# Combinatorial Causal Bandits without Graph Skeleton

**Shi Feng*** SHIFENG-THU@OUTLOOK.COM
*Harvard University, MA, USA*

**Nuoya Xiong*** NUOYAX@ANDREW.CMU.EDU
*Carnegie Mellon University, PA, USA*

**Wei Chen** WEIC@MICROSOFT.COM
*Microsoft Research, Beijing, China*

**Editors:** Vu Nguyen and Hsuan-Tien Lin

## Abstract

In combinatorial causal bandits (CCB), the learning agent chooses a subset of variables in each round to intervene and collects feedback from the observed variables to minimize expected regret or sample complexity. Previous works study this problem in both general causal models and binary generalized linear models (BGLMs). However, all of them require prior knowledge of causal graph structure or unrealistic assumptions. This paper studies the CCB problem without the graph structure on binary general causal models and BGLMs. We first provide an exponential lower bound of cumulative regrets for the CCB problem on general causal models. To overcome the exponentially large space of parameters, we then consider the CCB problem on BGLMs. We design a regret minimization algorithm for BGLMs even without the graph skeleton and show that it still achieves $O(\sqrt{T}\ln T)$ expected regret, as long as the causal graph satisfies a weight gap assumption. This asymptotic regret is the same as the state-of-art algorithms relying on the graph structure. Moreover, we propose another algorithm with $O(T^{\frac{2}{3}}\ln T)$ regret to remove the weight gap assumption.

**Keywords:** Causal Bandits; Online Learning; Multi-armed Bandits; Causal Inference

## 1. Introduction

The multi-armed bandits (MAB) problem is a classical model in sequential decision-making (Robbins, 1952; Auer et al., 2002; Bubeck et al., 2012). In each round, the learning agent chooses an arm and observes the reward feedback corresponding to that arm, with the goal of either maximizing the cumulative reward over $T$ rounds (regret minimization), or minimizing the sample complexity to find the arm closest to the optimal one (pure exploration). MAB can be extended to have more structures among arms and reward functions, which leads to more advanced learning techniques. Such structured bandit problems include combinatorial bandits (Chen et al., 2013, 2016), linear bandits (Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013; Li et al., 2017), and sparse linear bandits (Abbasi-Yadkori et al., 2012).

In this paper, we study another structured bandit problem called causal bandits, which is first proposed by Lattimore et al. (2016). It consists of a causal graph $G = (\boldsymbol{X} \cup \{Y\}, E)$ indicating the causal relationship among the observed variables. In each round, the learning agent selects one or a few variables in $\boldsymbol{X}$ to intervene, gains the reward as the output of $Y$,

* Equal Contribution.

and observes the values of all variables in $\boldsymbol{X} \cup \{Y\}$. The use of causal bandits is possible in a variety of contexts that involve causal relationships, including medical drug testing, performance tuning, policy making, scientific experimental process, etc.

In all previous literature except Lu et al. (2021); Konobeev et al. (2023), the structure of the causal graph is known, but the underlying probability distributions governing the causal model are unknown. Lu et al. (2021) further assume that the graph structure is unknown and the learning agent can only see the graph skeleton. Here, graph skeleton is also called essential graph (Gámez et al., 2013) and represents all the edges in $G$ without the directional information. In our paper, we further consider that the graph skeleton is unknown and remove the unrealistic assumption that $Y$ only has a single parent in Lu et al. (2021); Konobeev et al. (2023). In many scenarios, the learning agent needs to learn the causal relationships between variables and thus needs to learn the graph without any prior information. For example, in policymaking for combating COVID-19, many possible factors like food supply, medical resources, vaccine research, public security, and public opinion may consequently impact the mortality rate. However, the causal relationships among these factors are not readily known and need to be clarified during the sequential decision-making process. Learning the causal graph from scratch while identifying the optimal intervention raises a new challenge to the learning problem.

For regret minimization, we study CCB under the BGLMs as in Feng and Chen (2023); Xiong and Chen (2023). Using a novel initialization phase, we could determine the ancestor structure of the causal graph for the BGLM when the minimum edge weight in the model satisfies a weight gap assumption. This is enough to perform a CCB algorithm based on maximum likelihood estimation on it (Feng and Chen, 2023). The resulting algorithm BGLM-OFU-Unknown achieves $O(\sqrt{T}\log T)$ regret, where $T$ is the time horizon. The big $O$ notation only holds for $T$ larger than a threshold so the weight gap assumption is hidden by the asymptotic notation. For binary linear models (BLMs), we can remove the weight gap assumption with the $O(T^{2/3})$ regret. The key idea is to measure the difference in the reward between the estimated graph (may be inaccurate) and the true graph. The algorithms we design for BLMs allow hidden variables and use linear regression instead of MLE to remove an assumption on parameters.

For pure exploration, we give some discussions on general causal models in Appendix H. If we allow the weight gap, a trivial solution exists. Without the weight gap, we give an adaptive algorithm for general causal model in the atomic setting.

In summary, our contribution includes: (a) providing an exponential lower bound of cumulative regret for CCB on general causal model, (b) proposing an $O(\sqrt{T}\ln T)$ cumulative regret CCB algorithm BGLM-OFU-Unknown for BGLMs without graph skeleton (with the weight gap assumption), (c) proposing an $O(T^{\frac{2}{3}}\ln T)$ cumulative regret CCB algorithm BLM-LR-Unknown for BLMs without graph skeleton and the weight gap assumption, (d) conducting a numerical experiment in Appendix G for BGLM-OFU-Unknown and BLM-LR-Unknown and giving intuitions on how to choose between them, (e) giving the first discussion in Appendix H including algorithms and lower bounds on the pure exploration of causal bandits on general causal models and atomic intervention without knowing the graph structure.

## 2. Related Works

In this section, we introduce two related lines of research.

### 2.1. Causal Bandits

The causal bandits problem is first proposed by Lattimore et al. (2016). They discuss the simple regret for parallel graphs and general graphs with known probability distributions $P(\boldsymbol{Pa}(Y)|a)$ for any action $a$. In this context, $\boldsymbol{Pa}(Y)$ represents the parent nodes of $Y$. Sen et al. (2017); Nair et al. (2021); Maiti et al. (2021) generalize the simple regret study for causal bandits to more general causal graphs and soft interventions. Lu et al. (2020); Nair et al. (2021); Maiti et al. (2021) consider cumulative regret for causal bandits problem. However, all of these studies are not designed for combinatorial action set and has exponentially large regret or sample complexity with respect to the graph size if the actions are combinatorial. Yabe et al. (2018); Feng and Chen (2023); Xiong and Chen (2023); Varici et al. (2022) consider combinatorial action set for causal bandits problem. Among them, Feng and Chen (2023) are the first to remove the requirement of $T > \sum_{X\in\boldsymbol{X}} 2^{|\boldsymbol{Pa}(X)|}$ and proposes practical CCB algorithms on BGLMs with $O(\sqrt{T}\ln T)$ regret. Xiong and Chen (2023) simultaneously propose CCB algorithms on BGLMs as well as general causal models with polynomial sample complexity with respect to the graph size. Varici et al. (2022) further include soft interventions in the CCB problem, but their work is on linear structural equation models (SEM). Lee and Bareinboim (2018, 2019, 2020) propose several CCB algorithms on general causal bandits problem, but they focus on empirical studies while we provide theoretical regret analysis. All of the above works require the learning agent to know the graph structure in advance. Lu et al. (2021) are the first to work on causal bandits without graph structure. However, their algorithm is limited to the case of $|\boldsymbol{Pa}(Y)| = 1$ for the atomic setting, and thus the main technical issue degenerates to finding the particular parent of $Y$ so that one could intervene on this node for the optimal reward. Recently, Konobeev et al. (2023) has eliminated the need for prior knowledge of the graph skeleton as required in Lu et al. (2021). However, their approach is still limited to designing bandit algorithms for the atomic setting with $|\boldsymbol{Pa}(Y)| = 1$. Furthermore, their algorithm may experience exponentially large regret when $1/\min_{X\in\boldsymbol{Anc}(Y),x\in\text{supp}(X),y\in\text{supp}(Y)} |P(Y = y|X = x) - P(Y = y)|$ is exponentially large in relation to the graph size. In this context, $\boldsymbol{Anc}(Y)$ denotes the ancestors of $Y$ and supp represents the support of a random variable. Recently, Malek et al. (2023) also studied the causal bandits problem without a graph; however, their objective is different from ours. Instead of minimizing regret, they aim to find a near-optimal intervention in the fewest number of exploration rounds.

### 2.2. Social Network and Causality

Causal models have intrinsic connections with influence propagation in social networks. Feng and Chen (2021) study the identifiability in the Independent Cascade (IC) propagation model as a causal model. The BGLM studied in this paper contains the IC model and linear threshold (LT) model in a DAG as special cases, and is also related to the general threshold model proposed by Kempe et al. (2003). Moreover, Feng and Chen (2023); Xiong and Chen (2023) also study causal bandits on BGLMs to avoid the exponentially large parameter space

of general causal models. These papers borrow some techniques and ideas from influence maximization literature, including Li et al. (2020) and Zhang et al. (2022). However, in our BGLM CCB problem, the graph skeleton is unknown, and we need adaptation and integration of previous techniques together with some new ingredients.

## 3. Model

We utilize capital letters $(U, X, Y \ldots)$ to represent variables and their corresponding lowercase letters to indicate their values, as was frequently done in earlier causal inference literature (see, for example, (Pearl, 2009b,a; Pearl and Mackenzie, 2018)). To express a group or a vector of variables or values, we use boldface characters like $\boldsymbol{X}$ and $\boldsymbol{x}$. For a vector $x \in \mathbb{R}^d$, the weighted $\ell_2$-norm associated with a positive-definite matrix $A$ is defined by $\|x\|_A = \sqrt{x^\intercal A x}$.

**C**ausal Models. A *causal graph* $G = (\boldsymbol{X} \cup \{Y\}, E)$ is a directed acyclic graph consisting of intervenable variables $\boldsymbol{X}$, a special target node $Y$ without outgoing edges, and the set of directed edges $E$ connecting nodes in $\boldsymbol{X} \cup \{Y\}$. Denote $n = |\boldsymbol{X}|$ as the number of nodes in $\boldsymbol{X}$. For simplicity, in this paper we consider all variables in $\boldsymbol{X} \cup \{Y\}$ are $(0, 1)$-binary random variables. In our main text, all the variables in $\boldsymbol{X} \cup \{Y\}$ are known and their values can be observed but the edges in $E$ are unknown and cannot be directly observed. We refer to the in-neighbor nodes of a node $X$ in $G$ as the *parents* of $X$, denoted by $\boldsymbol{Pa}(X)$, and the values of these parent random variables as $\boldsymbol{pa}(X)$. According to the definition of causal Bayesian model (Galles and Pearl, 1995; Pearl, 2009b), the probability distribution $P(X|\boldsymbol{Pa}(X))$ is used to represent the causal relationship between $X$ and its parents for every conceivable value combination of $\boldsymbol{Pa}(X)$. Moreover, we define the ancestors of a node $X \in \boldsymbol{X} \cup \{Y\}$ by $\boldsymbol{Anc}(X)$.

We mainly study the *Markovian* causal graph $G$ in this paper, which means that there are no hidden variables in $G$ and every observed variable $X$ has some randomness that is not brought on by any other variables.[1] In this study, we dedicate random variable $X_1$ to be a special variable that always takes the value 1 and is a parent of all other observed random variables in order to model the self-activation effect of the Markovian model. In essence, this represents the initial probability for each node, ensuring that even when all parent nodes of a node except $X_1$ are set to 0, the given node still possesses a probability of being 1.

In this paper, we study a special causal model called binary generalized linear model (BGLM). Specifically, in BGLM, we have $P(X = 1|\boldsymbol{Pa}(X) = \boldsymbol{pa}(X)) = f_X(\boldsymbol{\theta}^*_X \cdot \boldsymbol{pa}(X)) + \varepsilon_X$, where $f_X$ is a monotone increasing function, $\boldsymbol{\theta}^*_X$ is an unknown weight vector in $[0, 1]^{|\boldsymbol{Pa}(X)|}$, and $\varepsilon_X$ is a zero-mean sub-Gaussian noise that ensures the probability does not exceed 1 or equivalently, $\varepsilon_X \le 1 - \max_{\boldsymbol{pa}(X) \in \{0,1\}^{|\boldsymbol{Pa}(X)|}} f_X(\boldsymbol{pa}(X) \cdot \boldsymbol{\theta}^*_X)$. The bounded epsilon follows the convention established by GLM (Li et al., 2020) and provides randomness for linear models in our paper. We use the notation $\theta^*_{X',X}$ to denote the entry in the vector $\boldsymbol{\theta}^*_X$ that corresponds to node $X' \in \boldsymbol{Pa}(X)$, $\boldsymbol{\theta}^*$ to denote the vector of all the weights, and $\Theta$ to denote the feasible domain for the weights. We also use the notation $\boldsymbol{\varepsilon}$ to represent all noise random variables $(\varepsilon_X)_{X \in \boldsymbol{X} \cup Y}$.

1. In Section 6 we mention that our algorithm for BLM can also work for models with hidden variables.

We also study binary linear model (BLM) and linear model in this paper. In BLMs, all $f_X$'s are identity functions, so $P(X = 1|\boldsymbol{Pa}(X) = \boldsymbol{pa}(X)) = \boldsymbol{\theta}^*_X \cdot \boldsymbol{pa}(X) + \varepsilon_X$. When we remove the noise variable $\varepsilon_X$, BLM coincides with the *linear threshold (LT)* model for influence cascades (Kempe et al., 2003) in a DAG. In linear models, we remove the randomness of conditional probabilities, so $X = \boldsymbol{\theta}^*_X \cdot \boldsymbol{pa}(X) + \varepsilon_X$.

For the unknown causal graph, there is an important parameter $\theta^*_{\min} = \min_{(X',X)\in E} \theta^*_{X',X}$, which represents the minimum weight gap for all edges. Intuitively, this minimum gap measures the difficulty for the algorithm to discover the edge and its correct direction. When the gap is relatively large, we can expect to discover the whole graph accurately during the learning process; When the gap is very small, we cannot guarantee to discover the graph directly and we must come up with another way to solve the causal bandit problem on an inaccurate model.

**C**ombinatorial Causal Bandits. The problem of combinatorial causal bandits (CCB) was first introduced by Feng and Chen (2023) and describes the following online learning task. The intervention can be performed on all variables except $X_1$ and $Y$ and is denoted by the do-operator *do* following earlier causal inference literature (Pearl, 2009b,a; Pearl and Mackenzie, 2018). The action set is defined by $\mathcal{A} \subseteq \{do(\boldsymbol{S} = \boldsymbol{s})\}_{\boldsymbol{S}\subseteq \boldsymbol{X}\setminus\{X_1\}, \boldsymbol{s}\in\{0,1\}^{|\boldsymbol{S}|}}$. The expected reward $Y$ under an intervention on $\boldsymbol{S} \subseteq \boldsymbol{X}\setminus\{X_1\}$ is denoted as $\mathbb{E}[Y|do(\boldsymbol{S} = \boldsymbol{s})]$. A learning agent runs an algorithm $\pi$ for $T$ rounds, taking parameter initializations and feedback from causal propagation as inputs, and outputting the selected interventions in all rounds. In particular, an *atomic intervention* intervenes on only one node, i.e. $|\boldsymbol{S}| = 1$. In this paper, we assume the null intervention $do()$ and atomic interventions $do(X = x)$ are always included in our action set $\mathcal{A}$, because they are needed to discover the graph structure.

The performance of the agent could be measured by the *regret* of the algorithm $\pi$. The regret $R^\pi(T)$ in our context is the difference between the cumulative reward using algorithm $\pi$ and the expected cumulative reward of choosing best action $do(\boldsymbol{S}^* = \boldsymbol{s}^*)$. Here, $do(\boldsymbol{S}^* = \boldsymbol{s}^*) \in \operatorname{argmax}_{do(\boldsymbol{S}=\boldsymbol{s})\in\mathcal{A}} \mathbb{E}[Y|do(\boldsymbol{S})]$. Formally, we have

$$R^\pi(T) = \mathbb{E}\left[\sum_{t=1}^{T}(\mathbb{E}[Y|do(\boldsymbol{S}^* = \boldsymbol{s}^*)] - \mathbb{E}[Y|do(\boldsymbol{S}^\pi_t = \boldsymbol{s}^\pi_t)])\right], \tag{1}$$

where $\boldsymbol{S}^\pi_t$ and $\boldsymbol{s}^\pi_t$ are the intervention set and intervention values selected by algorithm $\pi$ in round $t$ respectively. The expectation is from the randomness of the causal model and the algorithm $\pi$.

In this paper, we mainly focus on the regret minimization problem, and we will discuss the pure exploration problem and its sample complexity in the Section H. We defer the definition of sample complexity to that section.

## 4. Lower Bound on General Binary Causal Model

In this section, we explain why we only consider BGLM and BLM instead of the general binary causal model in the combinatorial causal bandit setting. In this context, a general binary causal model refers to a causal Bayesian model, in which all variables are restricted to either 0 or 1 values. Both BGLM and BLM are special cases of this model. Note that

in the general case both the number of actions and the number of parameters of the causal model are exponentially large to the size of the graph. The following theorem shows that in the general binary causal model, the regret bound must be exponential to the size of the graph when $T$ is sufficiently large, or simply linear to $T$ when $T$ is not large enough. This means that we cannot avoid the exponential factor for the general case, and thus justify our consideration of the BGLM and BLM settings with only a polynomial number of parameters relative to $n$.

**Theorem 1 (Binary Model Lower Bound)** *Recall that $n = |\boldsymbol{X}|$. For any algorithm, when $T \geq \frac{16(2^n-1)}{3}$, there exists a precise bandit instance of general binary causal model $\mathcal{T}$ such that*

$$\mathbb{E}_{\mathcal{T}}[R(T)] \geq \frac{\sqrt{2^n T}}{8e}.$$

*Moreover, when $T \leq \frac{16(2^n-1)}{3}$, there exists a precise bandit instance of general binary causal model $\mathcal{T}$ that*

$$\mathbb{E}_{\mathcal{T}}[R(T)] \geq \frac{T}{16e}.$$

The lower bound contains two parts. The first part shows that the asymptotic regret cannot avoid an exponential term $2^n$ when $T$ is large. The second part states that if $T$ is not exponentially large, the regret will be linear at the worst case. The proof technique of this lower bound is similar to but not the same as previous classical bandit, because the existence of observation $do()$ and atomic intervention $do(X_i = 1)$ may provide more information. To our best knowledge, this result is the first regret lower bound on the general causal model considering the potential role of observation and atomic intervention. The result shows that in the general binary causal model setting, it is impossible to avoid the exponential term in the cumulative regret even with the observations on null and atomic interventions. The proof of lower bound is provided in Appendix E.5.

The main idea is to consider the action set $\boldsymbol{A} = \{do(), do(X = x), do(\boldsymbol{X} = \boldsymbol{x})\}$ for all node $X$, $x \in \{0, 1\}$, $\boldsymbol{x} \in \{0, 1\}^n$ be the null intervention, atomic interventions and actions that intervene all nodes. The causal graph we use is a parallel graph where all nodes in $\boldsymbol{X}$ directly points to $Y$ with no other edges in the graph, and each node $X_i \in \boldsymbol{X}$ has probability $P(X_i = 1) = P(X_i = 0) = 0.5$. Intuitively, under this condition the null intervention and atomic interventions can provide limited information to the agent. This fact shows that observations and atomic interventions may not be conducive to our learning process in the worst case on the general binary causal model.

## 5. BGLM CCB without Graph Skeleton but with Minimum Weight Gap

In this section, we propose an algorithm for causal bandits on Markovian BGLMs based on maximum likelihood estimation (MLE) without any prior knowledge of the graph skeleton.

Our idea is to try to discover the causal graph structure and then apply the recent CCB algorithm with known graph structure (Feng and Chen, 2023). We discover the graph structure by using atomic interventions in individual variables. However, there are

a few challenges we need to face on graph discovery. First, it could be very difficult to exactly identify all parent-child relationships, since some grand-parent nodes may also have strong causal influence to its grand-child nodes. Fortunately, we find that it is enough to identify ancestor-descendant relationships instead of parent-child relationships, since we can artificially add an edge with 0 weight between each pair of ancestor and descendant without impacting the causal propagation results. Another challenge is the minimum weight gap. When the weight of an edge is very small, we need to perform more atomic interventions to identify its existence and its direction. Hence, we design an initialization phase with the number of rounds proportional to the total round number $T$ and promise that the ancestor-descendant relationship can always be identified correctly with a large probability when $T$ is sufficiently large.

Following Li et al. (2017); Feng and Chen (2023); Xiong and Chen (2023), we have three assumptions:

**Assumption 1** *For every $X \in \boldsymbol{X} \cup \{Y\}$, $f_X$ is twice differentiable. Its first and second order derivatives are upper-bounded by $L_{f_X}^{(1)} > 0$ and $L_{f_X}^{(2)} > 0$.*

Let $\kappa = \inf_{X \in \boldsymbol{X} \cup \{Y\}, \boldsymbol{v} \in [0,1]^{|\boldsymbol{Pa}(X)|}, ||\boldsymbol{\theta} - \boldsymbol{\theta}_X^*|| \leq 1} f_X'(\boldsymbol{v} \cdot \boldsymbol{\theta})$.

**Assumption 2** *We have $\kappa > 0$.*

**Assumption 3** *There exists a constant $\zeta > 0$ such that for any $X \in \boldsymbol{X} \cup \{Y\}$ and $X' \in \boldsymbol{Anc}(X)$, for any value vector $\boldsymbol{v} \in \{0,1\}^{|\boldsymbol{Anc}(X) \setminus \{X', X_1\}|}$, the following inequalities hold:*

$$\Pr_{\boldsymbol{\varepsilon}, \boldsymbol{X}, Y} \left( X' = 1 | \boldsymbol{Anc}(X) \setminus \{X', X_1\} = \boldsymbol{v} \right) \geq \zeta, \tag{2}$$

$$\Pr_{\boldsymbol{\varepsilon}, \boldsymbol{X}, Y} \left( X' = 0 | \boldsymbol{Anc}(X) \setminus \{X', X_1\} = \boldsymbol{v} \right) \geq \zeta. \tag{3}$$

Assumptions 1 and 2 are the classical assumptions in generalized linear model (Li et al., 2017). Assumption 3 makes sure that each ancestor node of $X$ has some freedom to become 0 and 1 with a non-zero probability, even when the values of all other ancestors of $X$ are fixed, and it is originally given in Feng and Chen (2023) with additional justifications. For BLMs and continuous linear models, we propose an algorithm based on linear regression without the need of this assumption in Appendix D. Furthermore, we suppose that $\text{Range}(f_X) = \mathbb{R}$. As in Feng et al. (2024), in Appendix A we demonstrate that any function $f_X$ can be transformed into a function with range $\mathbb{R}$ without affecting the propagation of BGLM.

To discover the ancestors of all variables, we need to perform an extra initialization phase (see Algorithm 1). We denote the total number of rounds by $T$ and arbitrary constants $c_0, c_1$ to make sure that $c_0 T^{1/2} \in \mathbb{N}^+$ for simplicity of our writing. In the initialization phase, from $X_1$ to $X_n$, we intervene each of them to 1 and 0 for $c_0 T^{1/2}$ times respectively. We denote the value of $X$ in the $t^{th}$ round by $X^{(t)}$. For every two variables $X_i, X_j \in \mathbf{X} \setminus \{X_1\}$, if

$$\frac{1}{c_0\sqrt{T}} \sum_{k=1}^{c_0\sqrt{T}} \left( X_j^{\left(2ic_0\sqrt{T}+k\right)} - X_j^{\left((2i+1)c_0\sqrt{T}+k\right)} \right) > c_1 T^{-\frac{1}{5}}, \tag{4}$$

**Algorithm 1:** BGLM-OFU-Unknown for BGLM CCB Problem

**Input:** Graph $G = (\boldsymbol{X} \cup \{Y\}, E)$, action set $\mathcal{A}$, parameters $L_{f_X}^{(1)}, L_{f_X}^{(2)}, \kappa, \zeta$ in Assumption 1, 2 and 3, $c$ in Lecué and Mendelson's inequality (Nie, 2022), positive constants $c_0$ and $c_1$ for initialization phase such that $c_0\sqrt{T} \in \mathbb{N}^+$.
/* Initialization Phase: */
Initialize $T_0 \leftarrow 2(n-1)c_0T^{1/2}$.
Do each intervention among $do(X_2 = 1), do(X_2 = 0), \ldots, do(X_n = 1), do(X_n = 0)$ for $c_0T^{1/2}$ times in order and observe the feedback $(\mathbf{X}_t, Y_t)$, $1 \le t \le T_0$.
Compute the ancestors $\widehat{\boldsymbol{Anc}}(X)$, $X \in \boldsymbol{X} \cup \{Y\}$ by BGLM-Ancestors$((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{T_0}, Y_{T_0}), c_0, c_1)$ (see Algorithm 2).
/* Parameters Initialization: */
Initialize $\delta \leftarrow \frac{1}{3n\sqrt{T}}$, $R \leftarrow \lceil \frac{512n(L_{f_X}^{(2)})^2}{\kappa^4}(n^2 + \ln\frac{1}{\delta})\rceil$, $T_1 \leftarrow T_0 + \max\left\{\frac{c}{\zeta^2}\ln\frac{1}{\delta}, \frac{(8n^2-6)R}{\zeta}\right\}$ and $\rho \leftarrow \frac{3}{\kappa}\sqrt{\log(1/\delta)}$.
Do no intervention on BGLM $G$ for $T_1 - T_0$ rounds and observe feedback $(\boldsymbol{X}_t, Y_t), T_0+1 \le t \le T_1$.
/* Iterative Phase: */
**for** $t = T_1 + 1, T_1 + 2, \ldots, T$ **do**
  $\{\hat{\boldsymbol{\theta}}_{t-1,X}, M_{t-1,X}\}_{X \in \boldsymbol{X}\cup\{Y\}} =$ BGLM-Estimate$((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{t-1}, Y_{t-1}))$ (see Algorithm 5 in Appendix B).
  Compute the confidence ellipsoid $\mathcal{C}_{t,X} = \{\boldsymbol{\theta}'_X \in [0,1]^{|\widehat{\boldsymbol{Anc}}(X)|} : \left\|\boldsymbol{\theta}'_X - \hat{\boldsymbol{\theta}}_{t-1,X}\right\|_{M_{t-1,X}} \le \rho\}$ for any node $X \in \boldsymbol{X} \cup \{Y\}$.
  Adopt $\text{argmax}_{do(\boldsymbol{S}=\boldsymbol{s}) \in \mathcal{A}, \boldsymbol{\theta}'_{t,X} \in \mathcal{C}_{t,X}} \mathbb{E}[Y|do(\boldsymbol{S} = \boldsymbol{s})]$ as $(\boldsymbol{S}_t, \boldsymbol{s}_t, \tilde{\boldsymbol{\theta}}_t)$.
  Intervene all the nodes in $\boldsymbol{S}_t$ to $\boldsymbol{s}_t$ and observe the feedback $(\boldsymbol{X}_t, Y_t)$.
**end for**

we set $X_i$ as an ancestor of $X_j$. Here, $X_j^{(2ic_0\sqrt{T}+k)}$'s with $k \in [c_0\sqrt{T}]$ are the values of $X_j$ in the rounds that $do(X_i = 1)$ is chosen; $X_j^{((2i+1)c_0\sqrt{T}+k)}, k \in [c_0\sqrt{T}]$ are the values of $X_j$ in the rounds that $do(X_i = 0)$ is chosen. Specifically, if $X_i$ is not an ancestor of $X_j$, the value of $X_j$ is not impacted by intervention on $X_i$. Simultaneously, if $X_i \in \boldsymbol{Pa}(X_j)$, the value of $X_j$ is notably impacted by $do(X_i)$ so the difference of $X_j$ under $do(X_i = 1), do(X_i = 0)$ can be used as a discriminator for the ancestor-descendant relationship between $X_i$ and $X_j$. This is formally shown by Lemma 2.

**Lemma 2** *Let $G$ be a BGLM with parameter $\boldsymbol{\theta}^*$ that satisfies Assumption 2. Recall that $\theta^*_{\min} = \min_{(X',X)\in E}\theta^*_{X',X}$. If $X_i \in \boldsymbol{Pa}(X_j)$, we have $\mathbb{E}[X_j|do(X_i = 1)] - \mathbb{E}[X_j|do(X_i = 0)] \ge \kappa\theta^*_{X_i,X_j} \ge \kappa\theta^*_{\min}$; if $X_i$ is not an ancestor of $X_j$, we have $\mathbb{E}[X_j|do(X_i = 1)] = \mathbb{E}[X_j|do(X_i = 0)]$.*

We use the above idea to implement the procedure in Algorithm 2, and then put this procedure in the initial phase and integrate this step into BGLM-OFU proposed by Feng and Chen (2023), to obtain our main algorithm, BGLM-OFU-Unknown (Algorithm 1).

Notice that each term in Eq. (4) is a random sample of $\mathbb{E}[X_j|do(X_i = 1)] - \mathbb{E}[X_j|do(X_i = 0)]$, which means that the left-hand side of Eq. (4) is just an estimation of $\mathbb{E}[X_j|do(X_i =$

**Algorithm 2:** BGLM-Ancestors

**Input:** Observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{T_0}, Y_{T_0})$, positive constants $c_0$ and $c_1$.
**Output:** $\widehat{\boldsymbol{Anc}}(X)$, ancestors of $X$, $X \in \boldsymbol{X} \cup \{Y\}$.
For all $X \in \boldsymbol{X}$, $\widehat{\boldsymbol{Anc}}(X) = \emptyset$, $\widehat{\boldsymbol{Anc}}(Y) = \boldsymbol{X}$.
**for** $i \in \{2, 3, \ldots, n\}$ **do**
  **for** $j \in \{2, 3, \ldots, n\} \backslash \{i\}$ **do**
    **if** $\sum_{k=1}^{c_0\sqrt{T}} \left( X_j^{(2ic_0\sqrt{T}+k)} - X_j^{((2i+1)c_0\sqrt{T}+k)} \right) > c_0 c_1 T^{3/10}$ **then**
      Add $X_i$ into $\widehat{\boldsymbol{Anc}}(X_j)$.
    **end if**
  **end for**
**end for**
Recompute the transitive closure of $\widehat{\boldsymbol{Anc}}(\cdot)$, i.e., if $X_i \in \widehat{\boldsymbol{Anc}}(X_j)$ and $X_j \in \widehat{\boldsymbol{Anc}}(X_\ell)$, then add $X_i$ to $\widehat{\boldsymbol{Anc}}(X_\ell)$.

$1)] - \mathbb{E}[X_j | do(X_i = 0)]$. Such expression can be bounded with high probability by concentration inequalities. Hence we can prove that Algorithm 2 identifies $X_i \in \boldsymbol{Anc}(X_j)$ with false positive rate and false negative rate both no more than $\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right)$ when $\theta^*_{\min} \geq 2c_1\kappa^{-1}T^{-1/5}$. Formally, we have the following lemma that shows the probability of correctness for Algorithm 2. For completeness, the proof of Lemma 3 is put in appendix.

**Lemma 3 (Positive Rate of BGLM-Order)** *Suppose Assumption 2 holds for BGLM $G$. In the initialization phase of Algorithm 1, Algorithm 2 finds a consistent ancestor-descendant relationship for $G$ with probability no less than $1 - 2\binom{n-1}{2}\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right)$ when $\theta^*_{\min} \geq 2c_1\kappa^{-1}T^{-1/5}$.*

We refer to the condition $\theta^*_{\min} \geq 2c_1\kappa^{-1}T^{-1/5}$ in this lemma as *weight gap assumption*. The number of initialization rounds in Algorithm 1 is $O(\sqrt{T})$. According to Lemma 3, the expected regret contributed by incorrectness of the ancestor-descendant relationship does not exceed $O\left(T\exp\left(-\frac{c_0 c_1^2 T^{1/10}}{2}\right)\right) = o(\sqrt{T})$. Therefore, after adding the initialization, the expected regret of BGLM-OFU-Unknown increases by no more than $o(\sqrt{T})$ over BGLM-OFU (Algorithm 1 in Feng and Chen (2023)). Similar to BGLM-OFU, during the iterative phase, MLE is employed to estimate all the parameters. Simultaneously, a pair oracle is utilized to identify the optimal parameter configuration and intervention set within the confidence ellipsoid. Thus we have the following theorem to show the regret of Algorithm 2, which is formally proved in appendix.

**Theorem 4 (Regret Bound of BGLM-OFU-Unknown)** *Under Assumptions 1, 2 and 3, the regret of BGLM-OFU-Unknown (Algorithms 1, 2 and 5) is bounded as*

$$R(T) = O\left(\frac{1}{\kappa} n^{\frac{3}{2}} L^{(1)}_{\max} \sqrt{T} \log T\right), \tag{5}$$

*where* $L^{(1)}_{\max} = \max_{X \in \boldsymbol{X} \cup \{Y\}} L^{(1)}_{f_X}$ *and the terms of* $o(\sqrt{T}\ln T)$ *are omitted, and the big O notation holds for* $T \geq 32\left(\frac{c_1}{\kappa\theta^*_{\min}}\right)^5$.

Compared to Feng and Chen (2023), Theorem 4 has the same asymptotic regret, The only additional assumption is $T \geq 32\left(c_1/(\kappa\theta^*_{\min})\right)^5$. Intuitively, this extra assumption guarantees that we can discover the ancestor-descendant relationship consistent with the true graph. Our result indicates that not knowing the causal graph does not provide substantial difficulty with the weight gap assumption.

**Remark 5** *Our BGLM-OFU-Unknown regret bound aligns with the regret bound of BGLM-OFU as presented in Feng and Chen (2023). Furthermore, the leading term* $O(\sqrt{T}\ln T)$ *is consistent with the regret bounds previously established for combinatorial bandit algorithms (Li et al., 2020; Zhang et al., 2022) and causal bandit algorithms (Lu et al., 2020).*

*Because Lemma 3 requires weight gap assumption, in the proof of this regret bound, we only consider the case of* $T \geq 32\left(c_1/(\kappa\theta^*_{\min})\right)^5$. *This limitation does not impact the asymptotic big O notation in our regret bound. However, when the round number* $T$ *is not that large, the regret can be linear with respect to* $T$. *We remove this weight gap assumption in Section 6 for the linear model setting. The* $c_0$ *and* $c_1$ *are two adjustable constants in practice. When* $T$ *is small, one could try a small* $c_0$ *to shorten the initialization phase, i.e., to make sure that* $T_0 \ll T$, *and a small* $c_1$ *to satisfy the weight gap assumption. When* $T$ *is large, one could consider larger* $c_0$ *and* $c_1$ *for a more accurate ancestor-descendant relationship. However, because* $\theta^*_{\min}$ *is unknown, one cannot promise that the weight gap assumption holds by manipulating* $c_1$, *i.e.,* $\theta^*_{\min}$ *may be too small for any practical* $T$ *given* $c_1$.

## 6. BLM CCB without Graph Skeleton and Weight Gap Assumption

In the previous section, we find that if $T > O((\theta^*_{\min})^{-5})$, we can get a valid upper bound. However, in reality, we have two challenges: 1) We do not know the real value of $\theta^*_{\min}$, and this makes it hard to know when an edge's direction is identified. 2) When $\theta^*_{\min} \to 0$, it makes it very difficult to estimate the graph accurately. To solve these challenges, we must both eliminate the dependence of $\theta^*_{\min}$ in our analysis, and think about how the result will be influenced by an inaccurate model. In this section, we give a causal bandit algorithm and show that the algorithm can always give $\tilde{O}(T^{2/3})$ regret. This sub-linear regret result shows that the challenge can be solved by some additional techniques.

In this section, we consider a special case of BGLM called Binary Linear Model (BLM), where $f_X$ becomes identity function. The linear structure allows us to release the Assumption 1-3 (Feng and Chen, 2023) and analyze the influence of an inaccurate model.

The main algorithm follows the BLM-LR algorithm in Feng and Chen (2023), which uses linear regression to estimate the weight $\boldsymbol{\theta}^*$, and the pseudocode is provided in Algorithm 3. We add a graph discovery process (Algorithm 4) in the initialization phase using $O(nT^{2/3}\log T)$ times rather than $O(nT^{1/2})$ in the previous section. For any edge $X' \to X$ with weight $\theta^*_{X',X} \geq T^{-1/3}$, with probability at least $1 - 1/T^2$, we expect to identify the edge's direction within $O(nT^{2/3}\log(T))$ samples for $do(X' = 1)$ and $do(X' = 0)$ by checking whether the difference $P(X \mid do(X' = 1)) - P(X \mid do(X' = 0))$ is large than $T^{-1/3}$. Since

**Algorithm 3:** BLM-LR-Unknown for BLM CCB Problem without Weight Gap

**Input:** Graph $G = (\boldsymbol{X} \cup \{Y\}, E)$, action set $\mathcal{A}$, positive constants $c_0$ and $c_1$ for initialization phase.
/* Initialization Phase: */
Initialize $T_0 \leftarrow 2(n-1)c_0 T^{2/3}\log(T)$.
Do each intervention among $do(X_2 = 1), do(X_2 = 0), \ldots, do(X_n = 1), do(X_n = 0)$ for $c_0 T^{2/3}$ times in order and observe the feedback $(\mathbf{X}_t, Y_t)$ for $1 \le t \le T_0$.
Compute the ancestors $\widehat{\boldsymbol{Anc}}(X)$, $X \in \boldsymbol{X} \cup \{Y\}$ by Nogap-BLM-Ancestors$((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{T_0}, Y_{T_0}), c_0, c_1)$ (see Algorithm 4).
/* Parameters Initialization: */
Initialize $\delta \leftarrow \frac{1}{n\sqrt{T}}$, $\rho_t \leftarrow \sqrt{n\log(1+tn) + 2\log\frac{1}{\delta}} + \sqrt{n}$ for $t = 0, 1, 2, \ldots, T$, $M_{T_0,X} \leftarrow \mathbf{I} \in \mathbb{R}^{|\widehat{\boldsymbol{Anc}}(X)| \times |\widehat{\boldsymbol{Anc}}(X)|}$, $\boldsymbol{b}_{T_0,X} \leftarrow \mathbf{0}^{|\widehat{\boldsymbol{Anc}}(X)|}$ for all $X \in \boldsymbol{X} \cup \{Y\}$ and $\hat{\boldsymbol{\theta}}_{T_0,X} \leftarrow 0 \in \mathbb{R}^{|\widehat{\boldsymbol{Anc}}(X)|}$ for all $X \in \boldsymbol{X} \cup \{Y\}$.
/* Iterative Phase: */
**for** $t = T_0 + 1, T_0 + 2, \ldots, T$ **do**
  Compute the confidence ellipsoid $\mathcal{C}_{t,X} = \{\boldsymbol{\theta}'_X \in [0,1]^{|\widehat{\boldsymbol{Anc}}(X)|} : \left\|\boldsymbol{\theta}'_X - \hat{\boldsymbol{\theta}}_{t-1,X}\right\|_{M_{t-1,X}} \le \rho_{t-1}\}$ for any node $X \in \boldsymbol{X} \cup \{Y\}$.
  Adopt $\text{argmax}_{do(\boldsymbol{S}=\boldsymbol{s}) \subseteq \mathcal{A}, \boldsymbol{\theta}'_{t,X} \in \mathcal{C}_{t,X}} \mathbb{E}[Y | do(\boldsymbol{S} = \boldsymbol{s})]$ as $(\boldsymbol{S}_t, \boldsymbol{s}_t, \tilde{\boldsymbol{\theta}}_t)$.
  Intervene all the nodes in $\boldsymbol{S}_t$ to $\boldsymbol{s}_t$ and observe the feedback $(\boldsymbol{X}_t, Y_t)$.
  **for** $X \in \boldsymbol{X} \cup \{Y\}$ **do**
    Construct data pair $(\boldsymbol{V}_{t,X}, X^{(t)})$ with $\boldsymbol{V}_{t,X}$ the vector of ancestors of $X$ in round $t$, and $X^{(t)}$ the value of $X$ in round $t$ if $X \notin S_t$.
    $M_{t,X} = M_{t-1,X} + \boldsymbol{V}_{t,X}\boldsymbol{V}_{t,X}^\intercal$, $\boldsymbol{b}_{t,X} = \boldsymbol{b}_{t-1,X} + X^{(t)}\boldsymbol{V}_{t,X}$, $\hat{\boldsymbol{\theta}}_{t,X} = M_{t,X}^{-1}\boldsymbol{b}_{t,X}$.
  **end for**
**end for**

**Algorithm 4:** Nogap-BLM-Ancestors

**Input:** Observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_{T_0}, Y_{T_0})$, positive constants $c_0$ and $c_1$.
**Output:** For all $X \in \boldsymbol{X} \cup \{Y\}$, $\widehat{\boldsymbol{Anc}}(X)$.
For all $X \in \boldsymbol{X}$, $\widehat{\boldsymbol{Anc}}(X) = \emptyset$, $\widehat{\boldsymbol{Anc}}(Y) = \boldsymbol{X}$.
**for** $i \in \{2, 3, \ldots, n\}$ **do**
  **for** $j \in \{2, 3, \ldots, n\} \backslash \{i\}$ **do**
    **if** $\sum_{k=1}^{c_0 T^{2/3}} \left( X_j^{\left(c_0(2i)T^{2/3}+k\right)} - X_j^{\left(c_0(2i+1)T^{2/3}+k\right)} \right) > c_0 c_1 T^{1/3}\log(T^2)$ **then**
      Add $X_i$ into $\widehat{\boldsymbol{Anc}}(X_j)$.
    **end if**
  **end for**
**end for**
Recompute the transitive closure of $\widehat{\boldsymbol{Anc}}(\cdot)$.

the above difference is always larger than $\theta^*_{X',X}$, after the initialization phase, the edge $X' \to X$ will be added to the graph if $\theta^*_{X',X} \ge T^{-1/3}$.

Moreover, if $X'$ is not an ancestor of $X$, we claim that it cannot be a estimated as an ancestor after the initialization phase. This is because in this case $P(X \mid do(X' = 1)) = P(X) = P(X \mid do(X' = 0))$. Denote the estimated graph $G'$ as the graph with edge $X' \to X$ for all $X' \in \widehat{\boldsymbol{Anc}}(X)$. We then have the following lemma.

**Lemma 6** *In Algorithm 3, if the constants $c_0$ and $c_1$ satisfy that $c_0 \geq \max\{\frac{1}{c_1^2}, \frac{1}{(1-c_1)^2}\}$, with probability at least $1-(n-1)(n-2)\frac{1}{T^{1/3}}$, after the initialization phase we have*

*1). If $X'$ is a true parent of $X$ in $G$ with weight $\theta^*_{X',X} \geq T^{-1/3}$, the edge $X' \to X$ will be identified and added to the estimated graph $G'$.*

*2). If $X'$ is not an ancestor of $X$ in $G$, $X' \to X$ will not be added into $G'$.*

The properties above together provide the analytic basis for the following observation, which plays a key role in our further analysis. Denote the estimated accuracy $r = T^{-1/3}$. We know the linear regression for $X$ will be performed on $X$ and all its possible ancestors $\widehat{\boldsymbol{Anc}}(X)$ we estimated. For the true parent node $X'$ in $G$ that is not contained in $\widehat{\boldsymbol{Anc}}(X)$, we have $\theta^*_{X',X} \leq r$. Suppose $\widehat{\boldsymbol{Anc}}(X) = \{X_1, X_2, \ldots, X_m\}$, and true parents which is not contained in $\widehat{\boldsymbol{Anc}}(X)$ are $X_{m+1}, \ldots, X_{m+k}$. Thus we have $\theta^*_{X_{m+i},X} \leq r$ for all $1 \leq i \leq k$.

Also, assume $X_1, \ldots, X_t (t < m)$ are true parents of $X$ in $G$. For $X_{m+i}$, by law of total expectation, the expectation of $X$ can be rewritten as

$$\begin{aligned}
&\mathbb{E}[X \mid X_1, \ldots, X_t] \\
&= \mathbb{E}_{X_{m+1},\ldots,X_{m+k}}[\mathbb{E}[X \mid X_1, \ldots, X_t, X_{m+1}, \ldots, X_{m+k}]] \\
&= \mathbb{E}_{X_{m+1},\ldots,X_{m+k}}\left[\sum_{i=1}^{t} \theta^*_{X_i,X} X_i + \sum_{i=m+1}^{m+k} \theta^*_{X_i,X} X_i\right] \\
&= \sum_{i=1}^{t} \theta^*_{X_i,X} X_i + \sum_{i=m+1}^{m+k} \theta^*_{X_i,X} \mathbb{E}[X_i] = \sum_{i=1}^{t} \theta^{*'}_{X_i,X} X_i,
\end{aligned}$$

where

$$\theta^{*'}_{X_i,X} = \theta^*_{X_i,X}, \quad i \geq 2, \tag{6}$$

$$\theta^{*'}_{X_1,X} = \theta^*_{X_1,X} + \sum_{i=m+1}^{m+k} \theta^*_{X_i,X} \mathbb{E}[X_i]. \tag{7}$$

Eq. (7) is because $X_1 = 1$ always holds. Then we have $|\theta^{*'}_{X_i,X} - \theta^*_{X_i,X}| \leq \sum_{i=m+1}^{m+k} \theta^*_{X_i,X} \leq kr \leq nr$, which shows that the difference between $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}$ is small if accuracy $r$ is small. In Eqs. (6) and (7), we employ the linear property of BLMs, which is the reason that we are only able to perform transformations from $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}^{*\prime}$ for BLMs. Let model $M'$ represent the model with graph $G'$ with weights $\boldsymbol{\theta}^{*'}$ defined above. The following lemma shows the key observation:

**Lemma 7** *The linear regression performed on graph $G'$ in Algorithm 3 (lines 13–16) gives the estimation $\hat{\boldsymbol{\theta}}'$ such that*

$$\|(\hat{\boldsymbol{\theta}'}_{t,X} - \boldsymbol{\theta}^{*'}_X)\|_{M_{t,X}} \leq \sqrt{n\log(1+tn) + 2\log(1/\delta)} + \sqrt{n},$$

*where $M_{t,X}$ is defined in Algorithm 3.*

This lemma shows that, the linear regression performed on the inaccurate estimated linear model $M'$ is equivalent to the regression for $\boldsymbol{\theta}^{*'}$. Note that this regression only gives us the approximation in some direction with respect to elliptical norm, allowing the variables to be dependent.

Based on claim above, we only need to measure the difference for $\mathbb{E}[Y \mid do(\boldsymbol{S} = 1)]$ on model $M$ and $M'$. The following lemma shows that the difference between two models can be bounded by our estimated accuracy $r$:

**Lemma 8** *$|\mathbb{E}_M[Y \mid do(S = \boldsymbol{1})] - \mathbb{E}_{M'}[Y \mid do(S = \boldsymbol{1})]| \leq n^2(n+1)r$, where $r$ is the estimated accuracy defined in the start of this section.*

The Lemma 8 gives us a way to bound our linear regression performance on the estimated model $M'$. Suppose our linear regression achieves $O(\sqrt{T})$ regret comparing to $\max_{\boldsymbol{S}} \mathbb{E}_{M'}[Y \mid do(\boldsymbol{S} = \mathbf{1})]$, based on our estimated accuracy $r = O(T^{-1/3})$, the regret for optimization error is $O(T^{2/3})$, which is the same order as the initialization phase. Moreover, it implies that we cannot set $r$ to a larger gap, such as $r = O(T^{-1/2})$, as doing so would result in the initialization phase regret becoming linearly proportional to $T$.

From these two lemmas, we can measure the error for initialization phase. Motivated by Explore-then-Commit framework, we can achieve sublinear regret without the weight gap assumption. The detailed proof is provided in Appendix E.2 and Appendix E.3.

**Theorem 9** *If $c_0 \geq \max\{\frac{1}{c_1^2}, \frac{1}{(1-c_1)^2}\}$, the regret of Algorithm 3 running on BLM is upper bounded as*

$$R(T) = O((n^3 T^{2/3}) \log T).$$

Theorem 9 states the regret of our algorithm without weight gap. The leading term of the result is $O(T^{2/3} \log T)$, which has higher order than $O(\sqrt{T} \log T)$, the regret of the previous Algorithm 2 and the BLM-LR algorithm in Feng and Chen (2023). This degradation in regret bound can be viewed as the cost of removing the weight gap assumption, which makes the accurate discovery of the causal graph extremely difficult. For a detailed discussion of the weight gap assumption, interested readers can refer to Appendix F. How to devise a $O(\sqrt{T} \log T)$ algorithm without weight gap assumption is still an open problem.

Using the transformation in Section 5.1 in Feng and Chen (2023), this algorithm can also work with hidden variables. The model our algorithm can work on allows hidden variables but disallows the graph structure where a hidden node has two paths to $X_i$ and $X_i$'s descendant $X_j$ and the paths contain only hidden nodes except the end points $X_i$ and $X_j$.

Moreover, observe that Algorithms 1 and 3 necessitate prior knowledge of the horizon $T$. To circumvent this constraint, the "Doubling Trick" can be employed, converting our algorithms into anytime algorithms without compromising the regret bounds.

## 7. Future Work

This paper is the first theoretical study on causal bandits without the graph skeleton. There are many future directions to extend this work. We believe that similar initialization methods and proof techniques can be used to design causal bandits algorithms for

other parametric models without the skeleton, like linear structural equation models (SEM). Moreover, how to provide an algorithm with $\widetilde{O}(\sqrt{T})$ regret without weight gap assumption is interesting and still open. One possibility is to investigate the utilization of feedback during the iterative phase of the BGLM-OFU-Unknown algorithm in order to enhance graph structure identification, which could potentially lead to an improved regret bound.

## Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and constructive feedback.

## References

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

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

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

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1): 1–122, 2012.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pages 151–159. PMLR, 2013.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

Shi Feng and Wei Chen. Causal inference for influence propagation—identifiability of the independent cascade model. In *International Conference on Computational Data and Social Networks*, pages 15–26. Springer, 2021.

Shi Feng and Wei Chen. Combinatorial causal bandits. *arXiv preprint arXiv:2206.01995*, 2022.

Shi Feng and Wei Chen. Combinatorial causal bandits. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI)*, February 2023. URL https://www.microsoft.com/en-us/research/publication/combinatorial-causal-bandits/.

Shi Feng, Nuoya Xiong, Zhijie Zhang, and Wei Chen. A correction of pseudo log-likelihood method. *arXiv preprint arXiv:2403.18127*, 2024.

David Galles and Judea Pearl. Testing identifiability of causal effects. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 185–195, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

José A Gámez, Serafín Moral, and Antonio Salmerón Cerdan. *Advances in Bayesian networks*, volume 146. Springer, 2013.

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

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

Mikhail Konobeev, Jalal Etesami, and Negar Kiyavash. Causal bandits without graph learning. *arXiv preprint arXiv:2301.11401*, 2023.

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

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

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

Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in neural information processing systems*, 33: 8565–8576, 2020.

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

Shuai Li, Fang Kong, Kejie Tang, Qizhi Li, and Wei Chen. Online influence maximization under linear threshold model. *Advances in Neural Information Processing Systems*, 33: 1192–1204, 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*, pages 141–150. PMLR, 2020.

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

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

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

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

Zipei Nie. Matrix anti-concentration inequalities with applications. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 568–581, 2022.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146, 2009a. *do*10.1214/09-SS057. URL https://doi.org/10.1214/09-SS057.

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

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 3–11, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

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*, pages 3057–3066. PMLR, 2017.

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

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

Nuoya Xiong and Wei Chen. Combinatorial pure exploration of causal bandits. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, May 2023.

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