$$

$$= \mu_{a_l^t}^t - \Delta_{a_h^t} + \frac{\max\{\Delta_{a_h^t}, \varepsilon/2\}}{4} \tag{21}$$

$$\leq \mu_{a_l^t}^t - \frac{3\Delta_{a_h^t}}{4} + \varepsilon/4 \tag{22}$$

$$\leq \hat{\mu}_{a_l^t}^t + \frac{\max\{\Delta_{a^*}, \varepsilon/2\}}{4} - \frac{3\Delta_{a_h^t}}{4} + \varepsilon/4 \tag{23}$$

$$\leq \hat{\mu}_{a_l^t}^t + \varepsilon/2 - \frac{\Delta_{a_h^t}}{2}. \tag{24}$$

From the definition of a_h^t, we know $\varepsilon > \Delta_{a_h^t} \geq \Delta_{a^*}, \beta_{a_h^t}(t) \leq \varepsilon/4, \beta_{a_l^t}(t) \leq \varepsilon/4$. Then $\hat{\mu}_{a_l^t}^t + \beta_{a_l^t}(t) + \beta_{a_h^t}(t) \leq \hat{\mu}_{a_l^t} + \varepsilon/2 \leq \hat{\mu}_{a_h^t}^t + \varepsilon$, which means the algorithm stops at round $t+1$.

Now we can assume $a^* \neq a_l^t, a^* \neq a_h^t$. Then

$$\mu_{a_l^t} + 2\beta_{a_l^t}(t) \geq \hat{\mu}_{a_l^t} + \beta_{a_l^t}(t) \geq \hat{\mu}_{a^*} + \beta_{a^*}(T_{a^*}(t)) \geq \mu_{a^*} = \mu_{a_l^t} + \Delta_{a_l^t}. \tag{25}$$

Thus

$$\Delta_{a_l^t} \leq 2\beta_{a_l^t}(t) \leq \frac{\max\{\Delta_{a_l^t}, \varepsilon/2\}}{2}, \tag{26}$$

which leads to $\Delta_{a_l^t} \leq \varepsilon/2, \beta_{a_l^t}(t) \leq \varepsilon/8$. Since

Also,

$$\mu_{a_h^t} + \beta_{a_h^t}(t) \geq \hat{\mu}_{a_h^t} \geq \hat{\mu}_{a_l^t} \geq \mu_{a^*} - \beta_{a_l^t}(t) = \mu_{a_h^t} + \Delta_{a_h^t} - \beta_{a_l^t}(t), \tag{27}$$

which leads to

$$\frac{\max\{\Delta_{a_h^t}, \varepsilon/2\}}{4} \geq \Delta_{a_h^t} - \varepsilon/8, \tag{28}$$

and $\Delta_{a_h^t} \leq \varepsilon/2, \beta_{a_h^t}(t) \leq \varepsilon/8$. Hence $\hat{\mu}_{a_l^t}^t + \beta_{a_l^t}(t) + \beta_{a_h^t}(t) \leq \hat{\mu}_{a_l^t} + \varepsilon/2 \leq \hat{\mu}_{a_h^t}^t + \varepsilon$, which means the algorithm stops at round $t+1$. □

Denote $N_a(t)$ as the value of variable N_a at round t. So by Lemma 6, when $t \geq T_1$, at each round at least one intervention will be performed on some actions a with $\beta_a(t) \geq \frac{\max\{\Delta_a, \varepsilon/2\}}{4}$, which implies that $q_a < \frac{1}{H_{m_{\varepsilon,\Delta}^{(L)}} \cdot \max\{\Delta_a, \varepsilon/2\}^2}$, and $N_a(t) \leq \frac{64}{\max\{\Delta_a, \varepsilon/2\}^2} \log \frac{|\boldsymbol{A}| \log(2t)}{\delta}$ (Since $\beta_a(t) \leq \beta_I^a(t) = 2\sqrt{\frac{1}{t} \log \frac{|\boldsymbol{A}| \log(2t)}{\delta}}$). Denote the set of these arms as M, so we have

$$T - T_1 \leq \sum_{a \in M} \frac{64}{\max\{\Delta_a, \varepsilon/2\}^2} \log \frac{|\boldsymbol{A}| \log(2t)}{\delta}$$
$$\leq 64(H_{m_{\varepsilon,\Delta}^{(L)}}) \log \frac{|\boldsymbol{A}| \log(2t)}{\delta},$$

Hence

$$T \leq \frac{1152(M^{(1)})^2}{\kappa^2 \eta} H_{m_{\varepsilon,\Delta}^{(L)}} \log \frac{3nT^2}{\delta} + 64 H_{m_{\varepsilon,\Delta}^{(L)}} \log \frac{|\boldsymbol{A}| \log(2T)}{\delta}.$$

Now we prove a sample complexity bound for Algorithm 1 by the lemma above:

Lemma 7. *If* $T = NQ\log\frac{3nT^2}{\delta} + 64Q\log\frac{|\boldsymbol{A}|\log(2T)}{\delta}$ *for some constant* N*, then* $T = O(Q\log(Q|\boldsymbol{A}|/\delta))$.

Proof. Since $f(x) = x - NQ\log\frac{3nx^2}{\delta} - 64Q\log\frac{|\boldsymbol{A}|\log(2x)}{\delta}$ is a increasing function when $T \geq 64Q$, we only need to show that there exists a constant $C \geq 64$ such that $f(CQ\log\frac{Q|\boldsymbol{A}|}{\delta}) \geq f(T) = 0$. Then

$$
\begin{aligned}
f(CQ\log\frac{Q|\boldsymbol{A}|}{\delta}) &= CQ\log\frac{Q|\boldsymbol{A}|}{\delta} - NQ\log\frac{3n}{\delta} - 2NQ\log(\frac{CQ\log(Q|\boldsymbol{A}|/\delta)}{\delta}) \\
&- 64Q\log\frac{|\boldsymbol{A}|\log(2(CQ\log(Q|\boldsymbol{A}|/\delta)))}{\delta} \\
&\geq (C - 2N\log C - N)Q\log\frac{Q|\boldsymbol{A}|}{\delta} - 64Q\log\frac{|\boldsymbol{A}|}{\delta} \\
&- (64+2N)Q\log(\log 2C + \log(Q\log Q|\boldsymbol{A}|/\delta)) \\
&\geq (C - 2N\log C - N - 64)Q\log\frac{Q|\boldsymbol{A}|}{\delta} \\
&- (64+2N)Q\log(\log 2CQ) - (64+2N)Q\log(\log Q|\boldsymbol{A}|/\delta) \qquad (29)\\
&\geq (C - 2N\log C - N - 192 - (64+2N)\log\log 2C)Q\log(Q|\boldsymbol{A}|/\delta). \qquad (30)
\end{aligned}
$$

The equation (29) and (30) are based on $\log(x+y) \leq \log(xy) = \log x + \log y$ when $x, y \geq 2$. Then choose C such that $C - 2N\log C - N - 192 - (64+2N)\log\log 2C \geq 0$, so we complete the proof. ∎

Hence, by Lemma 7 with $N = \frac{1152(M^{(1)})^2}{\kappa^2\eta}$, we know the total sample complexity is

$$T = O(H_{m^{(L)}_{\varepsilon,\Delta}}\log\frac{H_{m^{(L)}_{\varepsilon,\Delta}}|\boldsymbol{A}|}{\delta}).$$

Finally, we prove the correctness of our algorithm. Since the stopping rule is $\hat{\mu}^t_{a^t_l} + \beta_{a^t_l}(t) \leq \hat{\mu}^t_{a^t_h} - \beta_{a^t_h}(t) + \varepsilon$, if $a^* \neq a^t_h$, we have

$$
\begin{aligned}
\mu_{a^t_h} + \varepsilon \geq \hat{\mu}_{a^t_h} - \beta_{a^t_h}(t) + \varepsilon &\geq \hat{\mu}_{a^t_l} + \beta_{a^t_l}(t) \qquad (31)\\
&\geq \hat{\mu}_{a^*} + \beta_{a^*}(T_{a^*}(t)) \qquad (32)\\
&\geq \mu_{a^*}. \qquad (33)
\end{aligned}
$$

Hence either $a^* = a^t_h$ or a^t_h is ε-optimal arm. Thus, we complete the proof.

∎

C.2 Proof of Theorem 2

Proof. In this proof, we denote $T_{a,\boldsymbol{z}}(t), T_a(t), N_a(t)$ are the value of $T_{a,\boldsymbol{z}}, T_a, N_a$ respectively. For conveniece, we prove CCPE-General(G, ε, δ) outputs a ε-optimal arm with probability $1 - 3\delta$. For simplity, we denote $H_{m^{(G)}_{\varepsilon,\Delta}}$ as $H^{(G)}$. In round t, $T_{a,\boldsymbol{z}}(t) = \sum_{j=1}^t \mathbb{I}\{X_{j,i} = x, \boldsymbol{Pa}(X)_j = \boldsymbol{z}\}$, $\hat{q}_{a,\boldsymbol{z}} = \frac{T_{a,\boldsymbol{z}}(t)}{t}$. By Chernoff bound, at round t such that $q_{a,\boldsymbol{z}}(t) \geq \frac{6}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}$, with probability at most $\delta/3|\boldsymbol{A}|I_a$,

$$|\hat{q}_{a,\boldsymbol{z}} - q_{a,\boldsymbol{z}}| > \sqrt{\frac{6q_{a,\boldsymbol{z}}}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}.$$

Hence

$$\hat{q}_a = \min_{\boldsymbol{z}}\{\hat{q}_{a,\boldsymbol{z}}\} \leq \min_{\boldsymbol{z}}\{q_{a,\boldsymbol{z}} + \sqrt{\frac{6q_{a,\boldsymbol{z}}}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}\} = q_a + \sqrt{\frac{6q_a}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}. \qquad (34)$$

When $q_a \geq \frac{3}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}$, $f(x) = x - \sqrt{\frac{6x}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}$ is a increasing function.

$$\hat{q}_a \geq \min_{\boldsymbol{z}}\{q_{a,\boldsymbol{z}} - \sqrt{\frac{6q_{a,\boldsymbol{z}}}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}\} = q_a - \sqrt{\frac{6q_a}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}. \tag{35}$$

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, when t is large the true mean of an arm is out of range $[L^t_{O,a}, U^t_{O,a}]$ and $[L^t_{I,a}, U^t_{I,a}]$ respectively. Following anytime confidence bound,$\Pr\{F_3\} \leq \delta$. By Lemma 8 and 10 we prove $\Pr\{F_2\} \leq 3\delta$.

To prove the concentration bound, we need the following lemma, which is a Chernoff-type anytime confidence bound for Bernoulli variables. To our best knowledge, it is the first anytime confidence bound based on Chernoff inequality.

Lemma 8. *For $X_1, X_2, \cdots, X_n$ drawn from Bernoulli distribution with mean μ, denote $\bar{X} = \sum_{i=1}^n X_i$, then for all round t we have*

$$P(\overline{X} - \mu > 2\sqrt{\frac{3\mu}{t}\log\frac{20\log(2t)}{\delta}}, \forall t \geq \frac{3}{\mu}\log\frac{20\log(2t)}{\delta}) \leq 1-\delta.$$

The main proof is achieved by modification on part of Lemma 1 in Jamieson et al. (2013). For completeness, we provide the full proof here. Let $S_t = \sum_{i=1}^t (X_i - \mu)$ and $\phi(x) = \sqrt{3\mu x\log(\frac{\log(x)}{\delta})}$. We define the sequence $\{u_i\}_{i\geq 1}$ as follows: $u_0 = 1, u_{k+1} = \lceil (1+C)u_k \rceil$, where C is a constant. Then for simple union bound and Chernoff inequality, we have

$$\begin{aligned}
P(\exists k \geq 1 : S_{u_k} \geq \sqrt{1+C}\phi(u_k)) &\leq \sum_{k=1}^{\infty} \exp\left\{-\frac{(1+C)\cdot 3\mu u_k \log(\frac{\log(u_k)}{\delta})}{3\mu \cdot u_k}\right\} \\
&\leq \exp\left\{-(1+C)\log\left(\frac{\log(u_k)}{\delta}\right)\right\} \\
&\leq \sum_{k=1}^{\infty}\left(\frac{\delta}{k\log(1+C)}\right)^{1+C} \\
&\leq \left(1+\frac{1}{C}\right)\log\left(\frac{\delta}{\log(1+C)}\right)^{1+C}.
\end{aligned}$$

Then we proof Chernoff-type maximal Inequality:

$$P(\exists t \in [n], S_t \geq x) \leq \exp\left\{-\frac{x^2}{3\mu n}\right\}. \tag{36}$$

First, we know $\{S_t\}$ is a martingale and then $\{e^{S_t}\}$ is a non-negative submartingale. By Doob's submartingale inequality, we have

$$\begin{aligned}
P(\sup_{0\leq i\leq n} S_i \geq x) = P(\sup_{0\leq i\leq n} e^{S_i} \geq e^{sx}) \leq \frac{\mathbb{E}[e^{s\cdot S_n}]}{e^{sx}} &= \frac{(\mu e^{s\cdot(1-\mu)} + (1-\mu)e^{-s\mu})^n}{e^{sx}} \\
&= \frac{((1-\mu) + \mu e^s)^n}{e^{sx+sn\mu}} \\
&\leq \frac{e^{n\mu\cdot(e^s-1)}}{e^{sx+sn\mu}}.
\end{aligned}$$

Choose $s = \ln(1+\frac{x}{n\mu})$, by the proof of Chernoff bound with $\mu \geq \frac{3}{t}\log\frac{20\log(2t)}{\delta}$, we can easily get

$$P(\sup_{0\leq i\leq n} S_i \geq x) \leq \exp\left\{\frac{-x^2}{3\mu n}\right\}.$$

Now with this inequality, we can derive the lemma.

$$\begin{aligned}
&P(\exists t \in \{u_k+1, \cdots, u_{k+1}-1\} : S_t - S_{u_k} \geq \sqrt{C}\phi(u_{k+1})) \\
&\leq P(\exists t \in [u_{k+1}-u_k-1] : S_t \geq \sqrt{C}\phi(u_{k+1})) \\
&\leq \exp\left\{-C \cdot \frac{u_{k+1}}{u_{k+1}-u_k-1} \log\left(\frac{\log(u_{k+1})}{\delta}\right)\right\} \\
&\leq \exp\left\{-(1+C)\log\left(\frac{\log(u_{k+1})}{\delta}\right)\right\} \\
&\leq \left(\frac{\delta}{(k+1)\log(1+C)}\right)^{1+C} \\
&\leq \left(\frac{\delta}{\log(1+C)}\right)^{1+C}.
\end{aligned}$$

Now with probability at least $1-(2+1/C)\left(\frac{\delta}{\log(1+C)}\right)^{1+C}$, for $u_k \leq t \leq u_{k+1}$, we have

$$\begin{aligned}
S_t &= S_t - S_{u_k} + S_{u_k} \\
&\leq \sqrt{C}\phi(u_{k+1}) + \sqrt{1+C}\phi(u_k) \\
&\leq (1+\sqrt{C})\phi((1+C)t).
\end{aligned}$$

Now denote $\delta' = (2+1/C)\left(\frac{\delta}{\log(1+C)}\right)^{1+C}$, $\delta = \log(1+C)\left(\frac{C\delta'}{2+C}\right)^{\frac{1}{1+C}}$, we have with probability $1-\delta'$

$$P\left(S_t \geq (1+\sqrt{C})\sqrt{3(1+C)\mu t \log\left(\left(\frac{2+C}{C\delta'}\right)^{\frac{1}{1+C}} \cdot \frac{\log(1+C)t}{\log(1+C)}\right)}\right) \leq 1-\delta'.$$

Choose $C = 0.25$, and note that $\frac{\log(1.25t)}{\log(1.25)} \leq \frac{\log(2t)}{\log 2}$, $(2.25/0.25)^{0.8}/\log(2) < 10$ and $1.5 * \sqrt{1.25} < 2$, we complete the lemma's proof.

Lemma 9. *Denote $T_{a,\boldsymbol{z},l}(t)$ is the number of observations from round 1 to round t in which $\boldsymbol{Z}_{a,i} = \boldsymbol{z}_i, X_i = x_i, i \leq l-1$. Then we have $T_{a,\boldsymbol{z},l}(t) \geq 2^{k-l+1+|\boldsymbol{Z}_{a,l}|}T_{a,\boldsymbol{z}}(t)$.*

Proof. The proof is straightforward. Since $T_{a,\boldsymbol{z}}(t)$ is the number of observations from round 1 to round t in which $\boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, 1 \leq i \leq k$. Hence the number of observations for $\boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i$ for $i \leq l-1$ is at least $2^{|\boldsymbol{Z}_l|} \cdot 2^{k-(l-1)} \cdot T_{a,\boldsymbol{z}}(t) = 2^{k-l+1+|\boldsymbol{Z}_l|}T_{a,\boldsymbol{z}}(t)$ □

Lemma 10. *With probability $1-3\delta$, for all round t,*

$$|\hat{\mu}_{obs,a} - \mu_a| < 8\sqrt{\frac{1}{T_a(t)}\log\frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}, \tag{37}$$

where $I_a = 2^{|\boldsymbol{Z}_a|}$.

Proof. If $T_a(t) \leq 12\log\frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}$, then the right term of (37) is greater than 1, and this lemma always holds. In this proof, we denote $\boldsymbol{Z}_{a,i}$ as $\boldsymbol{Z}_i$ for simplity. By classical anytime confidence bound, we know with probability $1-\delta/(k \cdot I_a)$, for all round t we have

$$|r_{a,\boldsymbol{z}}(t) - P(Y=1 \mid X=x, \boldsymbol{Z}_a = \boldsymbol{z})| \leq \sqrt{\frac{4}{T_{a,\boldsymbol{z}}(t)}\log\frac{|\boldsymbol{A}|\log(2t)}{\delta}}.$$

First, let $s(t) = 20kZ_aI_a|\boldsymbol{A}|\log(2t)$, if $t < \frac{6}{q_a}\log(s(t)/\delta)$, then let $Q = \frac{6}{q_a}\log(s(t)/\delta)$, based on $T_a(t) \geq 12\log\frac{s(t)}{\delta}$, then

$$P\left(t < \frac{6}{q_a}\log(1/\delta)\right) \leq P\left(T_a(Q) \geq 12\log\frac{s(t)}{\delta}\right).$$

Thus by Chernoff bound, we know

$$P\left(T_a(Q) \geq 12\log\frac{s(t)}{\delta}\right) = P\left(\hat{q}_a(Q) \geq 2q_a\right) \leq \delta,$$

where $\hat{q}_a(Q) = \frac{T_a(Q)}{Q}$.

Hence with probability at least $1-\delta$, now we have $t \geq \frac{6}{q_a}\log(s(t)/\delta)$. Also, since $\hat{P}(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1) = T_{a,\boldsymbol{z},l}(t)/t$, by Chernoff bound, when $t \geq \frac{6}{q_a}\log(s(t)/\delta)$, with probability $1-\exp\{-\frac{P(\boldsymbol{Z}_i=z_i,X_i=x_i,i\leq l-1)\cdot t}{3}\} \geq 1-\delta$, we have

$$\hat{P}(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1) \leq 2P(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1).$$

Now by Lemma 8 and Lemma 9, with probability $1-\delta/(k\cdot I_a)$, since

$$\begin{aligned} P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1) &\geq \frac{q_a}{P(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1)} \\ &\geq \frac{q_a}{2\hat{P}(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1)} \\ &\geq \frac{q_a t}{2T_{a,\boldsymbol{z},l}(t)} \\ &\geq \frac{3}{T_{a,\boldsymbol{z},l}(t)}\log\frac{20kI_a|\boldsymbol{A}|\log(2t)}{\delta}. \end{aligned}$$

By Lemma 8 we have

$$\begin{aligned} &|p_{a,\boldsymbol{z},l}(t) - P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)| \\ \leq\ & 2\sqrt{\frac{3P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)}{T_{a,\boldsymbol{z},l}(t)}\log\frac{20kI_a|\boldsymbol{A}|\log(2t)}{\delta}} \\ \leq\ & 2\sqrt{\frac{3P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)}{2^{k-l+|\boldsymbol{Z}_l|+1}T_{a,\boldsymbol{z}}(t)}\log\frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}. \end{aligned}$$

Thus by union bound, with probability $1-\delta$, we have

$$\begin{aligned} &\sum_{\boldsymbol{z}_l}(p_{a,\boldsymbol{z},l}(t) - P(\boldsymbol{Z}_l = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)) \\ \leq & \sum_{\boldsymbol{z}:p_{a,\boldsymbol{z},l}(t)\geq P(\boldsymbol{Z}_l=\boldsymbol{z}_l|\boldsymbol{Z}_i=\boldsymbol{z}_i,X_i=x_i,i\leq l-1)}(p_{a,\boldsymbol{z},l}(t) - P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)) \\ = & \frac{1}{2}\sum_{\boldsymbol{z}}|p_{a,\boldsymbol{z},l}(t) - P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)| \qquad (38) \\ \leq & \sum_{\boldsymbol{z}}\sqrt{\frac{3P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)}{2^{k-l+|\boldsymbol{Z}_l|+1}T_{a,\boldsymbol{z}}(t)}\log\frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}. \end{aligned}$$

The equation (38) is because $\sum_{\boldsymbol{z}_k} p_{a,\boldsymbol{z},k}(t) = \sum_{\boldsymbol{z}} P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1) = 1$. Now we denote

$$\hat{P}_{a,\boldsymbol{z},l}(t) = p_{a,\boldsymbol{z},1}(t)\cdots p_{a,\boldsymbol{z},k}(t),$$

$$P_{a,\boldsymbol{z},l} = P(\boldsymbol{Z}_1 = \boldsymbol{z}_1) \cdots P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le l-1).$$

Hence we get

$$\begin{aligned}
&\hat{\mu}_{obs,a} \qquad (39)\\
&= \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k}(t)\\
&\le \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1)\\
&\quad + \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \hat{P}_{a,\boldsymbol{z},k-1}(t) \sqrt{\frac{3P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1)}{2^{|\boldsymbol{Z}_k|+1} T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}\\
&\le \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1)\\
&\quad + \sum_{\boldsymbol{z}} \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot \sqrt{\frac{3P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1))}{2^{|\boldsymbol{Z}_k|+1} T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}\\
&\le \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1)\\
&\quad + \sum_{\boldsymbol{z}_k} \cdot \sqrt{\frac{3P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1))}{2^{|\boldsymbol{Z}_k|+1} T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}\\
&\le \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1)\\
&\quad + \sqrt{\frac{3 \cdot 2^{|\boldsymbol{Z}_k|}}{2 \cdot 2^{|\boldsymbol{Z}_k|} T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}} \qquad \text{(Cauchy-Schwarz Inequality)}\\
&\le \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \le k-1) \qquad (40)\\
&\quad + \sqrt{\frac{3}{2 \cdot T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}\\
&\le \cdots\\
&\le \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) P_{a,\boldsymbol{z},k} + \sum_{i=1}^{k} \sqrt{\frac{3}{2^i T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}\\
&\le \mu_a + \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\frac{3}{T_{a,\boldsymbol{z}}(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}} + \sqrt{\frac{4}{T_{a,\boldsymbol{z}}(t)} \log \frac{\log(2t)}{\delta}}.\\
&\le \mu_a + 8\sqrt{\frac{1}{T_a(t)} \log \frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}}.
\end{aligned}$$

The above inequality holds for probability $1 - 3\delta$. □

Thus by union bound, $\Pr\{F_2\} \le 3\delta$. In later proof, we will always assume that F_1, F_2 and F_3 don't happen. In this case, true mean $\mu_a \in [L_a^t, U_a^t]$ for all rounds t. Denote $T_1 = 2048H^{(G)} \log(20k|\boldsymbol{A}|H^{(G)} \log(2T_1)/\delta)$, then when $t \ge T_1$, for all arm a such that $q_a \ge \frac{1}{H^{(G)} \cdot \max\{\Delta_a, \varepsilon/2\}^2}$, note that $I_a \le \frac{1}{q_a} \le H^{(G)}$, we have

$$q_a \ge \frac{1}{H_{m^{(G)}_{\varepsilon,\Delta}} \cdot \max\{\Delta_a, \varepsilon/2\}^2} \ge \frac{3}{t} \log \frac{6|\boldsymbol{A}|I_a}{\delta}.$$

Since F_1 doesn't happen, by (35), $|\hat{q}_a - q_a| \le \sqrt{\frac{6q_a}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}}$ and

$$\sqrt{\frac{6q_a}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}} \le \frac{q_a}{2},$$

we have $\hat{q}_a \ge q_a - \sqrt{\frac{6q_a}{t}\log\frac{6|\boldsymbol{A}|I_a}{\delta}} \ge \frac{1}{2H^{(G)}\cdot\max\{\Delta_a,\varepsilon/2\}^2}$.

Hence

$$T_a(t) = \hat{q}_a \cdot t \ge \frac{1024}{\max\{\Delta_a,\varepsilon/2\}^2}\log\frac{20kZ_aI_a|\boldsymbol{A}|\log(2t)}{\delta}.$$

Thus

$$\begin{aligned}\beta_O(T_a(t)) &= \sqrt{\frac{64}{T_a(t)}\log\frac{20kZ_aI_a|\boldsymbol{A}|\log(2T_a(t))}{\delta}}\\ &\le \frac{\max\{\Delta_a,\varepsilon/2\}}{4},\end{aligned}$$

and by Lemma 10, we know the estimation lies in the confidence interval. Now we prove the main theorem. The following lemma provides the upper bound of sample complexity

Lemma 11. *With probability 1-5δ, the algorithm 3 takes at most $\lceil T\rceil$ rounds such that $T \ge 2112H^{(G)}\log\frac{20H^{(G)}|\boldsymbol{A}|\log(2t)}{\delta}$.*

Proof. In the proof we assume F_1, F_2 and F_3 don't happen. The probability for these events are $1-5\delta$. Assume when $t = \lceil T\rceil$, the algorithm don't terminate at t rounds.

Then since $f(x) = \frac{x}{\log(20k|\boldsymbol{A}|H^{(G)}\log(2t)/\delta)}$ is a increasing function, $t \ge 2048H^{(G)}\log\frac{20H^{(G)}k|\boldsymbol{A}|\log(2t)}{\delta}$ for any $t \in [T_1, T]$. Then from above, for arm a such that $q_a \ge \frac{1}{H^{(G)}\max\{\Delta_a,\varepsilon/2\}^2}$, we have $\beta_a(t) \le \beta_O(T_a(t)) \le \frac{\max\{\Delta_a,\varepsilon/2\}}{4}$. Then by Lemma 6, at each round at least one intervention will be performed on some arm a with $\beta_I(N_a(t)) \ge \beta_a(t) \ge \frac{\max\{\Delta_a,\varepsilon/2\}}{4}$, which implies that $N_a(t) \le \frac{64}{\max\{\Delta_a,\varepsilon/2\}^2}\log\frac{H^{(G)}\log(2t)}{\delta}$. Since these arms are M, we have $|M| \le m_{\varepsilon,\Delta}$ and

$$\begin{aligned}T - T_1 &\le \sum_{a\in S}\frac{64}{\max\{\Delta_a,\varepsilon/2\}^2}\log\left(\frac{20H^{(G)}k|\boldsymbol{A}|\log(2T)}{\delta}\right)\\ &\le 64H^{(G)}\log\left(\frac{20H^{(G)}k|\boldsymbol{A}|\log(2T)}{\delta}\right).\end{aligned}$$

Hence

$$T \le T_1 + 64H^{(G)}\log\left(\frac{20H^{(G)}k|\boldsymbol{A}|\log(2T)}{\delta}\right) \le 2112H^{(G)}\log\left(\frac{20H^{(G)}k|\boldsymbol{A}|\log(2T)}{\delta}\right),$$

which completes the proof of Lemma. 11. □

Lemma 12. *Suppose $T = NQ\log(\frac{20k|\boldsymbol{A}|Q\log(2T)}{\delta})$, then $T = O(Q\log(\frac{|\boldsymbol{A}|Q}{\delta}))$.*

Proof. Similar to Lemma 7, for $f(x) = x - NQ\log(\frac{20k|\boldsymbol{A}|Q\log(2T)}{\delta})$we only need to show that there exists a constant C such that $f(CQ\log\frac{Q|\boldsymbol{A}|}{\delta}) \ge f(T) = 0$.

We have

$$\begin{aligned}f(CQ\log\frac{Q|\boldsymbol{A}|}{\delta}) &= CQ\log\frac{Q|\boldsymbol{A}|}{\delta} - NQ\log(\frac{20k|\boldsymbol{A}|Q\log(2CQ\log\frac{Q|\boldsymbol{A}|}{\delta})}{\delta})\\ &\le CQ\log\frac{Q|\boldsymbol{A}|}{\delta} - 2NQ\log(\frac{20|\boldsymbol{A}|Q\log(2CQ\log\frac{Q|\boldsymbol{A}|}{\delta})}{\delta})\end{aligned}$$

$$
\begin{aligned}
&= (C-2N)Q\log\frac{Q|\boldsymbol{A}|}{\delta} - \log(\log 40CQ\log\frac{Q|\boldsymbol{A}|}{\delta})) \\
&\geq (C-2N)Q\log\frac{Q|\boldsymbol{A}|}{\delta} - \log(\log 40C)\cdot\log(Q\log\frac{Q|\boldsymbol{A}|}{\delta}) \\
&\geq (C-2N-\log\log 40C)Q\log\frac{Q|\boldsymbol{A}|}{\delta}.
\end{aligned}
$$

Thus we choose C such that $C-2N-\log\log 40C \geq 0$, then we complete the Lemma 12. □

By the Lemma 12 above, with probability $1-5\delta$, we have

$$
T = O\left(H^{(G)}\log\left(\frac{|\boldsymbol{A}|H^{(G)}}{\delta}\right)\right).
$$

The correctness has been proved in Section C.1, so we complete the proof of Theorem 2. □

C.3 Proof of Theorem 3

Proof. We consider a bandit instance ξ with $\boldsymbol{q}$ and probability distribution $P(X_1, X_2, \cdots, X_n, Y)$. Recall $\min_{\boldsymbol{x}\in\{0,1\}^n} P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}) = p_{min}, \max_{\boldsymbol{x}\in\{0,1\}^n} P(Y=1 \mid \boldsymbol{X}=\boldsymbol{x}) = p_{max}$ and $p_{max} + 2\Delta_{2n+1} + 2\varepsilon \leq 1$. For arm $a \in \boldsymbol{A}$ with $q_a \leq \frac{1}{H_{m_{\varepsilon,\Delta}-1}\cdot\max\{\Delta_a,\varepsilon/2\}^2}$, we denote the set of these arms are M. By definition of $m_{\varepsilon,\Delta}$, we know $|M| \geq m_{\varepsilon,\Delta}$. Then for $a = do(X_i = x) \neq \operatorname{argmin}_{a'\in M}\Delta'_a$ (if optimal arm $a^* \in M, a \neq a^*$), we construct bandit instance ξ'_a with probability distribution

$$
P'(Y \mid X_1,\cdots,X_n) = \begin{cases} P(Y \mid X_1,\cdots,X_n) & X_i \neq x \\ P(Y \mid X_1,\cdots,X_n) + 2(\Delta_a+\varepsilon) & X_i = x \end{cases}
$$

Thus for arm a with $q_a \leq \frac{1}{H^{(P)}_{m_{\varepsilon,\Delta}}\cdot\max\{\Delta_a,\varepsilon/2\}^2}$. Denote $a_{min} = \operatorname{argmin}_{a'\in S}\max\{\Delta_{a'},\varepsilon/2\}$, (We break the tie arbitrarily),for $a \neq a_{min}$, $q_a \leq \frac{1}{2}$.

$$
\begin{aligned}
P'(Y \mid do(X_i = x)) &= P'(Y \mid X_i = x) \\
&= \sum_{\boldsymbol{x}_{-i}} P'(Y \mid X_i = x, \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i})P'(\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) \\
&= \sum_{\boldsymbol{x}_{-i}} (P(Y \mid X_i = x, \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) + 2(\Delta_a+\varepsilon))P'(\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) \\
&= \sum_{\boldsymbol{x}_{-i}} (P(Y \mid X_i = x, \boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) + 2(\Delta_a+\varepsilon))P(\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) \\
&= P(Y \mid X_i = x) + 2(\Delta_a+\varepsilon) \\
&= P(Y \mid do(X_i = x)) + 2(\Delta_a+\varepsilon),
\end{aligned}
$$

Now we consider other arms $a' = do(X_j = x') \in \boldsymbol{A}$, we have

$$
\begin{aligned}
P'(Y \mid do(X_j = x')) &= P'(Y \mid X_j = x') \\
&= \sum_{\boldsymbol{x}_{-j}} P'(Y \mid X_j = x', \boldsymbol{X}_{-j} = \boldsymbol{x}_{-j})P'(\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}) \\
&= \sum_{\boldsymbol{x}_{-j}} P(Y \mid X_j = x', \boldsymbol{X}_{-j} = \boldsymbol{x}_{-j})P(\boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}) \\
&\quad + 2(\Delta_a+\varepsilon)\sum_{\boldsymbol{x}_{-j,-i}} P(\boldsymbol{X}_{-j,-i} = \boldsymbol{x}_{-j,-i}, X_i = x) \\
&= P(Y \mid X_j = x') + 2(\Delta_a+\varepsilon)\cdot P(X_i = x)\cdot\sum_{\boldsymbol{x}_{-j,-i}} P(\boldsymbol{X}_{-j,-i} = \boldsymbol{x}_{-j,-i}) \\
&= P(Y \mid X_j = x') + 2(\Delta_a+\varepsilon)\cdot q_a\cdot\sum_{\boldsymbol{x}_{-j,-i}} P(\boldsymbol{X}_{-j,-i} = \boldsymbol{x}_{-j,-i})
\end{aligned}
$$

$$\leq P(Y \mid X_j = x') + (\Delta_a + \varepsilon).$$

Also, if $a' = do()$, we have

$$\begin{aligned}
P'(Y \mid a') = P'(Y) &= \sum_{\boldsymbol{x}} P'(Y \mid \boldsymbol{X} = \boldsymbol{x}) P'(\boldsymbol{X} = \boldsymbol{x}) \\
&= \sum_{\boldsymbol{x}} P(Y \mid \boldsymbol{X} = \boldsymbol{x}) P(\boldsymbol{X} = \boldsymbol{x}) + 2(\Delta_a + \varepsilon) \cdot \sum_{\boldsymbol{x}_{-i}} P(\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}, X_i = x) \\
&= P(Y) + 2(\Delta_a + \varepsilon) \cdot P(X_i = x) \cdot \sum_{\boldsymbol{x}_{-i}} P(\boldsymbol{X}_{-i} = \boldsymbol{x}_{-i}) \\
&\leq P(Y) + (\Delta_a + \varepsilon).
\end{aligned}$$

Thus for all $a' \in A$, we have

$$\begin{aligned}
P'(Y \mid a') &\leq P(Y \mid a') + \Delta_a + \varepsilon \\
&\leq (P(Y \mid a) + \Delta_a + \varepsilon) + \Delta_a + \varepsilon \\
&= P(Y \mid a) + 2(\Delta_a + \varepsilon) \\
&= P'(Y \mid a),
\end{aligned}$$

which means that a is the best arm in bandit environment ξ'_a. Then denote the probability measure for ξ'_a and ξ as $\Pr_a$ and $\Pr$. Denote Y_t and $\boldsymbol{X}_t$ as the reward and observed value at time t. Define stopping time for the algorithm σ with respect to $\mathcal{F}_t$, Then from Lemma 19 in Kaufmann et al. (2016b), for any event $\zeta \in \mathcal{F}_\sigma$

$$\mathbb{E}_\xi \left[\sum_{t=1}^{\sigma} \log \left(\frac{\Pr(Y_t)}{\Pr_a(Y_t)} \right) \right] = d(\Pr(\zeta), \Pr_a(\zeta)),$$

where $d(x, y) = x \log(x/y) + (1-x) \log((1-x)/(1-y))$ is the binary relatively entropy.

Denote the output of our algorithm is a^o. Then since $a \neq a^*$, when we choose $\zeta = \{a^o = a\}$, we have $\Pr(\zeta) \leq \delta, \Pr_a(\zeta) \geq 1 - \delta$ and $d(\Pr(\zeta), \Pr_a(\zeta)) \geq \log(\frac{1}{2.4\delta})$

Now note that

$$\mathbb{E}_\xi \left[\sum_{t=1}^{\sigma} \log \left(\frac{\Pr(Y_t)}{\Pr_a(Y_t)} \right) \right] \tag{41}$$

$$= \sum_{t=1}^{\sigma} \mathbb{E}_\xi \left[\log \left(\frac{\Pr(Y_t)}{\Pr_a(Y_t)} \right) \right]$$

$$= \sum_{t=1}^{\sigma} \Pr((X_t)_i = x) \left(\sum_{y \in \{0,1\}} \Pr(Y_t = y \mid (X_t)_i = x) \log \frac{\Pr(Y_t = y \mid (X_t)_i = x)}{\Pr_a(Y_t = y \mid (X_t)_i = x)} \right) \tag{42}$$

Denote $B = \Pr(Y_t = y \mid (X_t)_i = x) \in [p_{min}, p_{max}]$

$$\begin{aligned}
&\sum_{y \in \{0,1\}} \Pr(Y_t = y \mid (X_t)_i = x) \log \frac{\Pr(Y_t = y \mid (X_t)_i = x)}{\Pr_a(Y_t = y \mid (X_t)_i = x)} \\
&= B \log \frac{B}{B + 2(\Delta_a + \varepsilon)} + (1 - B) \frac{1 - B}{1 - B - 2(\Delta_a + \varepsilon)} \\
&\leq \frac{-2B(\Delta_a + \varepsilon)}{B + 2\Delta_a} + \frac{2(1 - B)(\Delta_a + \varepsilon)}{1 - B - 2(\Delta_a + \varepsilon)} \\
&\leq \frac{4(\Delta_a + \varepsilon)^2}{(B + 2(\Delta_a + \varepsilon))(1 - B - 2(\Delta_a + \varepsilon))} \\
&\leq \frac{4(\Delta_a + \varepsilon)^2}{0.9 \cdot 0.1},
\end{aligned}$$

Then (42) becomes

$$\log\left(\frac{1}{2.4\delta}\right) \le \mathbb{E}_\xi\left[\sum_{t=1}^{\sigma}\log\left(\frac{\Pr(Y_t)}{\Pr_a(Y_t)}\right)\right] \le \frac{4(\Delta_a+\varepsilon)^2}{0.09}\sum_{t=1}^{\sigma}\Pr((X_t)_i = x)$$
$$\le \frac{16\max\{\Delta_a, \varepsilon/2\}^2}{0.09}\mathbb{E}_\xi[T_a(\sigma)],$$

where $T_a(\sigma)$ for $a = do(X_i = x)$ means the number of times that $X_i = x$. Suppose the sample complexity is T for ξ, denote $N_a(\sigma)$ be the number of times that $A_t = a$. We have

$$\mathbb{E}_\xi[N_a(\sigma) + q_a \cdot \sigma] \ge \mathbb{E}_\xi[T_a(\sigma)] \ge \frac{0.09}{16\max\{\Delta_a, \varepsilon/2\}^2}\log\left(\frac{1}{2.4\delta}\right).$$

By summing over all $a = do(X_i = x) \in M, a \ne a_{min}$, we get

$$0.09\sum_{\substack{a\in M\setminus\{do()\}\\ a\ne a_{min}}}\frac{1}{16\max\{\Delta_a,\varepsilon/2\}^2}\log\left(\frac{1}{2.4\delta}\right) \le \sum_{\substack{a\in M\setminus\{do()\}\\ a\ne a_{min}}}\mathbb{E}_\xi[N_a(\sigma) + q_a\cdot\sigma] \tag{43}$$

$$\le \mathbb{E}_\xi\left[\sigma + \sum_{\substack{a\in M\setminus\{do()\}\\ a\ne a_{min}}} q_a\cdot\sigma\right] \tag{44}$$

$$\le \mathbb{E}_\xi[\sigma]\left(1 + \sum_{\substack{a\in M\setminus\{do()\}\\ a\ne a_{min}}}\frac{1}{\max\{\Delta_a,\varepsilon/2\}^2 H_{m_{\varepsilon,\Delta}-1}}\right). \tag{45}$$

Denote

$$Q = \sum_{\substack{a\in M\setminus\{do()\}\\ a\ne a_{min}}}\frac{1}{\max\{\Delta_a,\varepsilon/2\}^2} \ge H_{m_{\varepsilon,\Delta}-1} - \min_{i<m_{\varepsilon,\Delta}}\frac{1}{\max\{\Delta_{a_i},\varepsilon/2\}^2} - \frac{1}{\max\{\Delta_{do()},\varepsilon/2\}^2},$$

then

$$\mathbb{E}_\xi[\sigma] \ge \frac{0.09Q\log\left(\frac{1}{2.4\delta}\right)}{1 + Q/H_{m_{\varepsilon,\Delta}-1}}$$
$$\ge \frac{0.09}{2}\left(H_{m_{\varepsilon,\Delta}-1} - \min_{i<m_{\varepsilon,\Delta}}\frac{1}{\max\{\Delta_{a_i},\varepsilon/2\}^2} - \frac{1}{\max\{\Delta_{do(),\varepsilon/2}\}^2}\right)\log\left(\frac{1}{2.4\delta}\right).$$

□

D Some Proofs of Lemma

D.1 Proof of Lemma 1

Proof. By definition, we only need to show that $|\{a \mid q_a \cdot \max\{\Delta_a, \varepsilon/2\}^2 < 1/H_{2m}\}| \le 2m$. Assume it does not hold, then $q_{a_i}\cdot\max\{\Delta_{a_i},\varepsilon/2\}^2 < \frac{1}{H_{2m}}$ for $i = 1, 2, \cdots, 2m+1$. Then for $i \ge m+1, m+2, \cdots, 2m+1$, we have

$$q_{a_i} < \frac{1}{H_{2m}\cdot\max\{\Delta_{a_i},\varepsilon/2\}^2} < \frac{1}{\sum_{j=1}^{m}\frac{1}{\max\{\Delta_{a_j},\varepsilon/2\}^2}\cdot\max\{\Delta_{a_i},\varepsilon/2\}^2} \le \frac{1}{m}.$$

The inequality above implies the $|\{a \mid q_a < \frac{1}{m}\}| \ge m+1$, which leads to a contradiction for the definition of m. □

D.2 THE EXISTENCE FOR ADMISSIBLE SEQUENCE IN GRAPHS WITHOUT HIDDEN VARIABLES

In graph without hidden variables, the admissible-sequence is important for identifying the causal effect. Now we provide an algorithm to show how to find the admissible-sequence in this condition.

Theorem 4. *For causal graph $G = (\boldsymbol{X} \cup \{Y\}, E)$ without hidden variables, for a set $\boldsymbol{S} = \{X_1, \cdots, X_k\}$ and $X_1 \succeq X_2 \succeq \cdots \succeq X_k$, the admissible-sequence with respect to $\boldsymbol{S}$ and Y can be found by*

$$\boldsymbol{Z}_i = \boldsymbol{Pa}(X_i) \setminus (\boldsymbol{Z}_1 \cup \cdots \cup \boldsymbol{Z}_{i-1} \cup X_1 \cup \cdots \cup X_{i-1}).$$

Proof. The proof is straightforward. First, $\boldsymbol{Z}_i \subseteq \boldsymbol{Pa}(X_i)$ consists of nondescendants of $\{X_i, X_{i+1}, \cdots, X_k\}$ by topological order. Second, we need to prove

$$(Y \perp\!\!\!\perp X_i \mid X_1, \cdots, X_{i-1}, \boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_i)_{G_{\underline{X}_i, \overline{X}_{i+1}, \cdots, \overline{X}_k}}. \tag{46}$$

We know that the $\boldsymbol{Pa}(X_i) \subseteq X_1 \cup X_2 \cdots X_{i-1} \cup \boldsymbol{Z}_1 \cup \cdots \boldsymbol{Z}_i$. Then it blocks all the backdoor path from X_i to Y. Also, since $X_1 \cup X_2 \cdots X_{i-1} \cup \boldsymbol{Z}_1 \cup \boldsymbol{Z}_2 \cdots \boldsymbol{Z}_i$ consists of nondescendants of X_i, it cannot block any forward path from X_i to Y. Also, for any forward path with colliders, namely, $X_i \to \cdots \to X' \leftarrow X'' \cdots Y$, the X' cannot be conditioned since it is a descendant for X_i. So conditioning on $X_1 \cup X_2 \cdots X_{i-1} \cup \boldsymbol{Z}_1 \cup \boldsymbol{Z}_2 \cdots \boldsymbol{Z}_i$ will not active any extra forward path. Hence, there is only original forward path from X_i to Y, which means that (46) holds. □

D.3 PROOF OF LEMMA 2

Lemma 2. *For an action $a = do(\boldsymbol{S} = \boldsymbol{s})$ and any two weight vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, we have*

$$|\sigma(\boldsymbol{\theta}, a) - \sigma(\boldsymbol{\theta}', a)| \leq \mathbb{E}_{\boldsymbol{e}} \left[\sum_{X \in N_{\boldsymbol{S},Y}} |\boldsymbol{V}_X^\top (\boldsymbol{\theta}_X - \boldsymbol{\theta}'_X)| M^{(1)} \right], \tag{6}$$

where $N_{\boldsymbol{S},Y}$ is the set of all nodes that lie on all possible paths from X_1 to Y excluding $\boldsymbol{S}$, $\boldsymbol{V}_X$ is the value vector of a sample of the parents of X according to parameter $\boldsymbol{\theta}$, $M^{(1)}$ is defined in Assumption 1, and the expectation is taken on the randomness of the noise term $\boldsymbol{e} = (e_X)_{X \in \boldsymbol{X} \cup \{Y\}}$ of causal model under parameter $\boldsymbol{\theta}$.

Proof. Note that our BGLM model is equivalent to a threshold model: For each node X, we randomly sample a threshold $\Gamma_X \in [0, 1]$, and if $f_X(\theta_X^T \boldsymbol{Pa}(X)) + e_X \geq \Gamma_X$, we let $X = 1$, which means it is activated. At timestep 1, the X_1 is activated, then at timestep $i \geq 2$, X_i is either activated (set it to 1) or deactivated (set it to 0). Then, the BGLM is equivalent to the propagating process above if we uniformly sample Γ_X for each node X, i.e. $\Gamma_X \sim \mathcal{U}[0, 1]$. Now we only need to show

$$|\sigma(\boldsymbol{\theta}, a) - \sigma(\boldsymbol{\theta}', a)| \leq \mathbb{E}_{\boldsymbol{e},\Gamma} \left[\sum_{X \in N_{\boldsymbol{S},Y}} |\boldsymbol{V}_X^\top (\boldsymbol{\theta}_X - \boldsymbol{\theta}'_X)| M^{(1)} \right], \tag{47}$$

Firstly, we have

$$|\sigma(\boldsymbol{\theta}, a) - \sigma(\boldsymbol{\theta}', a)| = \mathbb{E}_{\boldsymbol{e},\Gamma} \left[\mathbb{I}\{Y \text{is activated on } \boldsymbol{\theta} \neq \mathbb{I}\{Y \text{is activated on } \boldsymbol{\theta}'\}\} \right],$$

and we define the event $\mathcal{E}_0^{\boldsymbol{e}}(X)$ as

$$\mathcal{E}_0^{\boldsymbol{e}}(X) = \left\{ \Gamma | \mathbb{I}\{X \text{ is activated under } \Gamma, \boldsymbol{e}, \boldsymbol{\theta}\} \neq \mathbb{I}\{X \text{ is activated under } \Gamma, \boldsymbol{e}, \boldsymbol{\theta}'\} \right\}.$$

Hence

$$\left|\sigma(\boldsymbol{\theta}, a) - \sigma(\boldsymbol{\theta}', a)\right| \leq \mathbb{E}_{\boldsymbol{e}} \left[\Pr_{\Gamma \sim (\mathcal{U}[0,1])^n} \{\mathcal{E}_0^{\boldsymbol{e}}(Y)\} \right].$$

Now since only nodes in $N_{\boldsymbol{S},Y}$ will influence Y, we can only consider node X in $N_{\boldsymbol{S},Y}$.

Let $\phi^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma) = (\phi_0^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma) \subseteq S, \phi_1^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma), \cdots, \phi_n^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma))$ be the sequence of activated sets on $\boldsymbol{\theta}$, 0-mean noise $\boldsymbol{e}$ and threshold factor Γ. More specifically, $\phi_i(\boldsymbol{\theta},\Gamma)$ is the set of nodes activated by time step i. For every node $X \in N_{\boldsymbol{S},Y}$, we define the event that X is the first node that has different activation under $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ as below:

$$\mathcal{E}_1^{\boldsymbol{e}}(X) = \{\Gamma | \exists i \in [n], \forall i' < i, \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma) = \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma), X \in (\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)\backslash\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma) \cup (\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma)\backslash\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)))\}.$$

Then we have $\mathcal{E}_0^{\boldsymbol{e}}(Y) \subseteq \cup_{X \in N_{\boldsymbol{S},Y}} \mathcal{E}_1^{\boldsymbol{e}}(X)$. We also define other events:

$$\begin{aligned}
&\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i) = \{\Gamma | \forall i' < i, \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma) = \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma), X \notin \phi_{i-1}^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)\},\\
&\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i) = \{\Gamma | \forall i' < i, \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma) = \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma), X \in \phi_i^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)\backslash\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma)\},\\
&\mathcal{E}_{2,2}^{\boldsymbol{e}}(X,i) = \{\Gamma | \forall i' < i, \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma) = \phi_{i'}^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma), X \in \phi_i^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma)\backslash\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)\},\\
&\mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i) = \{\Gamma | X \in \phi_i^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)\backslash\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma)\}\\
&\mathcal{E}_{3,2}^{\boldsymbol{e}}(X,i) = \{\Gamma | X \in \phi_i^{\boldsymbol{e}}(\boldsymbol{\theta}',\Gamma)\backslash\phi_i^{\boldsymbol{e}}(\boldsymbol{\theta},\Gamma)\}.
\end{aligned}$$

Then since $\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i)$ and $\mathcal{E}_{2,2}^{\boldsymbol{e}}(X,i)$ are exclusive, we have

$$\Pr_{\Gamma}\{\mathcal{E}_1^{\boldsymbol{e}}(X)\} = \sum_{i=1}^{n} \Pr_{\Gamma}\{\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i)\} + \sum_{i=1}^{n} \Pr_{\Gamma}\{\mathcal{E}_{2,2}^{\boldsymbol{e}}(X,i)\}.$$

Now we need to bound the two terms above. First, consider $\Pr_{\Gamma}\{\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i)\}$, we set Γ_{-X} is the vector with all value $\Gamma_{X'}$ of node $X' \neq X$, then we also define the corresponding sub-event $\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) \subset \mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i)$ as the event with value Γ_{-X}. Define $\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) \subset \mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i)$, $\mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) \subset \mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i)$, $\mathcal{E}_{3,2}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) \subset \mathcal{E}_{3,2}^{\boldsymbol{e}}(X,i)$ in a similar way.

From definition, $\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) = \mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) \cup \mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})$, then we have

$$\Pr_{\Gamma}\{\mathcal{E}_{2,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X})\} = \Pr_{\Gamma}\{\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i)\} \cdot \Pr_{\Gamma}\{\mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) | \mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})\}.$$

Thus, by the definition of BGLM, in $\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})$, the value of Γ_X must lie in an interval with highest value 1. Denote it as $[\mathcal{W}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X}), 1]$, then

$$\Pr_{\Gamma_X \sim \mathcal{U}[0,1]} \mathcal{E}_{2,0}^{\boldsymbol{e}}(X) = 1 - \mathcal{W}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X}).$$

Now we consider

$$\Pr_{\Gamma_X}\{\mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) | \mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})\}.$$

We first assume $\mathcal{W}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) < 1$, otherwise our statement holds trivially. Then we denote that the nodes activated at timestep t under $\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})$ as $\phi_t^{\boldsymbol{e}}(\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X}))$. If the conditional event above holds, we have

$$f_X\left(\sum_{X' \in \phi_{i-1}^{\boldsymbol{e}}(\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})) \cap N(X)} \boldsymbol{\theta}_{X',X}\right) + e_X < \Gamma_X \leq f_X\left(\sum_{X' \in \phi_{i-1}^{\boldsymbol{e}}(\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})) \cap N(X)} \boldsymbol{\theta}'_{X',X}\right) + e_X,$$

or

$$f_X\left(\sum_{X' \in \phi_{i-1}^{\boldsymbol{e}}(\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})) \cap N(X)} \boldsymbol{\theta}_{X',X}\right) + e_X \geq \Gamma_X > f_X\left(\sum_{X' \in \phi_{i-1}^{\boldsymbol{e}}(\mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})) \cap N(X)} \boldsymbol{\theta}'_{X',X}\right) + e_X,$$

where $\theta_{X',X}$ is the element corresponding to X' in θ_X.

Thus,

$$\Pr_{\Gamma_X \sim \mathcal{U}[0,1]}\{\mathcal{E}_{3,1}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) \cup \mathcal{E}_{3,2}^{\boldsymbol{e}}(X,i,\Gamma_{-X}) | \mathcal{E}_{2,0}^{\boldsymbol{e}}(X,i,\Gamma_{-X})\}$$

$$= \frac{|f_X\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}_{X',X}\right) - f_X\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}'_{X',X}\right)|}{1-\mathcal{W}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X})}$$

Thus we have

$$\begin{aligned}
&\Pr_{\Gamma}\{\mathcal{E}^{\boldsymbol{e}}_{2,1}(X,i,\Gamma_{-X})\cup\mathcal{E}^{\boldsymbol{e}}_{2,2}(X,i,\Gamma_{-X})\}\\
&=\Pr_{\Gamma}\{\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i)\}\cdot\Pr_{\Gamma}\{\mathcal{E}^{\boldsymbol{e}}_{3,1}(X,i,\Gamma_{-X})\cup\mathcal{E}^{\boldsymbol{e}}_{3,2}(X,i,\Gamma_{-X})|\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X})\}\\
&=\left|f_X\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}_{X',X}\right)-f_X\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}'_{X',X}\right)\right|\\
&\le M^{(1)}\left|\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}_{X',X}\right)-\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}'_{X',X}\right)\right|.
\end{aligned}$$

When $\mathcal{E}^{\boldsymbol{e}}_{2,0}(X)=\emptyset$, both two sides are zero, so it holds in general.

Now we define $\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i,\Gamma_{-X})=\{\Gamma\mid\Gamma=(\Gamma_X,\Gamma_{-X})\mid X\notin\phi^{\boldsymbol{e}}_{i-1}(\boldsymbol{\theta},\Gamma)\}$, then $\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X})\subseteq\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i,\Gamma_{-X})$. In addition, when $\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X})\neq\emptyset$, $\phi^{\boldsymbol{e}}_{i'}(\mathcal{E}^{\boldsymbol{e}}_{2,0}(X,i,\Gamma_{-X}))=\phi^{\boldsymbol{e}}_{i'}(\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i,\Gamma_{-X}))$ for all $i'<i$. Thus we have

$$\begin{aligned}
&\Pr_{\Gamma}\{\mathcal{E}^{\boldsymbol{e}}_{2,1}(X,i,\Gamma_{-X})\cup\mathcal{E}^{\boldsymbol{e}}_{2,2}(X,i,\Gamma_{-X})\}\\
&\le M^{(1)}\left|\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}_{X',X}\right)-\left(\sum_{X'\in\phi^{\boldsymbol{e}}_{i-1}(\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i,\Gamma_{-X}))\cap N(X)}\boldsymbol{\theta}'_{X',X}\right)\right|.
\end{aligned}$$

Now we can get

$$\begin{aligned}
\Pr_{\Gamma}\{\mathcal{E}^{\boldsymbol{e}}_{1}(X)\} &= \int_{\Gamma_{-X}}\sum_{i=1}^{n}\Pr_{\gamma_X\sim\mathcal{U}[0,1]}\{\mathcal{E}^{\boldsymbol{e}}_{2,1}(X,i,\Gamma_{-X})\cup\mathcal{E}^{\boldsymbol{e}}_{2,2}(X,i,\Gamma_{-X})\}\mathrm{d}\Gamma_{-X}\\
&=\int_{\Gamma_{-X}}\Pr_{\gamma_X\sim\mathcal{U}[0,1]}\{\mathcal{E}^{\boldsymbol{e}}_{2,1}(X,i^*,\Gamma_{-X})\cup\mathcal{E}^{\boldsymbol{e}}_{2,2}(X,i^*,\Gamma_{-X})\}\mathrm{d}\Gamma_{-X}\\
&\le\int_{\Gamma_{-X}}\left|\sum_{X'\in\phi^{\boldsymbol{e}}_{i^*-1}(\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i^*,\Gamma_{-X}))\cap N(X)}(\theta_{X',X}-\theta'_{X',X})\right|M^{(1)}\mathrm{d}\Gamma_{-X}\\
&=\mathbb{E}_{\Gamma_{-X}}\left[\left|\sum_{X'\in\phi^{\boldsymbol{e}}_{i^*-1}(\mathcal{E}^{\boldsymbol{e}}_{4,0}(X,i^*,\Gamma_{-X}))\cap N(X)}(\theta_{X',X}-\theta'_{X',X})\right|\right]M^{(1)}\\
&=\mathbb{E}_{\Gamma_{-X}}\left[|\boldsymbol{V}_X(\theta_X-\theta'_X)|M^{(1)}\right],
\end{aligned}$$

where i^* is the topological order of X in graph G, and the second inequality is because $\mathcal{E}^{\boldsymbol{e}}_{2,1}(X,i,\Gamma_{-X})\neq\emptyset$ only when $i=i^*$. Summing over all node $X\in N_{\boldsymbol{S},Y}$, we complete the proof.

□

D.4 Proof of Lemma 3

Lemma 3. *For one node $X\in\boldsymbol{X}\cup\{Y\}$, assume Assumption 1 and 2 holds, and*

$$\lambda_{min}(M_{t,X})\ge\frac{512D(M^{(2)})^2}{\kappa^4}(D^2+\ln\frac{3nt^2}{\delta}),$$

with probability $1 - \delta/nt^2$*, for any vector* $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$*, at all rounds* t *the estimator* $\hat{\boldsymbol{\theta}}_{t,X}$ *in Algorithm 2 satisfy*

$$|\boldsymbol{v}^\top(\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*_X)| \leq \frac{3}{\kappa}\sqrt{\log(3nt^2/\delta)}||v||_{M^{-1}_{t,X}}.$$

Proof. The all proof is very similar to the proof in Feng & Chen (2022), for the completeness, we provide them here. Note that $\hat{\boldsymbol{\theta}}_{t,X}$ satisfies $\nabla L_{t,X}(\hat{\boldsymbol{\theta}}_{t,X}) = 0$, where

$$\nabla L_{t,X}(\boldsymbol{\theta}_X) = \sum_{i=1}^{t}[X^t - f_X(\boldsymbol{V}^T_{i,X}\boldsymbol{\theta}_X)]\boldsymbol{V}_{i,X}.$$

Define $G(\boldsymbol{\theta}_X) = \sum_{i=1}^{t}(f_X(\boldsymbol{V}^T_{i,X}\boldsymbol{\theta}_X) - f_X(\boldsymbol{V}^T_{i,X}\boldsymbol{\theta}^*_X))\boldsymbol{V}_{i,X}$. Thus $G(\boldsymbol{\theta}^*_X) = 0$ and $G(\hat{\boldsymbol{\theta}}_{t,X}) = \sum_{i=1}^{t}\varepsilon'_{i,X}\boldsymbol{V}_{i,X}$, where $\varepsilon'_{i,X} = X^i - f_X(\boldsymbol{V}^T_{i,X}\boldsymbol{\theta}^*_X)$. Now note that $\mathbb{E}[\varepsilon'_{i,X}|\boldsymbol{V}_{i,X}] = 0$ and $\varepsilon'_{i,X} = X^i - f_X(\boldsymbol{V}^T_{i,X}\boldsymbol{\theta}^*_X) \in [-1, 1]$, then $\varepsilon'_{i,X}$ is 1-subgaussian. Let $Z = G(\hat{\boldsymbol{\theta}}_{t,X}) = \sum_{i=1}^{t}\varepsilon'_{i,X}\boldsymbol{V}_{i,X}$

Step 1: Consistency of $\hat{\boldsymbol{\theta}}_{t,X}$ For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$, $\exists\bar{\boldsymbol{\theta}} = s\boldsymbol{\theta}_1 + (1-s)\boldsymbol{\theta}_2, 0 < s < 1$ such that

$$\begin{aligned} G(\boldsymbol{\theta}_1) - G(\boldsymbol{\theta}_2) &= \left[\sum_{i=1}^{t}\dot{f}_X(\boldsymbol{V}^T_{i,X}\bar{\boldsymbol{\theta}})\boldsymbol{V}_{i,X}\boldsymbol{V}^T_{i,X}\right](\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\ &\triangleq F(\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2). \end{aligned}$$

Since f is strictly increasing, $\dot{f} > 0$, then $G(\boldsymbol{\theta})$ is an injection and G^{-1} is well-defined.

Now let $\mathcal{B}_\eta = \{\boldsymbol{\theta} \mid ||\boldsymbol{\theta} - \boldsymbol{\theta}^*|| \leq \eta\}$, then define $\kappa_\eta := \inf_{\boldsymbol{\theta}\in\mathcal{B}_\eta, X\neq 0}\dot{f}(X^T\boldsymbol{\theta}) > 0$. The following lemma helps our proof, and it can be found in Lemma A of Yin & Zhao (2005):

Lemma 13 (Yin & Zhao (2005))**.** $\{\boldsymbol{\theta} \mid ||G(\boldsymbol{\theta})||_{M^{-1}_{t,X}} \leq \kappa_\eta\eta\sqrt{\lambda_{min}(M_{t,X})}\} \subseteq \mathcal{B}_\eta$.

The next lemma provides an upper bound of $||Z||_{M^{-1}_{t,X}}$:

Lemma 14 (Zhang et al. (2022))**.** *For any* $\delta > 0$*, the event* $\mathcal{E}_G := \{||Z||_{M^{-1}_{t,X}} \leq 4\sqrt{|\boldsymbol{Pa}(X)| + \ln(1/\delta)}\}$ *holds with probability at least* $1 - \delta$*.*

By the above two lemmas, when $\mathcal{E}_G$ holds, for any $\eta \geq \frac{4}{\kappa_\eta}\sqrt{\frac{|\boldsymbol{Pa}(X)|+\ln(1/\delta)}{\lambda_{\min}(M_{t,X})}}$, we have $||\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*|| \leq \eta$. Choose $\eta = 1$, we know $1 \geq \frac{4}{\kappa}\sqrt{\frac{|\boldsymbol{Pa}(X)|+\ln(1/\delta)}{\lambda_{\min}(M_{t,X})}}$, then with probability $1 - \delta$ $||\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*|| \leq 1$.

Step 2: Normality of $\hat{\boldsymbol{\theta}}_{t,X}$. Now we assume $||\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*|| \leq 1$ holds. Define $\Delta = \hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}^*$, then $\exists s \in [0, 1]$ such that $Z = G(\hat{\boldsymbol{\theta}}_{t,X}) - G(\boldsymbol{\theta}^*_X) = (H + E)\Delta$, where $\bar{\boldsymbol{\theta}} = s\boldsymbol{\theta}^*_X + (1-s)\hat{\boldsymbol{\theta}}_{t,X}$, $H = F(\boldsymbol{\theta}^*_X) = \sum_{i=1}^{t}\dot{f}_X(\boldsymbol{V}^T_{i,X}\boldsymbol{\theta}^*_X)\boldsymbol{V}_{i,X}\boldsymbol{V}^T_{i,X}$ and $E = F(\bar{\boldsymbol{\theta}}) - F(\boldsymbol{\theta}^*_X)$. Then, according to mean value theorem, we have

$$\begin{aligned} E &= \sum_{i=1}^{t}(\dot{f}_X(\boldsymbol{V}_{i,X}\cdot\bar{\boldsymbol{\theta}}) - \dot{f}_X(\boldsymbol{V}_{i,X}\cdot\boldsymbol{\theta}^*_X))\boldsymbol{V}_{i,X}\boldsymbol{V}^T_{i,X} \\ &= \sum_{i=1}^{t}\ddot{f}_X(r_i)\boldsymbol{V}^T_{i,X}\Delta\boldsymbol{V}_{i,X}\boldsymbol{V}^T_{i,X} \\ &\leq \sum_{i=1}^{t}M^{(2)}\boldsymbol{V}^T_{i,X}\Delta\boldsymbol{V}_{i,X}\boldsymbol{V}^T_{i,X} \end{aligned}$$

for some $r_i \in \mathbb{R}$. Thus we have

$$\begin{aligned}
\boldsymbol{v}^T H^{-1/2} E H^{-1/2} \boldsymbol{v} &\leq \sum_{i=1}^{t} L_{f_X}^{(2)} ||\boldsymbol{V}_{i,X}|| ||\Delta|| ||\boldsymbol{v}^T H^{-1/2} \boldsymbol{V}_{i,X}||^2 \\
&\leq M^{(2)} \sqrt{|\boldsymbol{Pa}(X)|} ||\Delta|| (\boldsymbol{v}^T H^{-1/2} (\sum_{i=1}^{t} \boldsymbol{V}_{i,X} \boldsymbol{V}_{i,X}^T) H^{-1/2} \boldsymbol{v}) \\
&\leq \frac{M^{(2)} \sqrt{|\boldsymbol{Pa}(X)|}}{\kappa} ||\Delta|| ||\boldsymbol{v}||^2,
\end{aligned}$$

hence we know

$$\begin{aligned}
||H^{-1/2} E H^{-1/2}|| &\leq \frac{M^{(2)} \sqrt{|\boldsymbol{Pa}(X)|}}{\kappa} ||\Delta|| \\
&\leq \frac{4M^{(2)} \sqrt{|\boldsymbol{Pa}(X)|}}{\kappa^2} \sqrt{\frac{|\boldsymbol{Pa}(X)| + \ln \frac{1}{\delta}}{\lambda_{\min}(M_{t,X})}} \\
&\leq \frac{1}{2},
\end{aligned}$$

where the last inequality is because

$$\lambda_{\min}(M_{t,X}) \geq 512 \frac{(M^{(2)})^2}{\kappa^4} |\boldsymbol{Pa}(X)| \left(|\boldsymbol{Pa}(X)| + \ln \frac{1}{\delta} \right) > 64 \frac{(M^{(2)})^2}{\kappa^4} |\boldsymbol{Pa}(X)| \left(|\boldsymbol{Pa}(X)| + \ln \frac{1}{\delta} \right).$$

Now for any $\boldsymbol{v} \in \mathbb{R}^{|\boldsymbol{Pa}(X)|}$, we have

$$\begin{aligned}
\boldsymbol{v}^T (\hat{\boldsymbol{\theta}}_{t,X} - \boldsymbol{\theta}_X^*) &= \boldsymbol{v}^T (H + E)^{-1} Z \\
&= \boldsymbol{v}^T H^{-1} Z - \boldsymbol{v}^T H^{-1} E (H + E)^{-1} Z.
\end{aligned}$$

The second equality is correct from $H + E = F(\bar{\boldsymbol{\theta}}) \succeq \kappa M_{t,X} \succeq 0$.

Define $D \triangleq (\boldsymbol{V}_{1,X}, \boldsymbol{V}_{2,X}, \cdots, \boldsymbol{V}_{t,X})^T \in \mathbb{R}^{t \times |\boldsymbol{Pa}(X)|}$. Then $D^T D = \sum_{i=1}^{t} \boldsymbol{V}_{i,X} \boldsymbol{V}_{i,X}^T = M_{t,X}$. By the Hoeffding's inequality Hoeffding (1994),

$$\begin{aligned}
\Pr(\left|\boldsymbol{v}^T H^{-1} Z \geq a\right|) &\leq \exp\left(-\frac{a^2}{2||\boldsymbol{v}^T H^{-1} D^T||^2} \right) \\
&= \exp\left(-\frac{a^2}{2\boldsymbol{v}^T H^{-1} D^T D H^{-1} \boldsymbol{v}} \right) \\
&\leq \exp\left(-\frac{a^2 \kappa^2}{2||\boldsymbol{v}||_{M_{t,X}^{-1}}^2} \right).
\end{aligned}$$

The last inequality holds because $H \succeq \kappa M_{t,X} = \kappa D^T D$.

Thus with probability $1 - 2\delta$, $|\boldsymbol{v}^T H^{-1} Z| \leq \frac{\sqrt{2 \ln 1/\delta}}{\kappa} ||\boldsymbol{v}||_{M_{t,X}^{-1}}$.

For the second term, we know

$$\begin{aligned}
|\boldsymbol{v}^T H^{-1} E (H + E)^{-1} Z| &\leq ||\boldsymbol{v}||_{H^{-1}} ||H^{-\frac{1}{2}} E (H + E)^{-1} Z|| \\
&\leq ||\boldsymbol{v}||_{H^{-1}} ||H^{-\frac{1}{2}} E (H + E)^{-1} H^{\frac{1}{2}}|| \, ||Z||_{H^{-1}} \\
&\leq \frac{1}{\kappa} ||\boldsymbol{v}||_{M_{t,X}^{-1}} ||H^{-\frac{1}{2}} E (H + E)^{-1} H^{\frac{1}{2}}|| \, ||Z||_{M_{t,X}^{-1}}. \qquad (48)
\end{aligned}$$

Then we get

$$\begin{aligned}
||H^{-\frac{1}{2}} E (H + E)^{-1} H^{\frac{1}{2}}|| &= ||H^{-\frac{1}{2}} E (H^{-1} - H^{-1} E (H + E)^{-1})^{-1} H^{\frac{1}{2}}|| \\
&= ||H^{-\frac{1}{2}} E H^{\frac{1}{2}} + H^{-\frac{1}{2}} E H^{-1} E (H + E)^{-1} H^{\frac{1}{2}}||
\end{aligned}$$

$$\leq ||H^{-\frac{1}{2}}EH^{\frac{1}{2}}|| + ||H^{-\frac{1}{2}}EH^{-\frac{1}{2}}||\, ||H^{-\frac{1}{2}}E(H+E)^{-1}H^{\frac{1}{2}}||,$$

where the first inequality is derived by $(H+E)^{-1} = H^{-1} - H^{-1}E(H+E)^{-1}$.

Then we can get

$$||H^{-\frac{1}{2}}E(H+E)^{-1}H^{\frac{1}{2}}|| \leq \frac{||H^{-\frac{1}{2}}EH^{\frac{1}{2}}||}{1-||H^{-\frac{1}{2}}EH^{\frac{1}{2}}||} \leq 2||H^{-\frac{1}{2}}EH^{\frac{1}{2}}|| \leq \frac{8M^{(2)}\sqrt{|\boldsymbol{Pa}(X)|}}{\kappa^2}\sqrt{\frac{|\boldsymbol{Pa}(X)|+\ln\frac{1}{\delta}}{\lambda_{\min}(M_{t,X})}}$$

Thus by (48) and Lemma 14

$$\left|v^T H^{-1}E(H+E)^{-1}Z\right| \leq \frac{32L_{f_X}^{(2)}\sqrt{|\boldsymbol{Pa}(X)|}(|\boldsymbol{Pa}(X)|+\log\frac{1}{\delta})}{\kappa^3\sqrt{\lambda_{\min}(M_{t,X})}}||\boldsymbol{v}||_{M_{t,X}^{-1}}.$$

So we have

$$\left|\boldsymbol{v}^T(\hat{\boldsymbol{\theta}}_{t,X}-\boldsymbol{\theta}_X^*)\right| \leq \left(\frac{32L_{f_X}^{(2)}\sqrt{|\boldsymbol{Pa}(X)|}(|\boldsymbol{Pa}(X)|+\log\frac{1}{\delta})}{\kappa^3\sqrt{\lambda_{\min}(M_{t,X})}} + \frac{\sqrt{2\ln 1/\delta}}{\kappa}\right)||\boldsymbol{v}||_{M_{t,X}^{-1}} \leq \frac{3}{\kappa}\sqrt{\log(1/\delta)}||\boldsymbol{v}||_{M_{t,X}^{-1}},$$

where the last inequality is because

$$\lambda_{\min}(M_{t,X}) \geq \frac{512|\boldsymbol{Pa}(X)|(L_{f_X}^{(2)})^2}{\kappa^4}\left(|\boldsymbol{Pa}(X)|^2+\ln\frac{1}{\delta}\right).$$

By replace δ with $\delta/3nt^2$, we complete the proof. □

E EXPERIMENTS

In this section, we provide some experiments supporting our theoretical result for CCPE-BGLM and CCPE-General.

E.1 CCPE-BGLM

Experiment 1 First, we provide the experiments for our CCPE-BGLM algorithm. We construct a causal graph with 9 nodes $X_1, \cdots, X_8$ and X_0, such that $X_i \succeq X_{i+1}$. Then, we randomly choose two nodes in $X_1, \cdots, X_{i-1}$ and also X_0 to be the parent of $X_i (i \geq 1)$. Y has 4 parents, and they are randomly chosen in $\boldsymbol{X} = \{X_1 \cdots, X_8\}$. For X_0, we know $P(X_0 = 1) = 1$. For node X_i and their parent $X_i^{(1)}, X_i^{(2)}$, $P(X_i = 1) = 0.4X_0 + 0.1X_i^{(1)} + 0.1X_i^{(2)}$. (If $i = 2$, $P(X_i = 1) = 0.4X_0 + 0.1X_i^{(1)} = 0.4X_0 + 0.1X_1$; If $i = 1$, $P(X_1 = 1) = 0.4X_0$.) Suppose the parents of reward variable are $X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}$ The reward variable is defined by $P(Y = 1) = 0.3X^{(1)} + 0.3X^{(2)} + 0.3X^{(3)} + 0.05X^{(4)}$. The action set is $\{do(\boldsymbol{S} = \boldsymbol{1} \mid |\boldsymbol{S}| = 3, \boldsymbol{S} \subset \boldsymbol{X})\}$. Hence the optimal arm is $do(\{X^{(1)}, X^{(2)}, X^{(3)}\} = \boldsymbol{1})$.

We choose 4 algorithms in this experiment: LUCB in Kalyanakrishnan et al. (2012), lilUCB-heuristic in Jamieson et al. (2013), Propagating-Inference in Yabe et al. (2018) and our CCPE-BGLM. LUCB and lilUCB-heuristic are classical pure exploration algorithm. Because in previous causal bandit literature, Propagating-Inference is the only algorithm considering combinatorial action set without prior knowledge $P(\boldsymbol{Pa}(Y) \mid a)$ for action $a \in \boldsymbol{A}$, we choose it in this experiment. Note that the criteria of Propagating-Inference algorithm is simple regret, hence it cannot directly compare to our pure exploration algorithm. We choose to compare the error probability at some fixed time T instead. In this criteria, Propagating-Inference algorithm will have an extra knowledge of budget T while LUCB, lilUCB-heuristic and CCPE-BGLM not. To implement the Propagating Inference algorithm, we follows the modification in Yabe et al. (2018) to make this algorithm more efficient and accurate by setting $\lambda = 0$ and $\eta_A = 1/C$. (Defined and stated in Yabe et al. (2018).) For CCPE-BGLM, we ignore the condition that $t \geq \max\{\frac{cD}{\eta^2}\log\frac{nt^2}{\delta}, \frac{1024(M^{(2)})^2(4D^2-3)D}{\kappa^4\eta}(D^2+\ln\frac{3nt^2}{\delta})\}$, to make it more efficient. During our experiment, the error probability is smaller than other algorithm even if we ignore this condition. Also, to make this algorithm more efficient, we update observational confidence bound (Line 11) each 50 rounds. (This will not influence the proof of Theorem 1.) For LUCB, lilUCB-heuristic and CCPE-BGLM, we find the best exploration parameter α, α_I and α_O by grid search from $\{0.05, 0.1, \cdots, 1\}$. (Exploration parameter α for UCB-type algorithm is a constant multiplied in front of the confidence radius, which should be tuned in practice. e.g.(Li et al. (2010), Mason et al. (2020).) For this task, we find $\alpha = 0.3, \alpha_O = 0.05, \alpha_I = 0.4$. We choose $T = 50 + 50i$ for $0 \leq i \leq 9$. For each time T, we run 100 iterations and average the result.

As the Figure 4 shows, even if our algorithm does not know the budget T, our algorithm converges quicker than all other algorithms.

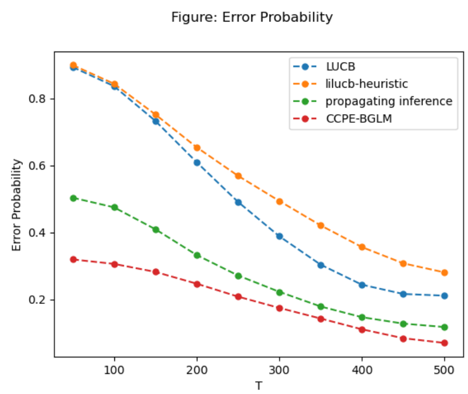


Figure 4: Error Probability for Experiment 1

E.2 CCPE-GENERAL

In this subsection, we provide the experiments for CCPE-General algorithm. We also choose 4 algorithms, LUCB, lilUCB-heuristic, Propagating-Inference and CCPE-General (called "adm_seq" in figure because it utilizes admissible sequence.). Since the Propagating-Inference cannot hold for general graph with hidden variables, we first compare them in graphs without hidden variables. Then we also compare LUCB, lilUCB-heuristic and our algorithm in the graphs with hidden variables.

Experiment 2 we construct the graph with 7 nodes $X_1, \cdots, X_7$ such that $X_i \succeq X_{i+1}$. Then, we randomly choose two nodes in $X_1, \cdots, X_{i-1}$ as parents of X_i. The reward variable Y has 5 parents $X^{(i)} = X_{i+2}, 1 \leq i \leq 5$. We choose $P(X_1 = 1) = 0.5$, $P(X_2 = 1) = 0.55$ if $X_0 = 1$ and otherwise $P(X_2 = 1) = 0.45$. For $i \geq 2$, and two parents $X_i^{(1)}, X_i^{(2)}$ of X_i, $P(X_i = 1) = 0.55$ if $X_i^{(1)} = X_i^{(2)}$ and otherwise $P(X_i = 1) = 0.45$. For reward variable Y,

$$P(Y=1) = \begin{cases} 0.9 & X_i^{(1)} = X_i^{(2)} = 1 \\ 0.7 + 0.05X_i^{(1)} + 0.05X_i^{(2)} & X_i^{(3)} = X_i^{(4)} = 1 \\ 0 & \text{Otherwise} \end{cases} \tag{49}$$

We define the action set $\{do(\boldsymbol{S} = \boldsymbol{s}) \mid |\boldsymbol{S}| = 2, \boldsymbol{s} \in \{0,1\}^2\}$, then the optimal arm is $do(\{X_i^{(1)}, X_i^{(2)}\} = \mathbf{1})$. We choose $\alpha_O = 0.25, \alpha_I = 0.4$, and exploration parameters α for LUCB and lilUCB are both 0.3. For each time $T = 150 + 50i$ for $0 \leq i \leq 9$, we run 100 times and average the result to get the error probability. The result is shown in Figure 5. We note that our algorithm performs almost the same as Propagating-Inference algorithm. Our CCPE-General algorithm is a fixed confidence algorithm without requirement for budget T, and our algorithm can be applied to causal graphs with hidden variables.

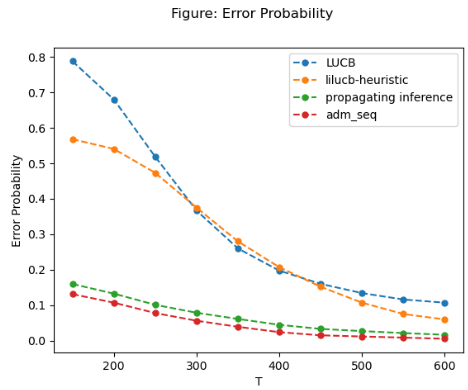
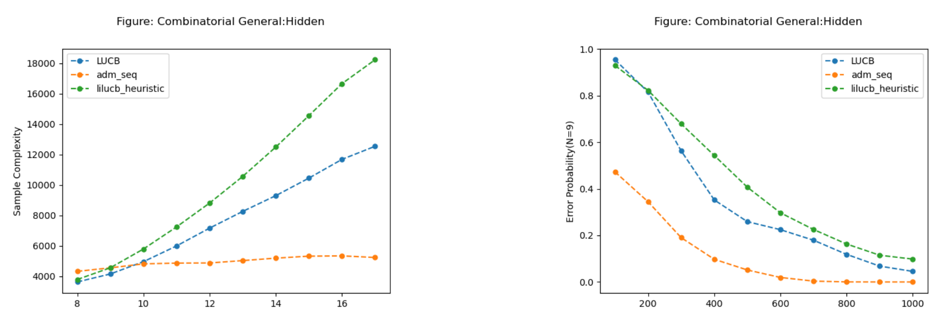


Figure 5: Error Probabiltiy for Experiment 2

Experiment 3 In this paragraph, we provide an experiment to show that CCPE-General algorithm can be applied to broader causal graphs with hidden variables. Since there is no previous algorithm working on both combinatorial action set and existence of hidden variable, we compare our result with LUCB and lilUCB-heuristic.

Our causal graph are constructed as follows: $\boldsymbol{X} = X_0, X_1, \cdots, X_{n+1}$, where $X_i \to Y$ and $X_1 \to X_i$ for $2 \le i \le n+1$, $X_1 \to X_0$. $\boldsymbol{U} = \{U_1, \cdots, U_n\}$ and $U_i \to X_{i+2}, U_i \to X_0$ for $2 \le i \le n+1$. For action set $\{do(\boldsymbol{S} = \boldsymbol{s}) \mid |\boldsymbol{S}| = 2, \boldsymbol{S} \subset \{X_2, \cdots, X_{n+1}\}\}$. Each U_i satisfies $P(U_i = 1) = 0.5$. $P(X_0 = 1) = \min\{\frac{1}{n}\sum_{i=0}^{n-1} U_i + 0.1X_1, 1\},$. $P(X_1 = 1) = 0.5$. $P(X_i = 1) = 0.5$ if $X_1 = U_{i-2} = 1$ and otherwise 0.4. For the reward variable Y, $P(Y = 1) = 0.4X_2 + 0.4X_3 + \frac{0.2}{n}\sum_{i=4}^{n+2} X_i$.

For this task, by grid search, we set $\alpha = 0.25$ for exploration parameter of LUCB and lilUCB, and $\alpha_O = 0.3, \alpha_I = 0.4$ for CCPE-General algorithm (In the figures below, we call our algorithm ”adm_seq” since it uses admissible sequence.) We compare the error probability and sample complexity for them. The results are shown in Figure 6(a) and Figure 6(b). Our CCPE-General algorithm wins in both metrics.



(a) Sample Complexity for Experiment 3

(b) Error Probability for Experiment 3

Figure 6: Experiment 3

F Fixed Budget Causal Bandit Algorithm

In this section, we provide a preliminary fixed budget causal bandit algorithm, which based on successive reject algorithm and our previous analysis for causal bandit. The previous causal bandit algorithm in fixed budget always directly estimate the observation threshold m. However, to derive a gap-dependent result, this method does not work. Our Causal Successive Reject avoids the estimation for observation threshold and get a better gap-dependent result. Note that $T_a(t) = \min_{\boldsymbol{z}} T_{a,\boldsymbol{z}}(t)$ for $\boldsymbol{z} \in \{0,1\}^{|Z_a|}$, then we can get the simple causal successive reject algorithm as follows:

Algorithm 4 Causal Successive Reject

```
Input: Causal Graph G, action set A, budget T, parameter ε.
Initialize t = 1, T_a = 0, μ̂_a = 0 for all arms a ∈ A. Define |A| = N. A_0 = A
Perform do() for T/2 times, and update T_a(t) for all a.
Set n_k = (T/2 − N) / (log N (N+1−k)) for k = 1, 2, ··· , n − 1.
for each phase k = 1, 2, ··· , n − 1: do
    for i = 1, 2, ··· , ⌈(N + 1 − k)(n_k − n_{k−1})⌉ do
        Perform intervention a for action with least T_a + N_a, and N_a = N_a + 1.
    end for
    Denote a_k = argmin_{a∈A_{k−1}} μ̂_a, where μ̂_a follows the same definition of Algorithm 2. A_k =
    A_{k−1}\{a_k}.
    if |A_k| = 1 then
        return A_k.
    end if
end for
```

Theorem 5. *The algorithm 4 will return the ε-optimal arm within error probability*

$$P(\mu_{a^o} < \mu_{a^*} - \varepsilon) \leq 4I_a N^2 \exp\left\{-\frac{\frac{T}{2} - N}{128\overline{\log} N H_3}\right\},$$

where $H_3 = \max_{k=1,2,\cdots,N}\{\alpha_k^{-1}(\max\{\Delta^{(k)}, \varepsilon\})^{-2}\}$, and α_k is defined by

$$\alpha_k = \begin{cases} \dfrac{1 + \sum_{i=k+1}^{N} \frac{1}{i}}{m} & \text{if } k > m \\ \dfrac{1}{k} + \dfrac{\sum_{i=m+1}^{N} \frac{1}{i}}{m} & \text{if } k \leq m \end{cases}$$

where m is defined in 1 with respect to q_a similar to Theorem 2

To show that our algorithm outperforms the classical successive reject and sequential halving algorithm, it is obvious that $H_3 \leq H_2$, where $H_2 = \max_{k=1,2,\cdots,N}\{k \max\{\Delta^{(k)}, \varepsilon\}^{-2}\}$, since $\alpha_k \geq \frac{1}{k}$.

Proof. We also denote $T_a(t), N_a(t)$ as the value of T_a and N_a at the round t. The main idea is that: Each stage we spend half budget to observe, and spend the remaining budget to supplement the arms which are not observed enough. The main idea of proof is to show that in each stage, each arm in A_k has $T_a(t) + D_a(t) \geq m_k$ times, which leads to a brand-new result.

Denote the set of arm $S = \{a \in \boldsymbol{A} \mid q_a < 1/m\}$, then $|S| \leq m$. First, for $a \notin S$, by chernoff bound, it has been observed by $\hat{q}_a \cdot T/2 \geq \frac{T}{4m}$ with probability $1 - \delta$, where $\delta = 6N \cdot \exp\{-\frac{T}{24m}\}$. Hence $T_a(T/2) \geq \frac{T}{4m}$ for $a \notin S$.

First we prove the following lemma:

Lemma 15. *After stage $1 \leq k \leq N - m$, all the arms in A_k must have $N_a(t) + T_a(t) \geq m_k$, where $m_k = \sum_{i=1}^{k} \frac{(N+1-i)(n_i - n_{i-1})}{2m} \geq \lceil \frac{\frac{T}{2} - N}{2\overline{\log} N \cdot m}(1 + \sum_{i=N+2-k}^{N} \frac{1}{i})\rceil$.*

Proof. Let $D_a(t) = T_a(t) + N_a(t)$. Denote $m_0 = n_0 = 0$, then For $a \notin S$, $T_a(t) \geq \frac{T}{4m}$. Thus number of arm a with $T_a(t) \leq \frac{T}{4m}$ is less than m. Then the intervention in stage k will only performed on $D_a(t) \leq \frac{T}{4m}$ unless all the arms have $D_a(t) \geq \frac{T}{4m} \geq m_k$ times.

If all arms have $D_a(t) \geq \frac{T}{4m} \geq m_k$ times, the lemma holds. If it is not true, the $(N + 1 - k)(n_k - n_{k-1})$ interventions will performed on at most m arms. Hence all the arms must have $N_a(t) \geq m_{k-1} + \frac{(N+1-k)(n_k - n_{k-1})}{m} = m_{k-1} + 2(m_k - m_{k-1}) \geq m_k$ times after stage $k \leq N - m$.

Then $m_k = \sum_{i=1}^{k} \frac{(N+1-i)(n_i - n_{i-1})}{2m} \geq \sum_{i=1}^{k-1} \frac{n_i}{2m} + \frac{(N+1-k)n_k}{2m} \geq \frac{\frac{T}{2}-N}{\log N \cdot 2m}(1 + \sum_{i=N+2-k}^{N} \frac{1}{i})$. □

Lemma 16. *After stage $k > N - m$, all the arms in A_k have $T_a(t) + N_a(t) \geq m'_k$, where $m'_k = \frac{\frac{T}{2}-N}{2\log N} \cdot \alpha_{N+1-k}$.*

Proof. For $a \in A_k$, in stage $k > N - m$, $|A_k| = N - k + 1 \leq m$. All the arms in A_k must have $T_a(t) + N_a(t) \geq m'_{k-1} + (N + 1 - k)(n_k - n_{k-1})/(N + 1 - k) = m'_{k-1} + n_k - n_{k-1}$ times, where $m'_{N-m} = m_{N-m}$.

Thus after stage $k > N - m$, all the arms in A_k must have

$$
\begin{aligned}
T_a(t) + N_a(t) &\geq m_{N-m} + \sum_{N-m<i\leq k} (n_i - n_{i-1}) \\
&= m_{N-m} + (n_k - n_{N-m}) \\
&\geq \frac{\frac{T}{2}-N}{\log N \cdot m}(1 + \sum_{i=m+2}^{N} \frac{1}{i}) + 1 + (\frac{\frac{T}{2}-N}{\log N}(\frac{1}{N+1-k} - \frac{1}{m+1})) - 1 \\
&= \frac{\frac{T}{2}-N}{\log N \cdot m}(\sum_{i=m+1}^{N} \frac{1}{i}) + (\frac{\frac{T}{2}-N}{\log N}(\frac{1}{N+1-k})) \\
&= \frac{\frac{T}{2}-N}{\log N}(\frac{1}{N+1-k} + \frac{1}{m(\sum_{i=m+1}^{N} \frac{1}{i})^{-1}}) \\
&= \frac{\frac{T}{2}-N}{\log N} \cdot \alpha_{N+1-k} \\
&\geq m'_k.
\end{aligned}
$$

□

Lemma 17. *In round t, with probability $1 - \frac{\delta}{8nt^3}$,*

$$
|\hat{\mu}_{obs,a} - \mu_a| < 4\sqrt{\frac{1}{T_a(t)} \log \frac{4I_a}{\delta}}. \tag{50}
$$

Proof. When $T_a(t) \geq 12 \log \frac{4I_a}{\delta}$, we know this lemma is trivial since $\mu_a, \hat{\mu}_{obs,a} \in [0, 1]$. Otherwise, if $t < \frac{6}{q_a}$, define $Q = \frac{6}{q_a} \log(\frac{4I_a}{\delta})$, based on $T_a(t) \geq 12 \log \frac{4I_a}{\delta}$, then

$$
P\left(t < \frac{6}{q_a} \log(1/\delta)\right) \leq P\left(T_a(Q) \geq 12 \log \frac{4I_a}{\delta}\right).
$$

Thus by Chernoff bound, we know

$$
P\left(T_a(Q) \geq 12 \log \frac{4I_a}{\delta}\right) = P\left(\hat{q}_a(Q) \geq 2q_a\right) \leq \delta,
$$

where $\hat{q}_a(Q) = \frac{T_a(Q)}{Q}$.

Hence with probability at least $1 - \delta$, now we have $t \geq \frac{6}{q_a} \log(4I_a/\delta)$. Also, since $\hat{P}(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1) = T_{a,\boldsymbol{z},l}(t)/t$, by Chernoff bound, when $t \geq \frac{6}{q_a} \log(4I_a/\delta)$, with probability $1 - \exp\{-\frac{P(\boldsymbol{Z}_i=z_i,X_i=x_i,i\leq l-1)\cdot t}{3}\} \geq 1 - \delta$, we have

$$
\hat{P}(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1) \leq 2P(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1).
$$

Now since

$$
\begin{aligned}
P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1) &\geq \frac{q_a}{P(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1)} \\
&\geq \frac{q_a}{2\hat{P}(\boldsymbol{Z}_i = z_i, X_i = x_i, i \leq l-1)} \\
&\geq \frac{q_a t}{2T_{a,\boldsymbol{z},l}(t)} \\
&\geq \frac{3}{T_{a,\boldsymbol{z},l}(t)} \log \frac{4I_a}{\delta}.
\end{aligned}
$$

By Hoeffding's inequality and Chernoff bound, for $a = do(X = x)$,

$$
|r_{a,\boldsymbol{z}}(t) - P(Y = 1 \mid \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{z} = \boldsymbol{z})| \leq \sqrt{\frac{1}{2T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}},
$$

$$
\begin{aligned}
&|p_{a,\boldsymbol{z},l}(t) - P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)| \\
&\quad \leq \sqrt{\frac{3P(\boldsymbol{Z}_l = \boldsymbol{z}_l \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq l-1)}{t} \log \frac{4I_a}{\delta}}
\end{aligned}
$$

at round $2t$ with probability $1 - 2Z\frac{\delta}{2I_a} = 1 - \delta$. Hence by Lemma 9, Eq (38), we get

$$
\begin{aligned}
&\hat{\mu}_{obs,a} \qquad (51)\\
&= \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k}(t) \\
&\leq \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1) \\
&\quad + \frac{1}{2}\sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t)\hat{P}_{a,\boldsymbol{z},k-1}(t)\sqrt{\frac{3P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1)}{2^{|\boldsymbol{Z}_k|+1}T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}} \\
&\leq \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1) \\
&\quad + \frac{1}{2}\sum_{\boldsymbol{z}} \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot \sqrt{\frac{3P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1))}{2^{|\boldsymbol{Z}_k|+1}T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}} \\
&\leq \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1) \\
&\quad + \frac{1}{2}\sum_{\boldsymbol{z}_k} \cdot \sqrt{\frac{3P(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1))}{2^{|\boldsymbol{Z}_k|+1}T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}} \\
&\leq \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1) \\
&\quad + \frac{1}{2}\sqrt{\frac{3 \cdot 2^{|\boldsymbol{Z}_k|}}{2 \cdot 2^{|\boldsymbol{Z}_k|}T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}} \qquad \text{(Cauchy-Schwarz Inequality)} \\
&\leq \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) \cdot \hat{P}_{a,\boldsymbol{z},k-1}(t) \cdot P_t(\boldsymbol{Z}_k = \boldsymbol{z}_k \mid \boldsymbol{Z}_i = \boldsymbol{z}_i, X_i = x_i, i \leq k-1) \qquad (52) \\
&\quad + \frac{1}{2}\sqrt{\frac{3}{2 \cdot T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}} \\
&\leq \cdots \\
&\leq \sum_{\boldsymbol{z}} r_{a,\boldsymbol{z}}(t) P_{a,\boldsymbol{z},k} + \frac{1}{2}\sum_{i=1}^{k}\sqrt{\frac{3}{2^i T_{a,\boldsymbol{z}}(t)} \log \frac{4I_a}{\delta}}
\end{aligned}
$$

$$\leq \mu_a + \frac{1}{2-\sqrt{2}}\sqrt{\frac{3}{T_{a,\boldsymbol{z}}(t)}\log\frac{4I_a}{\delta}} + \sqrt{\frac{1}{2T_{a,\boldsymbol{z}}(t)}\log\frac{2}{\delta}}.$$

$$\leq \mu_a + 4\sqrt{\frac{1}{T_a(t)}\log\frac{4I_a}{\delta}}.$$

□

Now we prove another lemma to bound the error probability of each stage.

Lemma 18. *For an arm $a \in \boldsymbol{A}$, $N_a(t) + T_a(t) = D_a(t)$, then we have*

$$P(|\hat{\mu}_a - \mu_a| > \epsilon) < 4I_a \exp\{-D_a(t)(\epsilon^2/32\}. \tag{53}$$

Proof. We know $N_a(t) \geq \frac{D_a(t)}{2}$ or $T_a(t) \geq \frac{N_a(t)}{2}$. When $N_a(t) \geq \frac{N_a(t)}{2}$, by Hoeffding's inequality, we know that

$$P(|\hat{\mu}_a - \mu_a| > \epsilon) < 2\exp\{-2N_a(t)(\epsilon/2)^2\} < 2\exp\{-D_a(t)(\epsilon/2)^2\}.$$

When $T_a(t) \geq \frac{D_a(t)}{2}$, by Lemma 17, we know

$$P(|\hat{\mu}_a - \mu_a| > \epsilon) < 4I_a \exp\{-N_a(t)\epsilon^2/16\} < 4I_a \exp\{-D_a(t)\epsilon^2/32\}.$$

Then we complete the proof. □

Hence the event that

$$\zeta = \left\{\forall i \in \{1,2,\cdots,N\}, \forall a \in \boldsymbol{A}_i, |\hat{\mu}_a - \mu_a| < \frac{1}{2}\max\{\Delta^{(N+1-i)}, \varepsilon\}\right\}$$

doesn't happen within probability at most

$$\sum_{i=1}^{n}\sum_{a\in\boldsymbol{A}_i} 4I_a \exp\{-m_i(\frac{\max\{(\Delta^{(N+1-i)})^2,\varepsilon\}}{2})^2/32\}$$
$$\leq \sum_{i=1}^{n}\sum_{a\in\boldsymbol{A}_i} 4I_a \exp\left\{-\alpha_{N+1-k}\cdot\max\{(\Delta^{(N+1-i)}),\varepsilon\}^2\frac{\frac{T}{2}-N}{128\overline{\log}N}\right\}$$
$$\leq 4I_a N^2 \exp\left\{-\frac{\frac{T}{2}-N}{128\overline{\log}N H_3}\right\},$$

where $H_3 = \max_{i=1,2,\cdots,n}\{\alpha_i^{-1}(\max\{\Delta^{(i)},\varepsilon\})^{-2}\}$.

Now we prove that under event ζ, the algorithm output a ε-optimal arm.

For each stage k, we prove that one of the following condition will be satisfied:

(1). All arms in A_{k-1} are $\varepsilon-$optimal.

(2). Stage k eliminate an non-optimal arm $a_k \neq a^*$

In fact, assume (1) does not hold, then there exists at least one arm which is not $\varepsilon-$optimal. Since $|A_{k-1}| = N+1-k$, there must exist an arm $a \in \boldsymbol{A}_{k-1}$ with $\mu_{a^*} - \mu_a \geq \max\{\varepsilon, \Delta^{(N+1-k)}\}$. Hence because of event ζ, after stage k, all arms in A_k satisfy $\hat{\mu}_a - \mu_a < \frac{1}{2}\max\{\Delta^{(N+1-k)},\varepsilon\}$. Hence

$$\hat{\mu}_a \leq \mu_a + \frac{1}{2}\max\{\Delta^{(N+1-k)},\varepsilon\}$$
$$\leq \mu_{a^*} - \max\{\Delta^{(N+1-k)},\varepsilon\} + \frac{1}{2}\max\{\Delta^{(N+1-k)},\varepsilon\}$$

$$
\begin{aligned}
&< \hat{\mu}_{a^*} + \max\{\Delta^{(N+1-k)}, \varepsilon\} - \max\{\Delta^{(N+1-k)}, \varepsilon\} \\
&= \hat{\mu}_{a^*}.
\end{aligned}
$$

So the optimal arm a^* will not be eliminated.

Hence if (2) always happen, the remaining arm will be the optimal arm. Otherwise, if (1) happens, the algorithm will return an ε-optimal arm. Hence we complete the proof. □