Causal Bandits for Linear Structural Equation Models

Burak Varıcı   VARICB@RPI.EDU
Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Karthikeyan Shanmugam *   KARTHIKEYANVS@GOOGLE.COM
Google Research India
Bangalore, India

Prasanna Sattigeri   PSATTIG@US.IBM.COM
MIT-IBM Watson AI Lab
IBM Research
Cambridge, MA 02142, USA

Ali Tajer   TAJER@ECSE.RPI.EDU
Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Editor: Daniel Roy

Abstract

This paper studies the problem of designing an optimal sequence of interventions in a causal graphical model to minimize cumulative regret with respect to the best intervention in hindsight. This is, naturally, posed as a causal bandit problem. The focus is on causal bandits for linear structural equation models (SEMs) and soft interventions. It is assumed that the graph's structure is known and has N nodes. Two linear mechanisms, one soft intervention and one observational, are assumed for each node, giving rise to 2^N possible interventions. The majority of the existing causal bandit algorithms assume that at least the interventional distributions of the reward node's parents are fully specified. However, there are 2^N such distributions (one corresponding to each intervention), acquiring which becomes prohibitive even in moderate-sized graphs. This paper dispenses with the assumption of knowing these distributions or their marginals. Two algorithms are proposed for the frequentist (UCB-based) and Bayesian (Thompson sampling-based) settings. The key idea of these algorithms is to avoid directly estimating the 2^N reward distributions and instead estimate the parameters that fully specify the SEMs (linear in N) and use them to compute the rewards. In both algorithms, under boundedness assumptions on noise and the parameter space, the cumulative regrets scale as $\tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT})$, where d is the graph's maximum degree, and L is the length of its longest causal path. Additionally, a minimax lower of $\Omega(d^{\frac{L}{2}-2}\sqrt{T})$ is presented, which suggests that the achievable and lower bounds conform in their scaling behavior with respect to the horizon T and graph parameters d and L.

Keywords: Causal bandits, multi-armed bandits, causality, linear SEMs, cumulative regret.

---

*. The project was started when the author was at IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA.

---

1. Introduction

Multi-armed bandit (MAB) settings provide a rich theoretical context for formalizing and analyzing sequential experimental design procedures. Each arm represents one experiment in a canonical MAB setting, the stochastic outcome of which is represented by a random reward. The objective of a learner is to design a sequence of arm selections (i.e., experiments) that maximizes the cumulative reward over a time horizon. Bandit problems have a growing list of applications in various domains such as marketing campaigns (Sawant et al., 2018), clinical trials (Liu et al., 2020), portfolio management (Shen et al., 2015), recommender systems (Zhou et al., 2017). Various assumptions on the statistical dependencies among the arms have led to different classes of problems, such as linear bandits (Abbasi-Yadkori et al., 2011; Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010), combinatorial bandits (Cesa-Bianchi and Lugosi, 2012), and contextual bandits (Tewari and Murphy, 2017). In this paper, we focus on *causal bandits* (Lattimore et al., 2016) as another instance of a bandit setting in which reward models are assumed to be the results of causal relationships.

Causal Bayesian networks represent the cause-effect relationships through a directed acyclic graph (DAG). Each node of the causal graph represents a random variable, the edges encode the statistical dependencies among them, and the directed edges signify causal relationships. Recently, the effectiveness of DAGs in encoding non-trivial dependencies among random variables has led to an interest in the study of causal bandits. In a causal bandit framework, *interventions* on the nodes of a DAG are modeled as arms, and the post-intervention stochastic observations are modeled as the arm rewards. Instead of a cumulative reward, one node's observation is generally regarded as the reward value. This node is usually selected to be one without any descendants. The structure of the underlying DAG induces statistical dependencies among the arms' rewards and their sequential selection. Furthermore, different interventions on different subsets of nodes result in distinct reward distribution for the reward node. In such a causal bandit framework, the objective is to minimize the cumulative regret with respect to the best intervention in hindsight.

Causal bandits are effective for modeling several complex real-world problems. For instance, in drug discovery, dosages of multiple drugs can be adjusted adaptively to identify a desirable clinical outcome (Liu et al., 2020). Likewise, in advertisements campaigns, an advertisement approach can adaptively adjust its strategies that can be modeled as interventions on their advertisement system to maximize their return on advertisement investment (Lu et al., 2020, 2021; Nair et al., 2021). In each of these applications, several variables affect the observed reward. In parallel, these variables also have cause-effect relationships among them. Hence, an optimal design of experiments in such settings often necessitate simultaneously performing interventions on multiple variables.

Literature review. Designing a causal bandit problem hinges critically on the extent of assumptions on the topology of the underlying DAG and the probability distribution of its random variables. The existing literature can be categorized based on different combinations of these assumptions.

The majority of the existing literature assumes that both the topology and the distributions of the reward node's parents are fully specified under all possible intervention models (Lattimore et al., 2016; Sen et al., 2017; Lu et al., 2020; Nair et al., 2021). Among these works, the initial studies posed the causal bandit problem as best arm identification, where the learner does not incur a regret for exploration (Bareinboim et al., 2015; Lattimore et al., 2016; Sen et al., 2017). The impact of exploration was further accounted for by Lu et al. (2020). Specifically, Lu et al. (2020) proposed algorithms that capitalize on the causal information to improve the cumulative regrets compared to the algorithms that do not use this information. Nair et al. (2021) extended the previous work

by providing an instance-dependent regret bound that, in the worst case, grows with the number of interventions provided. The shared assumption of these works is the a priori knowledge of interventional distributions of parents of the reward node. This is a rather strong assumption since the number of such distributions can be prohibitive. **In contrast, we assume that the interventional distributions, or their marginals, are unknown.**

The setting without the knowledge of the topology and the structure of the interventional distributions has been studied more recently (de Kroon et al., 2022; Lu et al., 2021; Bilodeau et al., 2022). However, the algorithm of de Kroon et al. (2022) relies on auxiliary separating set algorithms, and its improvement upon non-causal algorithms is shown only empirically. Lu et al. (2021) provide an improved regret result for a similar setting, though only for the atomic interventions. Bilodeau et al. (2022) consider a different case and establish adaptive regret guarantees with respect to an unknown causal structure. They assume access to an estimate of the interventional distributions, which can be learned offline. Each possible arm is played a certain number of times to ensure that these estimates are sufficiently accurate. Notably, no assumptions are made about the structure of the distributions. Furthermore, their proposed adaptive algorithm is agnostic to the causal structure and incurs sublinear regret even in the presence of unobserved confounders. The main distinction of our work compared to Bilodeau et al. (2022) is that their focus is on the adaptivity of the regret guarantees, and their algorithm's regret upper bound grows with the cardinality of the intervention space.

One setting between the previous two settings assumes that the topology is known while the statistical models are unknown. This setting was first studied by Yabe et al. (2018), and they obtained a simple regret guarantee that scales polynomially in the graph size. However, the scope of this study is limited to binary random variables and the best arm identification objective. Maiti et al. (2022) propose an algorithm that takes a semi-Markovian graph as input and achieves a simple regret almost optimal for a certain class of graphs. They also propose an algorithm that achieves an improved cumulative regret compared to the non-causal bandit algorithms when all variables are observed. The setting, however, is focused on atomic interventions on binary random variables. Unlike Maiti et al. (2022), in which the cardinality of the intervention space is at most N, in this paper, the intervention space consists of all subsets of the graph's nodes with cardinality 2^N. Feng and Chen (2023) posed the non-atomic interventions as a combinatorial bandit problem of finding the best set of nodes to intervene with a pre-specified cardinality. They assume that the causal mechanisms follow a generalized linear model with $\{0, 1\}$ Boolean variables. In their paper, they consider only hard interventions, unlike our case. Therefore, the Gram matrix associated with their local least-squares estimation is always about estimating the observational causal mechanism. In our case, due to the soft intervention model, eigenvalues of the Gram matrices for estimating the different interventional mechanisms are not straightforward to control, given that one needs to account for other simultaneous interventions. Xiong and Chen (2023) also consider combinatorial bandits and propose algorithms with gap-dependent sample complexity results for pure exploration problems. Notably, these papers (Yabe et al., 2018; Maiti et al., 2022; Xiong and Chen, 2023; Feng and Chen, 2023) consider only binary random variables. In this paper, we assume that only the DAG topology is known, which is a setup similar to that of Yabe et al. (2018) and Feng and Chen (2023). Nevertheless, we have significant differences in the intervention model, the space of interventions, and the causal mechanisms assumed.
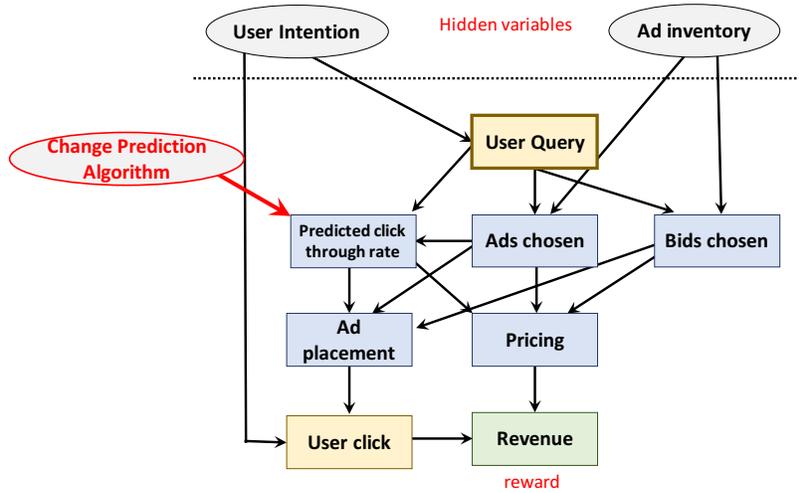
Figure 1: A computational advertising example borrowed from (Bottou et al., 2013). The interventions are changes in the prediction algorithm for the click rate. Since they do not remove any edges, they are soft interventions.

Motivation for soft interventions. We note that most existing causal bandit algorithms assume deterministic hard interventions, i.e., $do()$ interventions, for minimizing cumulative regret. The exceptions include the study by Sen et al. (2017), which considers soft interventions on a single node for best arm identification. Yabe et al. (2018) remark that it can also be extended to Boolean variables. For minimizing cumulative regret under unknown interventional distributions, which is the goal of this paper, the only exception is the adaptive algorithm of Bilodeau et al. (2022).

Under hard interventions, the variables are forced to take fixed values, and parental effects are completely absent. In contrast, soft interventions, which modify the conditional distributions of the target variables, can be more realistic in certain applications and have found recent interest in causal inference literature (Jaber et al., 2020; Varici et al., 2021). In this paper, we focus on soft interventions.

A motivating example of soft interventions is the computational advertising studied by Bottou et al. (2013). The elements of an advertising system (e.g., query, prices, and click rates) and their interactions can be modeled by a causal graph with a known structure, and the reward variable accounts for the resulting revenue. However, the strengths of the interactions between the nodes are unknown. Interventions are a set of click-rate prediction algorithms. Figure 1 illustrates the described model. The interventions are not hard in this causal system since they do not remove the causal effects.

Contributions. Motivated by the limitations of the existing works, our objective is to answer the following question: **Can we use only the graph structure without the knowledge of interventional distributions, allow soft interventions on continuous variables, and obtain cumulative regret guarantees that scale optimally with the horizon ($\sqrt{T}$) while not growing with the cardinality of the intervention space?** We answer this question affirmatively when the causal system follows a linear structural equation model (SEM) with unknown parameters, and the interventions inherit that structure.

We consider per-node interventional mechanisms, giving rise to 2^N possible intervention models. Standard linear bandit algorithms also are inapplicable since the feature vectors of arms are unknown in our case. We propose the **Lin**ear **SEM U**pper **C**onfidence **B**ound the (LinSEM-UCB) algorithm that achieves the cumulative regret $\tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT})$ in which d is the maximum degree, and L is the length of the longest causal path in the graph. Importantly, the number of interventions 2^N does not appear in our regret result. Instead, we have $d^{L+\frac{1}{2}}$ term that depends on the topology. We also propose the **Lin**ear **SEM T**hompson **S**ampling (LinSEM-TS) algorithm for the Bayesian setting, and its Bayesian regret scales similar to that of LinSEM-UCB, i.e., $\tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT})$. Finally, we establish a regret lower bound that is a constant factor of $d^{\frac{L}{2}-2}\sqrt{T}$, which matches the behavior of our achievable regret $\tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT})$ in terms of the horizon T and graph parameters d and L.

Key technical challenges. The central piece of our analysis pertains to characterizing the hitherto uninvestigated cumulative regret when we have continuous variables, an intervention space that is the power set of the graphs' nodes, and unknown interventional distributions. This analysis faces several challenges while developing bandit algorithms with regret guarantees. First, the linear SEM, i.e., weight matrices and the noise model, are unknown. Since these pieces of information are needed for specifying the interventional distributions, not knowing them renders these distributions unknown. A direct (naive) approach to regret analysis involves estimating all the unknown distributions, the complexity of which grows quickly. Motivated by the fact that estimating 2^N interventional distributions explicitly bears significant redundancy, to circumvent the computational challenge, we start by expressing the reward as a function of the intervention's weight matrix and the random noise vector, which is independent of the action. We note that the measurements from the nodes are observed *only after* an intervention is applied. Hence, these noise vectors need to be independent of the actions to reduce the dependency on interventions to the weight matrices. This relationship is subsequently leveraged to estimate the unknown weight matrices containing at most $2Nd$ parameters. Finally, these estimates are leveraged to estimate the rewards.

A second challenge is due to the non-linearity of the relationship between the reward values observed and the unknown weight matrices. Specifically, even though the causal model is linear, interventions have a compounding effect on the mean values of the nodes inducing non-linearity. This renders the problem fundamentally different from the linear bandit problems. The third challenge is due to applying soft interventions. Under soft interventions, unlike hard interventions, the effects of the intervention targets' ancestors on the reward node are not nullified. In addition, having continuous variables (instead of binary, e.g., Yabe et al. (2018); Feng and Chen (2023)) necessitates a careful examination of the compounded effect of the estimation errors through the causal paths.

Our regret analysis naturally decomposes the total regret into two separate terms, one accounting for the effect of the topology through the maximum degree and the longest causal path of the graph, and the other term captures the effect of the intervention space and the underlying data distributions. For establishing lower bounds, we construct two bandit instances that differ by just one edge but have very different expected rewards due to the aggregation of the difference in one edge through L layers.

Organization. The rest of the paper is organized as follows. Section 2 introduces the graphical model, intervention model, and the causal bandit model that is built on them. Performance measures and the quantities that will affect the regret guarantees are also defined. Section 3 outlines our approach to the problem, and establishes the estimation procedure. Section 4 develops an upper confidence bound-based algorithm and regret guarantees. Section 5 develops a Thompson sampling-

based algorithm and benefits from the results developed in Section 4 to present a similar regret result. Also, a more practical algorithm is proposed in Section 5 that is used in the numerical studies in Section 7. Section 6 presents a minimax lower bound on the regret which shows that the scaling behavior of our regret guarantees is necessary. The central pieces of the proofs are provided in the main body of the paper, and the rest of the proofs are relegated to the appendix.

2. Problem Setup

Notations. Vectors are denoted by upper-case letters, e.g., X, where X_i denotes the i-th element of vector X. Matrices are denoted by bold upper-case letters, e.g., $\mathbf{A}$, where $[\mathbf{A}]_i$ denotes the i-th column vector of $\mathbf{A}$ and $[\mathbf{A}]_{i,j}$ denotes the entry at row i and column j of $\mathbf{A}$. For a positive integer N, we define $[N] \triangleq \{1, \dots, N\}$. Sets and events are denoted by calligraphic letters. For a subset $\mathcal{S} \subseteq [N]$, we define $X_{\mathcal{S}} \triangleq X \odot \mathbf{1}(\mathcal{S})$, where $\odot$ denotes the Hadamard product and the N-dimensional binary vector $\mathbf{1}(\mathcal{S}) \in \{0, 1\}^N$ is specified such that its elements at the coordinates included in $\mathcal{S}$ are set to 1, and the rest are 0. Norm $\|\cdot\|$ denotes the Euclidean norm when applied on a vector and the spectral norm when applied to a matrix. For the latter, the spectral norm of a matrix is equal to the largest eigenvalue of the matrix. The $\mathbf{A}$-norm of a vector X associated with the positive semidefinite matrix $\mathbf{A}$ is denoted by $\|X\|_{\mathbf{A}} = \sqrt{X^\top \mathbf{A} X}$. We denote the singular values of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, where $M \geq N$, by

$$\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \cdots \geq \sigma_N(\mathbf{A}) \,. \tag{1}$$

In this paper, we often work with zero-padded vectors. As a result, the matrices that contain these vectors have non-trivial *null space* leading to zero singular values. In such cases, we use the *effective* smallest singular value that is non-zero. We denote the largest and smallest eigenvalues that correspond to effective dimensions of a positive semidefinite matrix $\mathbf{A}$ with rank k by

$$\sigma_{\max}(\mathbf{A}) \triangleq \sigma_1(\mathbf{A}) \,, \quad \text{and} \quad \sigma_{\min}(\mathbf{A}) \triangleq \sigma_k(\mathbf{A}) \,. \tag{2}$$

Similarly, for a square matrix of the form $\mathbf{V} = \mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{N \times N}$, we denote the largest and smallest eigenvalues by

$$\lambda_{\max}(\mathbf{V}) \triangleq \lambda_{\max}\left(\mathbf{A}\mathbf{A}^\top\right) = \sigma_{\max}^2(\mathbf{A}) \,, \tag{3}$$

$$\text{and} \quad \lambda_{\min}(\mathbf{V}) \triangleq \lambda_{\min}\left(\mathbf{A}\mathbf{A}^\top\right) = \sigma_{\min}^2(\mathbf{A}) \,. \tag{4}$$

Finally, the notation $\tilde{\mathcal{O}}$ is used to ignore constant and poly-logarithmic factors. We also have a table at Appendix A that summarizes commonly used notations throughout the paper.

Graphical model. We consider a directed acyclic graph with the known structure $\mathcal{G} \triangleq (\mathcal{V}, \mathcal{B})$, where $\mathcal{V} \triangleq [N]$ is the set of nodes, and $\mathcal{B}$ is the set of edges. The vector of random variables associated with the nodes is denoted by $X \triangleq (X_1, \dots, X_N)^\top$. We refer to the set of parents of node $i \in \mathcal{V}$ by $\mathrm{pa}(i)$. We denote the maximum degree of the graph by $d \triangleq \{\max_i |\mathrm{pa}(i)|\}$ and use L to denote the length of the longest directed path in the graph. For the causal model, we consider a linear SEM, according to which

$$X = \mathbf{H}^\top X + \nu + \epsilon \,, \tag{5}$$

6

where $\mathbf{H} \in \mathbb{R}^{N \times N}$ is the edge weights matrix, and it is strictly upper triangular, ν is the constant vector of unknown affine terms, ϵ accounts for the zero-mean model noise, and it is independent of X and ν. The term $\nu + \epsilon$ can be alternatively viewed as a noise term with an unknown mean value. In this model, $[\mathbf{H}]_i$ captures the weights of the causal effects of $\mathrm{pa}(i)$ on node i. Consequently, the edge weight $[\mathbf{H}]_{j,i}$ is non-zero if and only if $j \in \mathrm{pa}(i)$. The elements of the noise vector $\epsilon \triangleq (\epsilon_1, \ldots, \epsilon_N)^\top$ are statistically independent and are assumed to be 1-sub-Gaussian. The noise vector is assumed to satisfy $\|\epsilon\| \leq m_\epsilon$, where $m_\epsilon \in \mathbb{R}^+$ is an unknown constant.

Next, we re-arrange the terms in (5) such that the unknown terms, i.e., $\mathbf{H}$ and ν, can be presented more compactly. We create a *dummy* node $X_0 = 1$ and an associated dummy noise term $\epsilon_0 = 1$. Accordingly, we augment X and ϵ and redefine them as $X \triangleq (1, X_1, \ldots, X_N)^\top$ and $\epsilon \triangleq (1, \epsilon_1, \ldots, \epsilon_N)^\top$, respectively. Hence, for any node $i \in [N]$, the dummy node 0 acts as a parent with edge weight ν_i. Subsequently, we create the matrix $\mathbf{B} \in \mathbb{R}^{(N+1) \times (N+1)}$, in which we use 0 to refer to the index of the dummy node, i.e.,

$$[\mathbf{B}]_{i,j} = \begin{cases} [\mathbf{H}]_{i,j} & , \quad \forall i \in [N],\ \forall j \in [N]\,, \\ \nu_j & , \quad i = 0,\ \forall j \in [N]\,, \\ 0 & , \quad j = 0,\ \forall i \in [N] \cup \{0\}\,. \end{cases} \tag{6}$$

Hence, (5) becomes

$$X = \mathbf{B}^\top X + \epsilon\,. \tag{7}$$

Next, note that the constant $X_0 = 1$ acts as a parent for all nodes. Hence, we expand $\mathrm{pa}(i)$ with the dummy node and define

$$\overline{\mathrm{pa}}(i) \triangleq \mathrm{pa}(i) \cup \{0\}\,, \quad \forall i \in [N]\,. \tag{8}$$

Therefore, the effective maximum degree becomes $d + 1$, the effective longest path length becomes $L + 1$, and the weight between the dummy node 0 and node i is $[\mathbf{B}]_{0,i} = \nu_i$.

Soft intervention model. The conditional distribution of X_i given its parents is denoted by $\mathbb{P}(X_i | X_{\mathrm{pa}(i)})$. A *soft intervention on node i* refers to an action that induces a change in the conditional distribution $\mathbb{P}(X_i | X_{\mathrm{pa}(i)})$. An intervention can impact one or more nodes simultaneously, and we distinguish the intervention actions based on the set of nodes they impact. We denote the intervention space by $\mathcal{A} \triangleq 2^{\mathcal{V}}$.

To formalize the impact of our soft interventions, we model the effect of a soft intervention on node i as an alteration in $[\mathbf{B}]_i$ such that $[\mathbf{B}]_i$ changes to $[\mathbf{B}^*]_i \neq [\mathbf{B}]_i$. We define $\mathbf{B}^*$ as a weight matrix formed by columns $\{[\mathbf{B}^*]_i : i \in \mathcal{V}\}$. We refer to $\mathbf{B}$ and $\mathbf{B}^*$ as the observational and interventional weight matrices, respectively. We note that soft interventional changes can occur on either the weights $[\mathbf{H}]_i$ or the affine term ν_i. While the distribution of the noise variable ϵ_i can also change without affecting its mean, such a change would not affect the expected reward. Therefore, we only consider soft interventional changes that affect the expected reward, i.e., changes in $[\mathbf{H}]_i$ or the affine term ν_i. Since any change in the mean of ϵ_i can be attributed to a change in ν_i, ϵ_i can be assumed to remain zero-mean. Hence, our model covers all possible soft interventions on linear models that affect expected rewards.

Since an intervention can impact multiple nodes simultaneously, each intervention action imposes changes in multiple columns of $\mathbf{B}$. This leads to a distinct linear SEM model for each possible

intervention action $a \in \mathcal{A}$. To capture the impact of a specific intervention action $a \in \mathcal{A}$, we define $\mathbf{B}_a$ as its corresponding post-intervention weight matrix. $\mathbf{B}_a$ is constructed according to:

$$[\mathbf{B}_a]_i = \mathbb{1}_{\{i \in a\}}[\mathbf{B}^*]_i + \mathbb{1}_{\{i \notin a\}}[\mathbf{B}]_i . \tag{9}$$

We use $\mathbb{P}_a$ to denote the distribution of the random variables under intervention $a \in \mathcal{A}$ induced by linear SEMs in (7) with weight matrix $\mathbf{B}_a$. For given matrices $\mathbf{B}$ and $\mathbf{B}^*$ we define

$$m_B \triangleq \max_{i \in \mathcal{V}, a \in \mathcal{A}} \{\|[\mathbf{B}_a]_i\|\} . \tag{10}$$

Since the maximum degree of a node i is d and we have augmented the weight vectors with ν_i's, we have $\|[\mathbf{B}_a]_i\|_0 \leq d + 1$. The boundedness of $\|[\mathbf{B}_a]_i\|$ and $\|\epsilon\|$ in conjunction with the SEM in (7) imply that there exists a constant $m \in \mathbb{R}^+$ such that $\|X\| \leq m$. We formalize the boundedness assumptions as follows.

Assumption 1 (Boundedness) *For given matrices $\mathbf{B}$ and $\mathbf{B}^*$, $\max_{i \in \mathcal{V}, a \in \mathcal{A}}\{\|[\mathbf{B}_a]_i\|\} \leq m_B$. Furthermore, X satisfies $\|X\| \leq m$ for some known $m \in \mathbb{R}^+$.*

Causal bandit model. In the setting described, a learner performs a sequence of repeated interventions. Each intervention is represented by an arm, including the null intervention (i.e., a pure observation). The set of interventions can be abstracted by a multi-arm bandit setting with 2^N arms, one arm corresponding to one specific intervention. Following the related literature, e.g., (Lattimore et al., 2016; Sen et al., 2017; Lu et al., 2020; Nair et al., 2021; Yabe et al., 2018), we assume the structure of the graph $\mathcal{G}$ is given, and without loss of generality, we designate node N as the reward node and accordingly X_N as the reward variable. Similar to linear bandits, the reward X_N is a linear function of some other variables, namely the parents of the reward node, $X_{\overline{\mathrm{pa}}(N)}$. These parent variables $X_{\overline{\mathrm{pa}}(N)}$, in turn, depend on their causal ancestors according to the linear SEM in (7). Even though each parent contributes linearly to its immediate descendants, their compounding effects induce a non-linearity in the overall model such that X_N varies non-linearly with respect to the entries of $\mathbf{B}$ and $\mathbf{B}^*$. We denote the expected reward based on intervention (action) $a \in \mathcal{A}$ by

$$\mu_a \triangleq \mathbb{E}_a[X_N] , \tag{11}$$

where $\mathbb{E}_a$ denotes expectation under $\mathbb{P}_a$. Accordingly, we define the optimal action a^* as

$$a^* \triangleq \arg\max_{a \in \mathcal{A}} \mu_a . \tag{12}$$

The sequence of interventions over time is denoted by $\{a_t \in \mathcal{A} : t \in \mathbb{N}\}$. Upon intervention a_t in round t, the learner observes $X(t) \triangleq (1, X_1(t), \ldots, X_N(t))^\top \sim \mathbb{P}_{a_t}$ and collects the reward $X_N(t)$. We denote the random noise vector in round t by $\epsilon(t) \triangleq (1, \epsilon_1(t), \ldots, \epsilon_N(t))^\top$, which is independent of the intervention a_t. The learner does not know the observational or interventional matrices $\mathbf{B}$ and $\mathbf{B}^*$. More importantly, the interventional distributions of the parents of the reward variable, i.e., $\{\mathbb{P}_a(\overline{\mathrm{pa}}(N)) : a \in \mathcal{A}\}$, are unknown. This is in contrast to the earlier studies in (Lattimore et al., 2016; Sen et al., 2017; Lu et al., 2020; Nair et al., 2021), which assume that these distributions are known. Our setting is in line with those of the more recent studies in (Bilodeau et al., 2022; Lu et al., 2021; de Kroon et al., 2022), which dispense with knowing the distributions.

8

Let us denote the second moment of the parents of a node i under intervention $a \in \mathcal{A}$ by

$$\Sigma_{i,a} \triangleq \mathbb{E}_{X \sim \mathbb{P}_a} \left[X_{\overline{\mathrm{pa}}(i)} X_{\overline{\mathrm{pa}}(i)}^\top \right] . \tag{13}$$

This second-moment matrix is a function of intervention a, and the unknown weight matrices $\mathbf{B}$ and $\mathbf{B}^*$. Accordingly, we denote the lower and upper bounds on the minimum and maximum singular values of these moments by

$$\kappa_{i,\min} \triangleq \min_{a \in \mathcal{A}} \sigma_{\min} \left(\Sigma_{i,a} \right) , \quad \kappa_{\min} \triangleq \min_{i \in [N]} \kappa_{i,\min} , \tag{14}$$

$$\kappa_{i,\max} \triangleq \max_{a \in \mathcal{A}} \sigma_{\min} \left(\Sigma_{i,a} \right) , \quad \kappa_{\max} \triangleq \min_{i \in [N]} \kappa_{i,\max} . \tag{15}$$

We do not assume to know these moments. Note that having a zero singular value implies that there is a deterministic relationship among the elements of $X_{\overline{\mathrm{pa}}(i)}$. However, random variables of a causal model, such as the model described in this section, cannot have such a deterministic relationship. Furthermore, the addition of constant $X_0 = 1$ does not violate this property since it is the only non-random variable in vector X, and any random variable X_i still does not possess a deterministic relationship with the other elements of $X_{\overline{\mathrm{pa}}(i)}$. Hence, in our setting, we have $\kappa_{\min} > 0$. Furthermore, since $\|X\|^2 \leq m^2$, we have

$$0 < \kappa_{\min} \leq \kappa_{i,\min} \leq \kappa_{i,\max} \leq \kappa_{\max} \leq m^2 . \tag{16}$$

The learner's objective is to design a policy for sequentially selecting the interventions over time so that a measure of cumulative regret is minimized. For this purpose, we define $r(t) \triangleq \mu_{a^*} - \mu_{a_t}$ as the regret incurred at time t. In this paper, we consider the following two canonical expected cumulative regret measures.

1. **Frequentist regret:** By denoting the cumulative regret up to time T by $R(T) \triangleq \sum_{t=1}^T r(t)$, the expected cumulative regret is given by

$$\mathbb{E}[R(T)] \triangleq T\mu_{a^*} - \mathbb{E}\left[\sum_{t=1}^T X_N(t) \right] . \tag{17}$$

2. **Bayesian regret:** We define $\mathbf{W} \triangleq [\mathbf{B} \ \mathbf{B}^*] \in \mathbb{R}^{N \times 2N}$ to capture the entire parameterization of the observational and interventional distributions. We define the domain of $\mathbf{W}$ by $\mathcal{W} \triangleq \{\Theta \in R^{N \times 2N} : \|[\Theta]_i\| \leq 1, \ \forall i \in [2N]\}$. By denoting the cumulative regret associated with $\mathbf{W}$ up to time T by $R_{\mathbf{W}}(T)$, the Bayesian regret is given by

$$\mathrm{BR}(T) \triangleq \mathbb{E}_{\mathcal{W}} \mathbb{E}_\epsilon [R_{\mathbf{W}}(T)] , \tag{18}$$

where $\mathbb{E}_{\mathcal{W}}$ is the expectation with respect to the Bayesian prior over $\mathcal{W}$.

3. Causal Bandits Methodology

Before presenting the algorithms and performance guarantees, we note that despite similarities to the conventional bandit problems and the existing causal bandits' literature, treating our problem

necessitates a distinct approach to algorithm design and analysis. As discussed in Section 1, this is primarily due to the lack of information about the joint distribution of reward node's parents under different interventions $\{\mathbb{P}_a(\overline{\mathrm{pa}}(N)) : a \in \mathcal{A}\}$. Naively, the moments of these distributions can be estimated from the data. However, it becomes prohibitive even for moderate values of N, which gives rise to 2^N possible interventions.

We take a different approach and use the fact that reward variable X_N under intervention a is a function of the weight matrix $\mathbf{B}_a$ and the noise vector ϵ (which is independent of a). We show that the reward X_N is a linear function of the entries of ϵ, and the coefficients of the noise terms are non-linear functions of the entries of $\mathbf{B}_a$. Hence, causal relationships are captured by these coefficients. We compute these coefficients by solving each linear problem $X_i = X^\top [\mathbf{B}_a]_i + \epsilon_i$ separately, and combining the estimates for $[\mathbf{B}_a]_i$ according to the graph structure. We show how these steps work, propose algorithms, and provide their regret analysis in the rest of the paper.

Reward Modeling. In linear SEMs, each random variable X_i can be specified as a linear function of the exogenous noise variables ϵ via recursive substitution of the structural equations. This can easily be seen by rearranging (7) to obtain $X = (I_{N+1} - \mathbf{B}^\top)^{-1} \epsilon$. The second observation is that, $(I_{N+1} - \mathbf{B}^\top)^{-1}$ has a simple expansion since $\mathbf{B}$ is strictly upper triangular. Specifically,

$$\left(I_{N+1} - \mathbf{B}^\top \right)^{-1} = (I_{N+1} - \mathbf{B})^{-\top} \tag{19}$$

$$= \left(I_{N+1} + \sum_{i=1}^{\infty} \mathbf{B}^i \right)^\top \tag{20}$$

$$= \left(\sum_{\ell=0}^{L+1} \mathbf{B}^\ell \right)^\top , \tag{21}$$

where the last equality holds since B^ℓ becomes a zero matrix for $\ell \geq L + 2$. Finally, note that the entry $[\mathbf{B}^\ell]_{i,j}$ is the sum of the weighted products along ℓ-length directed paths from node j to node i. Hence, using (21), the multiplier of ϵ_i in the expansion of X_N has a simple description: the sum of the products of weights along a path that traces from an upstream node i and ends at the reward node N. The following lemma characterizes this linear relationship between the reward node X_N and the noise variables ϵ. The relationship is further simplified for the expected reward.

Lemma 1 *Consider the linear SEM associated with intervention a with weight matrix $\mathbf{B}_a$. The reward X_N is related to the noise vector ϵ via*

$$X_N = \sum_{\ell=0}^{L+1} \left\langle \left[\mathbf{B}_a^\ell\right]_N, \ \epsilon \right\rangle , \tag{22}$$

in which, L is defined as the length of the longest path in $\mathcal{G}$. Furthermore, since $\{\epsilon_i : i \in [N]\}$ have zero mean values and $\epsilon_0 = 1$, the expected reward under intervention a is

$$\mu_a = f(\mathbf{B}_a) \triangleq \sum_{\ell=1}^{L+1} \left[\mathbf{B}_a^\ell\right]_{0,N} . \tag{23}$$

Proof The first result immediately follows from (21) as follows:

$$X \stackrel{(7)}{=} (I_{N+1} - \mathbf{B}_a^\top)^{-1}\epsilon \stackrel{(21)}{=} \left(\sum_{\ell=0}^{L+1} \mathbf{B}_a^\ell\right)^\top \epsilon \,, \tag{24}$$

$$X_N = \sum_{\ell=0}^{L+1} \left\langle [\mathbf{B}_a^\ell]_N, \ \epsilon \right\rangle \,. \tag{25}$$

Note that ϵ and $\mathbf{B}_a$ are independent, the expectation of each ϵ_i is 0 for $i \in [N]$, and dummy noise constant $\epsilon_0 = 1$. Then, we obtain

$$\mu_a = \mathbb{E}[X_N] = \sum_{\ell=0}^{L+1} \mathbb{E}\left[\left\langle [\mathbf{B}_a^\ell]_N, \ \epsilon \right\rangle\right] \tag{26}$$

$$= \sum_{\ell=0}^{L+1} \sum_{i=0}^{N} \left([\mathbf{B}_a^\ell]_{i,N}\mathbb{E}[\epsilon_i]\right) \tag{27}$$

$$= \sum_{\ell=0}^{L+1} [\mathbf{B}_a^\ell]_{0,N} \,. \tag{28}$$

Note that $\mathbf{B}_a^0 = I_{N+1}$, and the summand for $\ell = 0$ in (28) is zero. Hence, by defining $f(\mathbf{B}_a) = \sum_{\ell=1}^{L+1} [\mathbf{B}_a^\ell]_{0,N}$, we obtain the desired result $\mu_a = f(\mathbf{B}_a)$. ∎

Lemma 1 indicates that, given $\mathbf{B}_a$, the expected reward can be computed from the sum of $L + 1$ components. Next, we focus on estimating the weight vectors, which are used for constructing the matrices $\{\mathbf{B}_a : a \in \mathcal{A}\}$.

Estimating parameter vectors under causal bandit model. We use the ordinary least-squares method to estimate vectors $\{[\mathbf{B}]_i, [\mathbf{B}^*]_i : i \in [N]\}$. For estimating $[\mathbf{B}]_i$ and $[\mathbf{B}^*]_i$, we should use the data samples from the rounds in which i is non-intervened, and intervened, respectively. Hence, our estimates at time t, which are denoted by $\{[\mathbf{B}(t)]_i, [\mathbf{B}^*(t)]_i : i \in [N]\}$, are computed according to:

$$[\mathbf{B}(t)]_i \triangleq [\mathbf{V}_i(t)]^{-1} \sum_{s \in [t]: i \notin a_s} X_{\overline{\mathrm{pa}}(i)}(s)X_i(s) \,, \tag{29}$$

$$\text{and} \qquad [\mathbf{B}^*(t)]_i \triangleq [\mathbf{V}_i^*(t)]^{-1} \sum_{s \in [t]: i \in a_s} X_{\overline{\mathrm{pa}}(i)}(s)X_i(s) \,, \tag{30}$$

where we have defined

$$\mathbf{V}_i(t) \triangleq \sum_{s \in [t]: i \notin a_s} X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s) + I_{N+1}, \tag{31}$$

$$\text{and} \qquad \mathbf{V}_i^*(t) \triangleq \sum_{s \in [t]: i \in a_s} X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s) + I_{N+1} \,. \tag{32}$$

Note that, first entry of $X_{\overline{\mathrm{pa}}(i)} = 1$ since dummy node 0 is contained in $\overline{\mathrm{pa}}(i) = \mathrm{pa}(i) \cup \{0\}$ for every node $i \in [N]$. Accordingly, we denote our estimate of $\mathbf{B}_a$ at time t by $\mathbf{B}_a(t)$. These estimates are formed similarly to (9) and according to:

$$[\mathbf{B}_a(t)]_i \triangleq \mathbb{1}_{\{i \in a\}}[\mathbf{B}^*(t)]_i + \mathbb{1}_{\{i \notin a\}}[\mathbf{B}(t)]_i \,. \tag{33}$$

11

Note that for estimating $[\mathbf{B}_a(t)]_i$, based on whether node i is contained in a, we use either the observational data (when $i \notin a$) or the interventional data (when $i \in a$). Furthermore, at time t, we construct two distinct $t \times (N+1)$ data matrices $\mathbf{D}_i(t)$ and $\mathbf{D}_i^*(t)$ from the observational and interventional data, respectively, to store the data from these two cases separately. Specifically, if node i is intervened at time $s \in [t]$, then the s-th row of $\mathbf{D}_i^*(t)$ stores $X_{\overline{\mathrm{pa}}(i)}^\top(s)$, and the s-th row of $\mathbf{D}_i(t)$ is a zero vector. This construction is reversed when i is not intervened at time s. Hence,

$$\left[\mathbf{D}_i^\top(t)\right]_s \triangleq \mathbb{1}_{\{i \notin a_s\}} X_{\overline{\mathrm{pa}}(i)}^\top(s) \,, \tag{34}$$

$$\text{and} \qquad \left[\mathbf{D}_i^{*\top}(t)\right]_s \triangleq \mathbb{1}_{\{i \in a_s\}} X_{\overline{\mathrm{pa}}(i)}^\top(s) \,. \tag{35}$$

Similarly to (9), we denote the relevant data and Gram matrices for node i under intervention a by

$$\mathbf{D}_{i,a}(t) \triangleq \mathbb{1}_{\{i \in a\}} \mathbf{D}_i^*(t) + \mathbb{1}_{\{i \notin a\}} \mathbf{D}_i(t) \,, \tag{36}$$

$$\text{and} \qquad \mathbf{V}_{i,a}(t) \triangleq \mathbb{1}_{\{i \in a\}} \mathbf{V}_i^*(t) + \mathbb{1}_{\{i \notin a\}} \mathbf{V}_i(t) \,. \tag{37}$$

The constructions in (34)-(37) yield

$$\mathbf{V}_{i,a}(t) = \mathbf{D}_{i,a}^\top(t) \mathbf{D}_{i,a}(t) + I_{N+1} \,. \tag{38}$$

Note that matrices $\mathbf{V}_i(t)$ and $\mathbf{V}_i^*(t)$ are positive definite with their smallest eigenvalue being at least 1 due to the regularization constant I_{N+1}. Hence, $\lambda_{\min}(\mathbf{V}_{i,a}(t)) \geq 1$.

Let $N_i^*(t)$ denote the number of times that node i is intervened up to time t, for each $i \in [N]$ and $t \in [T]$. Similarly, denote the number of times that i is not intervened by $N_i(t) = t - N_i^*(t)$. Formally,

$$N_i^*(t) \triangleq \sum_{s=1}^t \mathbb{1}_{\{i \in a_s\}} \,, \quad \text{and} \quad N_i(t) \triangleq \sum_{s=1}^t \mathbb{1}_{\{i \notin a_s\}} = t - N_i^*(t) \,. \tag{39}$$

Accordingly, for node i under intervention a we define

$$N_{i,a}(t) \triangleq \mathbb{1}_{\{i \in a\}} N_i^*(t) + \mathbb{1}_{\{i \notin a\}} N_i(t) \,, \tag{40}$$

and denote the estimation error at time t for matrix $\mathbf{B}_a$ and its columns by

$$\Delta_a(t) \triangleq \mathbf{B}_a(t) - \mathbf{B}_a \,, \quad \text{and} \quad [\Delta_a(t)]_i \triangleq [\mathbf{B}_a(t)]_i - [\mathbf{B}_a]_i \,, \qquad \forall i \in [N] \,. \tag{41}$$

Our analysis will show that (Lemma 1), regret analysis involves the powers of $\mathbf{B}_a(t)$ matrices. To get insight into the matrix powers, consider a node $j \in \mathcal{V} \setminus \overline{\mathrm{pa}}(i)$ that is not a parent of node i. By construction, vector $X_{\overline{\mathrm{pa}}(i)}$ has a zero at its j-th entry. Then, the j-th rows of $[\mathbf{V}_i(t)]^{-1}$ and $[\mathbf{V}_i^*(t)]^{-1}$ will consist of only zeros, except for their j-th entries, which are 1, accounting for the addition of I_{N+1}. Since $j \notin \overline{\mathrm{pa}}(i)$, we have $[X_{\overline{\mathrm{pa}}(i)}]_j(s) = 0$ and $[\mathbf{B}_a(t)]_{j,i} = 0$ based on (29). Similarly, $[\mathbf{B}^*]_i$ has non-zero entries at only entries $k \in \overline{\mathrm{pa}}(i)$, and $[\Delta_a(t)]_i$ will be at most $(d+1)$-sparse. We denote the estimation error of the ℓ-th power of $\mathbf{B}_a$ for each $\ell \in [L]$ and $i \in [N]$ by

$$\Delta_a^{(\ell)}(t) \triangleq \mathbf{B}_a^\ell(t) - \mathbf{B}_a^\ell \,, \quad \text{and} \quad \left[\Delta_a^{(\ell)}(t)\right]_i \triangleq \left[\mathbf{B}_a^\ell(t)\right]_i - \left[\mathbf{B}_a^\ell\right]_i \,. \tag{42}$$

Based on the approach described and the estimated quantities specified, in the next section, we present the main algorithm.

4. LinSEM-UCB Algorithm

Upper confidence bound (UCB)-based algorithms are effective in a wide range of bandit settings. Their general principle is to, sequentially and adaptively to the data, compute upper confidence bounds on the reward of each arm. A learner, subsequently, in each round selects the arm with the largest upper confidence bound. Confidence intervals for the estimated parameters are leveraged to compute these bounds. In our problem, we have $2N$ unknown weight vectors $\{[\mathbf{B}]_i, [\mathbf{B}^*]_i : i \in [N]\}$ that specify the weight matrices $\{\mathbf{B}_a : a \in \mathcal{A}\}$, which in turn, characterize the reward variables uniquely. Hence, we design a UCB-based algorithm by maintaining confidence intervals for these $2N$ vectors. We describe the details of the algorithm next.

Algorithm details. Algorithm 1 presents our main bandit algorithm referred to as **Lin**ear **SEM** **U**pper **C**onfidence **B**ound (LinSEM-UCB). As its inputs, it takes the graph structure $\mathcal{G}$ (set of parents for each node), action set $\mathcal{A}$, horizon T, and parameter β_T. We build the confidence intervals centered on the empirical estimates for observational and interventional weight vectors as follows:

$$\mathcal{C}_i(t) \triangleq \left\{ \theta \in \mathbb{R}^N : \|\theta\| \leq m_B, \|\theta - [\mathbf{B}(t-1)]_i\|_{\mathbf{V}_i(t-1)} \leq \beta_T \right\} , \tag{43}$$

$$\text{and} \quad \mathcal{C}_i^*(t) \triangleq \left\{ \theta \in \mathbb{R}^N : \|\theta\| \leq m_B, \|\theta - [\mathbf{B}^*(t-1)]_i\|_{\mathbf{V}_i^*(t-1)} \leq \beta_T \right\} , \tag{44}$$

where $\beta_T \in \mathbb{R}_+$ controls the size of the confidence intervals. Accordingly, we define the relevant confidence interval for node i under intervention a as

$$\mathcal{C}_{i,a}(t) \triangleq \mathbb{1}_{\{i \in a\}} \mathcal{C}_i^*(t) + \mathbb{1}_{\{i \notin a\}} \mathcal{C}_i(t) . \tag{45}$$

Based on these, we define the upper confidence bound for intervention a in round t as follows:

$$\mathrm{UCB}_a(t) \triangleq \begin{cases} \max_{\Theta \in \mathbb{R}^{N \times N}} & f(\Theta) \\ \text{s.t.} & [\Theta]_i \in \mathcal{C}_{i,a}(t) \quad \forall\, i \in [N] \end{cases} . \tag{46}$$

The LinSEM-UCB algorithm computes the upper confidence bounds in each round and plays the action that has the largest upper confidence bound. The estimates $\{[\mathbf{B}_a(t)]_i : i \in [N]\}$ are updated according to (29) and (30).

Regret analysis. The regret in round t depends on the closeness of the estimates $\{[\mathbf{B}(t)]_i, [\mathbf{B}^*(t)]_i : i \in [N]\}$ to the true parameters $\{[\mathbf{B}]_i, [\mathbf{B}^*]_i : i \in [N]\}$. The estimation errors will become smaller as more data samples are observed. Since β_T controls the size of the confidence sets, a proper choice of β_T would guarantee that the true parameters lie in the defined confidence intervals with high probability. Consider an intervention a, and suppose that constant β_T is chosen such that it ensures the closeness of the estimates of each vector $\{[\mathbf{B}_a(t)]_i : i \in [N]\}$ to the true parameters. Lemma 1 indicates that the expected reward μ_a is composed of $L + 1$ components, corresponding to the set of paths of length $\ell \in [L + 1]$. Therefore, the mismatch in the computed rewards can be decomposed into $L + 1$ components, where each term accounts for the estimation error of a certain path length. From (42) recall that the estimation error term corresponding to paths of length ℓ is denoted by $\Delta_a^{(\ell)}$. We do not form direct estimates for the error terms $\Delta_a^{(\ell)}$ and, subsequently, cannot directly find bounds on them. Hence, to find such bounds, we leverage the estimation error bounds of the individual estimation guarantees on weight vectors and then aggregate them according to (42). The

---

Algorithm 1 LinSEM-UCB

1: **Input:** Horizon T, causal graph $\mathcal{G}$, action set $\mathcal{A}$, parameter β_T.
2: **Initialization:** Initialize parameters for $2N$ linear problems:
3: $[\mathbf{B}(0)]_i = \mathbf{0}_{(N+1)\times 1}, [\mathbf{B}^*(0)]_i = \mathbf{0}_{(N+1)\times 1}, \ \forall i \in [N]$　　　$\triangleright$ initiliaze estimates for parameter vectors
4: $\mathbf{V}_i(0) = I_{N+1}, \ \mathbf{V}_i^*(0) = I_{N+1}, \ g_i(0) = \mathbf{0}_{(N+1)\times 1}, \ g_i^*(0) = \mathbf{0}_{(N+1)\times 1}$　　$\triangleright$ initiliaze auxiliary parameters
5: **for** $t = 1, 2, \ldots, T$ **do**
6:　　**for** $a \in \mathcal{A}$ **do**
7:　　　　Compute $\mathrm{UCB}_a(t)$ according to (46) .
8:　　**end for**
9:　　$a_t = \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_a(t)$　　　　　　　　$\triangleright$ select the action that maximizes UCB
10:　　Pull a_t, observe $X(t) = (1, X_1(t), \ldots, X_N(t))^\top$.
11:　　**for** $i \in \{1, \ldots, N\}$ **do**
12:　　　　**if** $i \in a_t$ **then**　　　　　　　　　　　$\triangleright$ update interventional parameters
13:　　　　　　$\mathbf{V}_i^*(t) = \mathbf{V}_i^*(t-1) + X_{\overline{\mathrm{pa}}(i)}(t)X_{\overline{\mathrm{pa}}(i)}^\top(t)$　and　$\mathbf{V}_i(t) = \mathbf{V}_i(t-1)$
14:　　　　　　$g_i^*(t) = g_i^*(t-1) + X_{\overline{\mathrm{pa}}(i)}(t)X_i(t)$　and　$g_i(t) = g_i(t-1)$
15:　　　　　　$[\mathbf{B}^*(t)]_i = [\mathbf{V}_i^*(t)]^{-1}g_i^*(t)$　and　$[\mathbf{B}(t)]_i = [\mathbf{B}(t-1)]_i$
16:　　　　**else**　　　　　　　　　　　　　　$\triangleright$ update observational parameters
17:　　　　　　$\mathbf{V}_i(t) = \mathbf{V}_i(t-1) + X_{\overline{\mathrm{pa}}(i)}(t)X_{\overline{\mathrm{pa}}(i)}^\top(t)$　　and　　$\mathbf{V}_i^*(t) = \mathbf{V}_i^*(t-1)$
18:　　　　　　$g_i(t) = g_i(t-1) + X_{\overline{\mathrm{pa}}(i)}(t)X_i(t)$　and　$g_i^*(t) = g_i^*(t-1)$
19:　　　　　　$[\mathbf{B}(t)]_i = [\mathbf{V}_i(t)]^{-1}g_i(t)$　and　$[\mathbf{B}^*(t)]_i = [\mathbf{B}^*(t-1)]_i$
20:　　　　**end if**
21:　　**end for**
22: **end for**

---

following lemma delineates a data-dependent bound on the estimation error $\Delta_a^{(\ell)}(t)$, and will be pivotal in our regret analysis.

Lemma 2 *If* $\|[\Delta_a(t)]_i\|_{\mathbf{V}_{i,a}(t)} \leq \beta_T$ *for all* $i \in [N]$ *and* $t \in [T]$, *then for all* $\ell \in [L+1]$ *we have*

$$\left\| \left[\Delta_a^{(\ell)}(t)\right]_N \right\|_{\mathbf{V}_{N,a}(t)} < (d+1)^{\frac{\ell-1}{2}}(\beta_T + m_B)^\ell \left[\lambda_{\max}^{1/2}\big(\mathbf{V}_{N,a}(t)\big) \max_{i \in [N]} \lambda_{\min}^{-1/2}\big(\mathbf{V}_{i,a}(t)\big) \right] . \quad (47)$$

Proof See Appendix B. ■

Besides the true parameters $\mathbf{B}_a$, and the estimated parameters $\mathbf{B}_a(t)$, there is one more set of parameters relevant for analyzing the performance of LinSEM-UCB, namely the parameters that attain the upper confidence bounds. The following corollary will be used in parallel to Lemma 2 for treating parameter estimations in our analysis.

Corollary 3 *For all* $\mathbf{A} \in \mathbb{R}^{N \times N}$ *define*

$$\Delta_\mathbf{A}(t) \triangleq \mathbf{A} - \mathbf{B}_a(t) , \quad and \quad \Delta_\mathbf{A}^{(\ell)}(t) \triangleq \mathbf{A}^\ell - \mathbf{B}_a^\ell(t) . \quad (48)$$

If $\forall i \in [N]$ *and* $\forall t \in [T]$, $\mathbf{A}$ *satisfies the following conditions:*

1. $\|[\mathbf{A}]_i\| \leq m_B$;

2. $\|[\mathbf{A}]_i\|_0 \leq d+1$;

3. $\|\Delta_{\mathbf{A}}(t)\|_0 \leq d+1$;

4. $\|[\Delta_{\mathbf{A}}(t)]_i\|_{\mathbf{V}_{i,a}(t)} \leq \beta_T$;

then for all $\ell \in [L]$ we have

$$\left\| [\Delta_{\mathbf{A}}^{(\ell)}(t)]_N \right\|_{\mathbf{V}_{N,a}(t)} < (d+1)^{\frac{\ell-1}{2}} (\beta_T + m_B)^{\ell} \left[\lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) \max_{i \in [N]} \lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) \right] . \quad (49)$$

Proof See Appendix B. ∎

For Lemma 2 to be useful in the regret analysis of LinSEM-UCB, we need to ensure that the conditions of its statement hold with high probability. We note that estimating $[\mathbf{B}_a]_i$ by regressing from $X_{\overline{\mathrm{pa}}(i)}$ to X_i is a linear problem. Then, the condition $\|[\Delta_a(t)]_i\|_{\mathbf{V}_{i,a}(t)} \leq \beta_T$ follows the form of a uniform confidence interval guarantee for the linear bandits, which has been studied extensively (Dani et al., 2008; Abbasi-Yadkori et al., 2011). Specifically, the results of Lattimore and Szepesvári (2020, Theorem 20.5) are relevant to our confidence intervals defined in (43) and (44). We adopt these results and modify them to account for the effect of including I_{N+1} (in the definition of $\mathbf{V}_{i,a}(t)$) as a regularizer for the least-squares estimator of $[\mathbf{B}_a]_i$. This is formalized in the next theorem.

Theorem 4 *Lattimore and Szepesvári (2020, Theorem 20.5) Let $\delta \in (0, 1)$. If $\|[\mathbf{B}_a]_i\| \leq m_B$ for a node i, then $\mathbb{P}(\exists t \in \mathbb{N} : [\mathbf{B}_a]_i \notin \mathcal{C}'_{i,a}(t)) \leq \delta$, where we have defined*

$$\mathcal{C}'_{i,a}(t) \triangleq \left\{ \theta \in \mathbb{R}^N : \|\theta - [\mathbf{B}_a(t-1)]_i\|_{\mathbf{V}_{i,a}(t-1)} \leq m_B + \sqrt{2\log(1/\delta) + \log(\det \mathbf{V}_{i,a}(t-1))} \right\} . \tag{50}$$

Note that the confidence set $\mathcal{C}'_{i,a}(t)$ resembles $\mathcal{C}_{i,a}(t)$ that we defined in (45). Since $\|[\mathbf{B}_a]_i\| \leq m_B$, the proper choice of β_T would make the both confidence sets equal.

Sketch of the result: We are ready to present the first part of the regret result. The events that do not satisfy the conditions of Lemma 2 will incur a constant term in the regret. Since ϵ is assumed to be 1-sub-Gaussian, we can use Lattimore and Szepesvári (2020, Theorem 20.5) to control the probability of such events. Subsequently, we will use Lemma 2 to analyze each of $L + 1$ components of the regret. The structure of the graph $\mathcal{G}$ will affect the regret through degree d and the longest causal path length L as implied by Lemma 2. Note that these results are for a time instance $t \in [T]$, but we aim to find bounds on the cumulative regret. Furthermore, the properties of the intervention space have not been accounted for yet. The term λ_T, which will be defined shortly, captures these effects and is analyzed in a separate result (Theorem 6).

Theorem 5 *Under Assumption 1, the regret of LinSEM-UCB is bounded by*

$$\mathbb{E}[R(T)] \leq 2m + 2(\beta_T + m_B)^{L+1}(d+1)^{\frac{L}{2}} \lambda_T , \tag{51}$$

where we have set

$$\beta_T = m_B + \sqrt{2\log(2NT) + (d+1)\log(1 + m^2 T/(d+1))} \,, \tag{52}$$

$$and \quad \lambda_T = \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\frac{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t)\right)}{\lambda_{\min}\left(\mathbf{V}_{N,a_t}(t)\right) \min_{i\in[N]} \lambda_{\min}\left(\mathbf{V}_{i,a_t}(t)\right)}}\right] \,. \tag{53}$$

Proof For the cumulative regret specified in (17) we have

$$\mathbb{E}[R(T)] = T\mu_{a^*} - \mathbb{E}\left[\sum_{t=1}^{T} X_N(t)\right] = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_{a^*} - \mu_{a_t})\right] \,. \tag{54}$$

From Lemma 1 we have $\mu_{a_t} = f(\mathbf{B}_{a_t})$. Therefore, $\mathbb{E}[R(T)]$ can be equivalently stated as

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\sum_{t=1}^{T}(f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t}))\right] \,. \tag{55}$$

Note that we were able to use Lemma 1 since $\epsilon(t)$ is independent of the data, which governs the choice of a_t. Since $\beta_T = m_B + \sqrt{2\log(2NT) + (d+1)\log(1 + m^2 T/(d+1))}$ is independent of $t \in [T]$, we use the shorthand β to replace it. Next, we define the events $\{\mathcal{E}_i, \mathcal{E}_i^* : \forall i \in [N]\}$ as

$$\mathcal{E}_i \triangleq \left\{\forall t \in [T] : \|[\mathbf{B}(t-1)]_i - [\mathbf{B}]_i\|_{\mathbf{V}_i(t-1)} \le \beta\right\} \,, \tag{56}$$

$$and \quad \mathcal{E}_i^* \triangleq \left\{\forall t \in [T] : \|[\mathbf{B}^*(t-1)]_i - [\mathbf{B}^*]_i\|_{\mathbf{V}_i^*(t-1)} \le \beta\right\} \,. \tag{57}$$

We will show that the specified choice of β ensures that the events $\{\mathcal{E}_i, \mathcal{E}_i^* : \forall i \in [N]\}$ hold with a high probability. In other words, the confidence intervals of UCB contain the true parameters with high probability. To this end, we first bound $\mathbf{V}_i(t)$. Since $\mathbf{V}_i(t)$ is a positive definite matrix, we can use the arithmetic-geometric mean inequality (AM-GM) to upper bound its determinant through its trace. Furthermore, since $\|X\| \le m$, the trace of $V_i(t)$ will be upper bounded by $d_i + Tm^2$. Therefore, we have

$$\det \mathbf{V}_i(t) \overset{(\text{AM}-\text{GM})}{\le} \left(\frac{1}{d_i}\text{tr}\left(\mathbf{V}_i(t)\right)\right)^{d_i} \le \left(1 + \frac{Tm^2}{d_i}\right)^{d_i} \le \left(1 + \frac{Tm^2}{d}\right)^{d} \,. \tag{58}$$

By noting that the norms of vectors $\{[\mathbf{B}]_i, [\mathbf{B}^*]_i : \forall i \in [N]\}$ are bounded by m_B (Assumption 1), and setting $\delta = \frac{1}{2NT}$, Lattimore and Szepesvári (2020, Theorem 20.5) yields

$$\mathbb{P}(\mathcal{E}_i^c) \le \frac{1}{2NT} \,, \quad \mathbb{P}(\mathcal{E}_i^{*c}) \le \frac{1}{2NT} \,, \quad \forall i \in [N] \,. \tag{59}$$

Let $\mathcal{E}_\cap$ denote the event that all of the events $\{\mathcal{E}_i, \mathcal{E}_i^* : i \in [N]\}$ occur simultaneously, i.e.,

$$\mathcal{E}_\cap \triangleq \left(\bigcap_{i=1}^{N} \mathcal{E}_i\right) \cap \left(\bigcap_{i=1}^{N} \mathcal{E}_i^*\right) \,. \tag{60}$$

16

By invoking the union bound we have

$$\mathbb{P}\left(\mathcal{E}_{\cap}^{\mathrm{c}}\right) \leq \sum_{i=1}^{N} \mathbb{P}(\mathcal{E}_i^{\mathrm{c}}) + \sum_{i=1}^{N} \mathbb{P}(\mathcal{E}_i^{*\mathrm{c}}) \overset{(59)}{\leq} \sum_{i=1}^{N} \left(\frac{1}{2NT} + \frac{1}{2NT} \right) = \frac{1}{T} . \tag{61}$$

Next, we analyze the regret under the complementary events $\mathcal{E}_{\cap}$ and $\mathcal{E}_{\cap}^{\mathrm{c}}$. Note that the regret at any time t can be at most $2m$ since $|X_N| \leq \|X\| \leq m$. Therefore, for the expected regret we have

$$\mathbb{E}[R(T)] = \mathbb{E}\left[\sum_{t=1}^{T} (f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t})) \right] \tag{62}$$

$$= \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cap}^{\mathrm{c}}} \sum_{t=1}^{T} \underbrace{(f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t}))}_{\leq 2m} \right] + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cap}} \sum_{t=1}^{T} (f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t})) \right] \tag{63}$$

$$\leq 2mT\mathbb{P}(\mathcal{E}_{\cap}^{\mathrm{c}}) + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cap}} \sum_{t=1}^{T} (f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t})) \right] \tag{64}$$

$$\overset{(61)}{\leq} 2m + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cap}} \sum_{t=1}^{T} (f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t})) \right] . \tag{65}$$

The algorithm selects $a_t = \arg\max_{a \in \mathcal{A}} \mathrm{UCB}_a(t)$ in round t. Let $\tilde{\mathbf{B}}_a$ denote the parameters that attain $\mathrm{UCB}_a(t)$, i.e., $f(\tilde{\mathbf{B}}_a) = \mathrm{UCB}_a(t)$. By definition, $\mathrm{UCB}_{a^*}(t) \leq \mathrm{UCB}_{a_t}(t)$. Under the event $\mathcal{E}_{\cap}$, we have

$$f(\mathbf{B}_{a^*}) \leq \mathrm{UCB}_{a^*}(t) \leq \mathrm{UCB}_{a_t}(t) = f(\tilde{\mathbf{B}}_{a_t}) , \tag{66}$$

$$\text{and} \qquad f(\mathbf{B}_{a^*}) - f(\mathbf{B}_{a_t}) \leq f(\tilde{\mathbf{B}}_{a_t}) - f(\mathbf{B}_{a_t}) . \tag{67}$$

For the term $f(\tilde{\mathbf{B}}_{a_t}) - f(\mathbf{B}_{a_t})$, based on the definition of f in Lemma 1, by applying the Cauchy-Schwarz (CS) inequality we have

$$f(\tilde{\mathbf{B}}_{a_t}) - f(\mathbf{B}_{a_t}) = \sum_{\ell=1}^{L+1} \left(\left[\tilde{\mathbf{B}}_{a_t}^{\ell} \right]_{0,N} - \left[\mathbf{B}_{a_t}^{\ell} \right]_{0,N} \right) \tag{68}$$

$$\leq \sum_{\ell=1}^{L+1} \left\| \left[\tilde{\mathbf{B}}_{a_t}^{\ell} \right]_{N} - \left[\mathbf{B}_{a_t}^{\ell} \right]_{N} \right\| , \tag{69}$$

$$\overset{(CS)}{\leq} \sum_{\ell=1}^{L+1} \left\| \left[\tilde{\mathbf{B}}_{a_t}^{\ell} \right]_{N} - \left[\mathbf{B}_{a_t}^{\ell} \right]_{N} \right\|_{\mathbf{V}_{N,a_t}(t)} \lambda_{\min}^{-1/2} \left(\mathbf{V}_{N,a_t}(t) \right) . \tag{70}$$

Note that $[\mathbf{B}_{a_t}(t)]_N$ is an estimate of $[\mathbf{B}_{a_t}]_N$, and $[\tilde{\mathbf{B}}_{a_t}]_N$ lies in the confidence interval that is centered on $[\mathbf{B}_{a_t}(t)]_N$. We decompose $\left[\tilde{\mathbf{B}}_{a_t}^{\ell} \right]_N - \left[\mathbf{B}_{a_t}^{\ell} \right]_N$ into two parts by adding and subtracting the term $\left[\mathbf{B}_{a_t}^{\ell}(t) \right]_N$ as follows

$$\left[\tilde{\mathbf{B}}_{a_t}^{\ell} \right]_N - \left[\mathbf{B}_{a_t}^{\ell} \right]_N = \left(\left[\tilde{\mathbf{B}}_{a_t}^{\ell} \right]_N - \left[\mathbf{B}_{a_t}^{\ell}(t) \right]_N \right) + \left(\left[\mathbf{B}_{a_t}^{\ell}(t) \right]_N - \left[\mathbf{B}_{a_t}^{\ell} \right]_N \right) , \tag{71}$$

17

and due to the triangle inequality, we have

$$\left\| \left[\tilde{\mathbf{B}}_{a_t}^\ell \right]_N - \left[\mathbf{B}_{a_t}^\ell \right]_N \right\|_{\mathbf{V}_{N,a_t}(t)} \tag{72}$$

$$\leq \left\| \left[\tilde{\mathbf{B}}_{a_t}^\ell \right]_N - \left[\mathbf{B}_{a_t}^\ell(t) \right]_N \right\|_{\mathbf{V}_{N,a_t}(t)} + \left\| \left[\mathbf{B}_{a_t}^\ell(t) \right]_N - \left[\mathbf{B}_{a_t}^\ell \right]_N \right\|_{\mathbf{V}_{N,a_t}(t)} . \tag{73}$$

Using (73) in the right-hand side of (65) we obtain

$$\mathbb{E}[R(T)] \overset{(67)}{\leq} 2m + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap} \sum_{t=1}^T (f(\tilde{\mathbf{B}}_{a_t}) - f(\mathbf{B}_{a_t})) \right] \tag{74}$$

$$\overset{(73)}{\leq} 2m + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap} \sum_{t=1}^T \sum_{\ell=1}^{L+1} \left\| \left[\tilde{\mathbf{B}}_{a_t}^\ell \right]_N - \left[\mathbf{B}_{a_t}^\ell(t) \right]_N \right\|_{\mathbf{V}_{N,a_t}(t)} \lambda_{\min}^{-1/2}\left(\mathbf{V}_{N,a_t}(t) \right) \right] \tag{75}$$

$$+ \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap} \sum_{t=1}^T \sum_{\ell=1}^{L+1} \left\| \left[\mathbf{B}_{a_t}^\ell(t) \right]_N - \left[\mathbf{B}_{a_t}^\ell \right]_N \right\|_{\mathbf{V}_{N,a_t}(t)} \lambda_{\min}^{-1/2}\left(\mathbf{V}_{N,a_t}(t) \right) \right] . \tag{76}$$

Under the event $\mathcal{E}_\cap$, the conditions of Lemma 2 are satisfied for matrices $\Delta_a^{(\ell)}(t)$ and $\mathbf{B}_{a_t}$. Similarly, the conditions of Corollary 3 are satisfied for matrices $\tilde{\mathbf{B}}_{a_t}^\ell - \mathbf{B}_{a_t}^\ell(t)$, and $\tilde{\mathbf{B}}_{a_t}$. Therefore, by applying Lemma 2 to each term in (76), and Corollary 3 to each term in (75), we obtain

$$\mathbb{E}[R(T)] \leq 2m + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap} \sum_{t=1}^T \sqrt{\frac{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t) \right)}{\lambda_{\min}\left(\mathbf{V}_{N,a_t}(t) \right) \min_{i\in[N]} \lambda_{\min}\left(\mathbf{V}_{i,a_t}(t) \right)}} \right.$$
$$\left. \times 2 \sum_{\ell=1}^{L+1} (d+1)^{\frac{\ell-1}{2}} (\beta + m_B)^\ell \right] \tag{77}$$

$$\leq 2m + \mathbb{E}\left[\sum_{t=1}^T \sqrt{\frac{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t) \right)}{\lambda_{\min}\left(\mathbf{V}_{N,a_t}(t) \right) \min_{i\in[N]} \lambda_{\min}\left(\mathbf{V}_{i,a_t}(t) \right)}} \right.$$
$$\left. \times 2 \sum_{\ell=1}^{L+1} (d+1)^{\frac{\ell-1}{2}} (\beta + m_B)^\ell \right] \tag{78}$$

$$= 2m + 2\lambda_T \frac{1}{\sqrt{d+1}} \sum_{\ell=1}^{L+1} ((\beta + m_B)\sqrt{d+1})^\ell , \tag{79}$$

in which, $\lambda_T \triangleq \mathbb{E}\left[\sum_{t=1}^T \sqrt{\frac{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t) \right)}{\lambda_{\min}\left(\mathbf{V}_{N,a_t}(t) \right) \min_{i\in[N]} \lambda_{\min}\left(\mathbf{V}_{i,a_t}(t) \right)}} \right]$. Note that, for $c \geq 1$,

$$\sum_{\ell=1}^{L+1} c^\ell = \frac{c^{L+2} - 1}{c - 1} - 1 \leq 2c^{L+1} . \tag{80}$$

Since $(\beta + m_B)\sqrt{d+1} \overset{(52)}{>} \sqrt{2}\sqrt{2\log 2} > 1$, by using (80) in (79) we obtain

$$\mathbb{E}[R(T)] \leq 2m + \frac{2\lambda_T}{\sqrt{d+1}} ((\beta + m_B)\sqrt{d+1})^{L+1} \tag{81}$$

$$= 2m + 2\lambda_T (\beta + m_B)^{L+1} (d+1)^{\frac{L}{2}} . \tag{82}$$

18

Next, we analyze the λ_T term in (53), which in conjunction with Theorem 5 characterizes our desired regret bound.

Theorem 6 *Under Assumption 1, λ_T specified as*

$$\lambda_T \triangleq \mathbb{E}\left[\sum_{t=1}^{T} \sqrt{\frac{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t)\right)}{\lambda_{\min}\left(\mathbf{V}_{N,a_t}(t)\right)\min_{i\in[N]}\lambda_{\min}\left(\mathbf{V}_{i,a_t}(t)\right)}}\right] , \tag{83}$$

is bounded according to

$$\lambda_T < \frac{4g(\tau)}{\sqrt{\kappa_{\min}}}\sqrt{NT} + 2\sqrt{2}(N+1)\tau g(\tau) + \frac{2\sqrt{2}N\sqrt{\tau}g(\tau)}{\sqrt{\kappa_{\min}}}\log\left(\frac{T}{2N}\right) + \frac{m}{T} + \frac{2m}{3} + 1 , \tag{84}$$

where $\tau = \frac{\alpha^2 m^4}{\kappa_{\min}^2}$, $\alpha = \sqrt{\frac{16}{3}\log((d+1)NT^{T/2}(T+1))}$, and $g(\tau) = \sqrt{2}(\sqrt{\tau\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} + 1)$. Furthermore, since $\alpha = \mathcal{O}(\sqrt{\log(T)}), \tau = \mathcal{O}(\log(T))$, and $g(\tau) = \mathcal{O}(\sqrt{\log(T)})$, we can write

$$\lambda_T = K_1\sqrt{NT} + K_2(\log(T))^2 + K_3 , \tag{85}$$

where $K_1 = \frac{4g(\tau)}{\sqrt{\kappa_{\min}}}$, and K_2 and K_3 are constants independent of T.

Proof We start by simplifying the notation for the quantity to bound. Note that α_T is a function of T and independent of a given $t \in [T]$. For simplicity, we use α as a shorthand for α_T. We also define $C_N(t)$ and $S(t)$ to compactly express λ_T as follows

$$C_N(t) \triangleq \sqrt{\frac{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t)\right)}{\lambda_{\min}\left(\mathbf{V}_{N,a_t}(t)\right)}} \overset{(38)}{=} \sqrt{\frac{\sigma_{\max}^2\left(\mathbf{D}_{N,a_t}(t)\right)+1}{\sigma_{\min}^2\left(\mathbf{D}_{N,a_t}(t)\right)+1}} , \tag{86}$$

$$S(t) \triangleq \frac{1}{\sqrt{\min_{i\in[N]}\lambda_{\min}\left(\mathbf{V}_{i,a_t}(t)\right)}} \overset{(38)}{=} \frac{1}{\sqrt{\min_{i\in[N]}\sigma_{\min}^2\left(\mathbf{D}_{i,a_t}(t)\right)+1}} , \tag{87}$$

based on which we have,

$$\lambda_T = \mathbb{E}\left[\sum_{t=1}^{T} C_N(t)S(t)\right] . \tag{88}$$

For bounding $\mathbb{E}\left[\sum_{t=1}^{T} C_N(t)S(t)\right]$, we will use upper and lower bounds for the maximum and minimum singular values of $\mathbf{D}_{i,a_t}(t)$. However, such bounds depend on the number of non-zero rows of $\mathbf{D}_{i,a_t}(t)$ matrices, which equals to values of the random variable $N_{i,a_t}(t)$. To start, define the constants

$$\varepsilon_n \triangleq \max\left\{\alpha m^2\sqrt{n}, \alpha^2 m^2\right\} , \quad \forall n \in [T] . \tag{89}$$

19

Then, for each triplet of $i \in [N], t \in [T]$, and $n \in [t]$, we define the error events $\mathcal{E}_{i,n}(t), \mathcal{E}_{i,n}^*(t)$ as:

$$\mathcal{E}_{i,n}(t) \triangleq \left\{ N_i(t) = n \quad \text{and} \right.$$
$$\left. \left\{ \sigma_{\min}(\mathbf{D}_i(t)) \le \max\left\{0, \sqrt{n\kappa_{\min}} - \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\} \text{ or } \sigma_{\max}(\mathbf{D}_i(t)) \ge \sqrt{n\kappa_{\max}} + \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\} \right\},$$
$$(90)$$

$$\mathcal{E}_{i,n}^*(t) \triangleq \left\{ N_i^*(t) = n \quad \text{and} \right.$$
$$\left. \left\{ \sigma_{\min}(\mathbf{D}_i^*(t)) \le \max\left\{0, \sqrt{n\kappa_{\min}} - \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\} \text{ or } \sigma_{\max}(\mathbf{D}_i^*(t)) \ge \sqrt{n\kappa_{\max}} + \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\} \right\}.$$
$$(91)$$

In other words, the event $\mathcal{E}_{i,n}(t)$ specifies the condition under which at least one of the terms $\sigma_{\min}(\mathbf{D}_i(t))$ and $\sigma_{\max}(\mathbf{D}_i(t))$ does not conform the lower and upper bounds that we construct. $\mathcal{E}_{i,n}^*(t)$ has the counterpart implications for singular values of $\mathbf{D}_i^*(t)$. The next result shows that events $\mathcal{E}_{i,n}(t)$ and $\mathcal{E}_{i,n}^*(t)$ occur with low probability.

Lemma 7 *The probability of the events $\mathcal{E}_{i,n}(t)$ and $\mathcal{E}_{i,n}^*(t)$ defined in* (90) *and* (91) *are upper bounded as*

$$\mathbb{P}(\mathcal{E}_{i,n}(t)) \le (d+1)\exp\left(-\frac{3\alpha^2}{16}\right) , \tag{92}$$

$$\text{and} \qquad \mathbb{P}(\mathcal{E}_{i,n}^*(t)) \le (d+1)\exp\left(-\frac{3\alpha^2}{16}\right) . \tag{93}$$

Proof We will prove the analysis for bounding $\mathbb{P}(\mathcal{E}_{i,n}(t))$ and analysis for $\mathbb{P}(\mathcal{E}_{i,n}^*(t))$ follows similarly. The core of the proof is using Freedman's concentration inequality for matrix martingales. We define the martingale sequence $\mathbf{Y}_i(k)$, with difference sequence $\mathbf{Z}_i(k)$, and the predictable quadratic variation of the process $\mathbf{W}_i(k)$ as follows

$$\mathbf{Z}_i(s) \triangleq \mathbb{1}_{\{i \notin a_s\}} \left(X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s) - \Sigma_{i,a_s} \right) , \quad \forall s \in [T] , \tag{94}$$

$$\mathbf{Y}_i(k) \triangleq \sum_{s=1}^k \mathbf{Z}_i(s) , \quad \forall k \in [T] , \tag{95}$$

$$\text{and} \qquad \mathbf{W}_i(k) \triangleq \sum_{s=1}^k \mathbb{E}[\mathbf{Z}_i^2(s) \mid \mathcal{F}_{s-1}] , \quad \forall k \in [T] , \tag{96}$$

where $\mathcal{F}_{s-1} \triangleq \sigma(a_1, X(1), \ldots, a_{s-1}, X(s-1), a_s)$. Under the event $\mathcal{E}_{i,n}(t)$ we have $N_i(t) = n$. We will show that, given $N_i(t) = n$, we have $\|\mathbf{W}_i(t)\| \le 2m^4 n$. Subsequently, given the event $\mathcal{E}_{i,n}(t)$, we will show that $\sigma_{\max}(\mathbf{Y}_i(t)) \ge \varepsilon_n$. The probability of these two events occurring together will be bounded by the matrix Freedman inequality. Finally, $\mathbb{P}(\mathcal{E}_{i,n}(t))$ will be upper bounded by the same probability. Detailed analysis is provided in Appendix B. ∎

Now that we have bounds on the probability of error events $\mathcal{E}_{i,n}(t)$ and $\mathcal{E}_{i,n}^*(t)$, we define the union error event $\mathcal{E}_\cup$ as

$$\mathcal{E}_\cup \triangleq \{\exists\, (i, t, n) : i \in [N], t \in [T], n \in [t],\ \mathcal{E}_{i,n}(t) \text{ or } \mathcal{E}_{i,n}^*(t)\}\,. \tag{97}$$

By taking a union bound and using Lemma 7 we have

$$\mathbb{P}(\mathcal{E}_\cup) \leq \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{n=1}^{t} (\mathbb{P}(\mathcal{E}_{i,n}(t)) + \mathbb{P}(\mathcal{E}_{i,n}^*(t))) \tag{98}$$

$$\leq NT(T+1)(d+1) \exp\left(-\frac{3\alpha^2}{16}\right) \tag{99}$$

Now we turn back to $\mathbb{E}\left[\sum_{t=1}^{T} C_N(t) S(t)\right]$ to analyze it under the complementary events $\mathcal{E}_\cup$ and $\mathcal{E}_\cup^c$.

$$\mathbb{E}\left[\sum_{t=1}^{T} C_N(t) S(t)\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup} \sum_{t=1}^{T} C_N(t) S(t)\right] + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup^c} \sum_{t=1}^{T} C_N(t) S(t)\right]\,. \tag{100}$$

Analyzing the second term will be more involved. Let us start with the first one.

Bounding $\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup} \sum_{t=1}^{T} C_N(t) S(t)\right]$. Since $\lambda_{\min}(\mathbf{V}_{i,a_t}(t)) \geq 1$, we have the following unconditional upper bound

$$C_N(t) S(t) = \sqrt{\frac{\lambda_{\max}(\mathbf{V}_{N,a_t}(t))}{\lambda_{\min}(\mathbf{V}_{N,a_t}(t))}} \cdot \frac{1}{\sqrt{\min_{i \in [N]} \lambda_{\min}(\mathbf{V}_{i,a_t}(t))}} \leq \sqrt{\lambda_{\max}(\mathbf{V}_{N,a_t}(t))}\,. \tag{101}$$

For finding an unconditional upper bound on $\lambda_{\max}(\mathbf{V}_{N,a_t}(t))$, we leverage $\|X\| \leq m$ as follows

$$\lambda_{\max}(\mathbf{V}_{N,a_t}(t)) = \lambda_{\max}\left(I_{N+1} + \sum_{s=1}^{t} \mathbb{1}_{\{N \in a_s\}} X_{\overline{\text{pa}}(i)}(s) X_{\overline{\text{pa}}(i)}^\top(s)\right) \tag{102}$$

$$\leq 1 + \sum_{s=1}^{t} \mathbb{1}_{\{N \in a_s\}} \lambda_{\max}\left(X_{\overline{\text{pa}}(i)}(s) X_{\overline{\text{pa}}(i)}^\top(s)\right) \tag{103}$$

$$\leq 1 + \sum_{s=1}^{t} \lambda_{\max}\left(X_{\overline{\text{pa}}(i)}(s) X_{\overline{\text{pa}}(i)}^\top(s)\right) \tag{104}$$

$$= 1 + \sum_{s=1}^{t} \left\|X_{\overline{\text{pa}}(i)}(s)\right\|^2 \tag{105}$$

$$\leq m^2 t + 1\,. \tag{106}$$

Therefore, the desired quantity is bounded by

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup}\sum_{t=1}^T C_N(t)S(t)\right] \overset{(101)}{\leq} \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup}\sum_{t=1}^T \sqrt{\lambda_{\max}\left(\mathbf{V}_{N,a_t}(t)\right)}\right] \tag{107}$$

$$\overset{(106)}{\leq} \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup}\sum_{t=1}^T \sqrt{m^2 t + 1}\right] \tag{108}$$

$$= \mathbb{E}[\mathbb{1}_{\mathcal{E}_\cup}]\sum_{t=1}^T \sqrt{m^2 t + 1} \tag{109}$$

$$= \mathbb{P}(\mathcal{E}_\cup)\sum_{t=1}^T \sqrt{m^2 t + 1}\,. \tag{110}$$

We have derived a bound for $\mathbb{P}(\mathcal{E}_\cup)$ at (99). The sum term is bounded as

$$\sum_{t=1}^T \sqrt{m^2 t + 1} \leq (m\sqrt{T} + 1) + \sum_{t=1}^{T-1}(m\sqrt{t} + 1) \tag{111}$$

$$\leq m\sqrt{T} + T + \int_{t=1}^T m\sqrt{t}\,dt \tag{112}$$

$$= m\sqrt{T} + T + \frac{2m}{3}(T^{3/2} - 1)\,. \tag{113}$$

By setting $\alpha = \sqrt{\frac{16}{3}\log((d+1)NT^{5/2}(T+1))}$, we obtain

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup}\sum_{t=1}^T \sqrt{m^2 t + 1}\right] \overset{(110)}{\leq} \mathbb{P}(\mathcal{E}_\cup)\sum_{t=1}^T \sqrt{m^2 t + 1} \tag{114}$$

$$\overset{(99)}{\leq} \underbrace{\frac{NT(T+1)(d+1)}{\exp(\log((d+1)NT^{5/2}(T+1)))}}_{=T^{-3/2}}\sum_{t=1}^T \sqrt{m^2 t + 1} \tag{115}$$

$$\overset{(113)}{\leq} T^{-3/2}\left(m\sqrt{T} + T + \frac{2m}{3}(T^{3/2} - 1)\right) \tag{116}$$

$$\leq \frac{m}{T} + \frac{2m}{3} + 1\,. \tag{117}$$

Bounding $\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cup^c}\sum_{t=1}^T C_N(t)S(t)\right]$. Given the event $\mathcal{E}_\cup^c$, all the events $\{\mathcal{E}_{i,n}^c(t), \mathcal{E}_{i,n}^{*\,c}(t) : i \in [N], t \in [T], n \in [t]\}$ hold. Therefore, we can use the following bounds on the singular values

$$\sigma_{\max}\left(\mathbf{D}_{i,a_t}(t)\right) \leq \sqrt{N_{i,a_t}(t)\kappa_{\max}} + \frac{\alpha m^2}{\sqrt{\kappa_{\min}}}\max\left\{1, \frac{\alpha}{\sqrt{N_{i,a_t}(t)}}\right\}, \tag{118}$$

$$\sigma_{\min}\left(\mathbf{D}_{i,a_t}(t)\right) \geq \max\left\{0, \sqrt{N_{i,a_t}(t)\kappa_{\min}} - \frac{\alpha m^2}{\sqrt{\kappa_{\min}}}\max\left\{1, \frac{\alpha}{\sqrt{N_{i,a_t}(t)}}\right\}\right\}\,. \tag{119}$$

Note that for values of $N_{i,a_t}(t)$ that are smaller than a certain threshold, the right-hand side of (119) becomes zero. The threshold above which this lower bound becomes non-zero will be critical in the following steps. Hence, we define the constant

$$\tau \triangleq \frac{\alpha^2 m^4}{\kappa_{\min}^2} \, . \tag{120}$$

When $N_{i,a_t}(t) \geq \tau$, we have $\sqrt{N_{i,a_t}(t)} \geq \frac{\alpha m^2}{\kappa_{\min}} \geq \alpha$ since $\kappa_{\min} \leq m^2$, in which case (119) reduces to

$$\sigma_{\min}\left(\mathbf{D}_{i,a_t}(t)\right) \geq \max\left\{0, \sqrt{N_{i,a_t}(t)\kappa_{\min}} - \sqrt{\tau\kappa_{\min}}\right\} \, . \tag{121}$$

To facilitate the analysis, we dispense with the square-root terms by using the following bounds

$$C_N(t) \overset{(86)}{=} \sqrt{\frac{\sigma_{\max}^2\left(\mathbf{D}_{N,a_t}(t)\right) + 1}{\sigma_{\min}^2\left(\mathbf{D}_{N,a_t}(t)\right) + 1}} \leq \sqrt{2}\frac{\sigma_{\max}\left(\mathbf{D}_{N,a_t}(t)\right) + 1}{\sigma_{\min}\left(\mathbf{D}_{N,a_t}(t)\right) + 1} \, , \tag{122}$$

$$S(t) \overset{(87)}{=} \frac{1}{\sqrt{\min_{i\in[N]} \sigma_{\min}^2\left(\mathbf{D}_{i,a_t}(t)\right) + 1}} \leq \frac{\sqrt{2}}{\min_{i\in[N]} \sigma_{\min}\left(\mathbf{D}_{i,a_t}(t)\right) + 1} \, . \tag{123}$$

Note that (122) follows from $\frac{x^2+1}{y^2+1} \leq 2(\frac{x+1}{y+1})^2$ when $x \geq y \geq 0$, and (123) follows from $2(x^2+1) > (x+1)^2$. Next, we define the following two functions of $x \in \mathbb{R}^+$:

$$g(x) \triangleq \sqrt{2}\frac{\sqrt{x\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} \max\left\{1, \frac{\alpha}{\sqrt{x}}\right\} + 1}{\max\left\{0, \sqrt{x\kappa_{\min}} - \sqrt{\tau\kappa_{\min}}\right\} + 1} \, , \tag{124}$$

$$h(x) \triangleq \frac{\sqrt{2}}{\max\left\{0, \sqrt{x\kappa_{\min}} - \sqrt{\tau\kappa_{\min}}\right\} + 1} \, . \tag{125}$$

Given the event $\mathcal{E}_{\cup}^{\mathrm{c}}$, we bound $C_N(t)$ and $S(t)$ in terms of the newly defined g and h functions as

$$\mathbb{1}_{\mathcal{E}_{\cup}^{\mathrm{c}}} C_N(t) \overset{(122)}{\leq} \sqrt{2}\frac{\sigma_{\max}\left(\mathbf{D}_{N,a_t}(t)\right) + 1}{\sigma_{\min}\left(\mathbf{D}_{N,a_t}(t)\right) + 1} \tag{126}$$

$$\overset{(118),(121)}{\leq} \sqrt{2}\frac{\sqrt{N_{N,a_t}(t)\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} \max\left\{1, \frac{\alpha}{\sqrt{N_{N,a_t}(t)}}\right\} + 1}{\max\left\{0, \sqrt{N_{N,a_t}(t)\kappa_{\min}} - \sqrt{\tau\kappa_{\min}}\right\} + 1} \tag{127}$$

$$= g(N_{N,a_t}(t)) \, , \tag{128}$$

$$\mathbb{1}_{\mathcal{E}_{\cup}^{\mathrm{c}}} S(t) \overset{(123)}{\leq} \frac{\sqrt{2}}{\min_{i\in[N]} \sigma_{\min}\left(\mathbf{D}_{i,a_t}(t)\right) + 1} \tag{129}$$

$$\overset{(121)}{\leq} \frac{\sqrt{2}}{\min_{i\in[N]} \max\left\{0, \sqrt{N_{i,a_t}(t)\kappa_{\min}} - \sqrt{\tau\kappa_{\min}}\right\} + 1} \tag{130}$$

$$= \max_{i\in[N]} h(N_{i,a_t}(t)) \, . \tag{131}$$

Plugging inequalities in (128) and (131) into $\mathbb{1}_{\mathcal{E}_{\cup}^c} \sum_{t=1}^T C_N(t)S(t)$, we have

$$\mathbb{1}_{\mathcal{E}_{\cup}^c} \sum_{t=1}^T C_N(t)S(t) \leq \sum_{t=1}^T g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) . \tag{132}$$

Note that $h(x)$ is a non-increasing function of x: it is equal to $\sqrt{2}$ for $x \leq \tau$, and it is decreasing for $x \geq \tau$. Furthermore, for $n \geq \tau$ values, we have $\sqrt{n} \geq \frac{\alpha m^2}{\kappa_{\min}} \geq \alpha$, and $\max\left\{1, \frac{\alpha}{\sqrt{n}}\right\} = 1$. Hence,

$$g(n) = \sqrt{2} \frac{\sqrt{n\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} + 1}{\sqrt{n\kappa_{\min}} - \sqrt{\tau\kappa_{\min}} + 1}, \quad \forall n \geq \tau, \tag{133}$$

which is also a decreasing function of n for $n \geq \tau$. To use this behavior of $g(n)$, we split the (132) into two parts as follows.

$$\sum_{t=1}^T g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) = \sum_{t=1}^T \mathbb{1}_{\{N_{N,a_t}(t) < \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t))$$

$$+ \sum_{t=1}^T \mathbb{1}_{\{N_{N,a_t}(t) \geq \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) . \tag{134}$$

We will bound each of the two summands next.

Bounding $\sum_{t=1}^T \mathbb{1}_{\{N_{N,a_t}(t) < \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t))$. Note that if $n < \tau$, $g(n)$ becomes

$$g(n) \overset{(124)}{=} \sqrt{2}\left(\sqrt{n\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} \max\left\{1, \frac{\alpha}{\sqrt{n}}\right\} + 1\right), \quad n < \tau . \tag{135}$$

By noting that $\max_{i \in [N]} h(N_{i,a_t}(t)) \leq \sqrt{2}$, we obtain

$$\sum_{t=1}^T \mathbb{1}_{\{N_{N,a_t}(t) < \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) \leq \sqrt{2} \sum_{t=1}^T \mathbb{1}_{\{N_{N,a_t}(t) < \tau\}} g(N_{N,a_t}(t)) \tag{136}$$

$$\leq \sqrt{2} \sum_{n=1}^{\tau-1} g(n) . \tag{137}$$

Substituting the expression of $g(n)$ in (135) into (137), and splitting it into two sums for $n \leq \lfloor \alpha^2 \rfloor$ and $n \geq \lfloor \alpha^2 \rfloor + 1$ cases, we obtain

$$\sqrt{2} \sum_{n-1}^{\tau-1} g(n) \overset{(135)}{=} 2 \sum_{n=1}^{\tau-1}\left(\sqrt{n\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} \max\left\{1, \frac{\alpha}{\sqrt{n}}\right\} + 1\right) \tag{138}$$

$$= 2\left(\tau - 1 + \sqrt{\kappa_{\max}} \sum_{n=1}^{\tau-1} \sqrt{n} + \sqrt{\tau\kappa_{\min}}\left(\sum_{n=1}^{\lfloor \alpha^2 \rfloor} \frac{\alpha}{\sqrt{n}} + \sum_{n=\lfloor \alpha^2 \rfloor+1}^{\tau-1} 1\right)\right) . \tag{139}$$

24

We bound the sum terms in (139) as follows:

$$\sum_{n=1}^{\tau-1} \sqrt{n} \leq \int_{n=1}^{\tau} \sqrt{n}\, dn = \frac{2}{3}(\tau^{3/2} - 1) \leq \frac{2}{3}\tau^{3/2} , \tag{140}$$

$$\text{and} \qquad \sum_{n=}^{\lfloor \alpha^2 \rfloor} \frac{\alpha}{\sqrt{n}} = 2\alpha \sum_{n=1}^{\lfloor \alpha^2 \rfloor} (\sqrt{n} - \sqrt{n-1}) = 2\alpha(\sqrt{\lfloor \alpha^2 \rfloor}) \leq 2\alpha^2 . \tag{141}$$

Plugging these results back to (139), and using $\tau \geq \alpha^2$, we obtain

$$\sqrt{2}\sum_{n=1}^{\tau-1} g(n) \leq 2\left(\tau - 1 + \sqrt{\kappa_{\max}}\sum_{n=1}^{\tau-1}\sqrt{n} + \sqrt{\tau\kappa_{\min}}\Big(\underbrace{\sum_{n=1}^{\lfloor \alpha^2 \rfloor}\frac{\alpha}{\sqrt{n}}}_{\leq 2\alpha^2} + \underbrace{\sum_{n=\lfloor \alpha^2 \rfloor+1}^{\tau-1} 1}_{\leq \tau - \alpha^2} \Big) \right) \tag{142}$$

$$\overset{(140)}{\leq} 2\left(\tau + \frac{2}{3}\sqrt{\kappa_{\max}}\tau^{3/2} + \sqrt{\tau\kappa_{\min}}(\tau + \underbrace{\alpha^2}_{\leq \tau}) \right) \tag{143}$$

$$\leq 2\left(\tau + \frac{2}{3}\sqrt{\kappa_{\max}}\tau^{3/2} + 2\sqrt{\kappa_{\min}}\tau^{3/2} \right) \tag{144}$$

$$< 4\tau(\sqrt{\kappa_{\max}\tau} + \sqrt{\kappa_{\min}\tau} + 1) \tag{145}$$

$$= 2\sqrt{2}\tau g(\tau) . \tag{146}$$

Hence, we have the following bound for the first summand

$$\sum_{t=1}^{T} \mathbb{1}_{\{N_{N,a_t}(t) < \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) \leq 2\sqrt{2}\tau g(\tau) , \tag{147}$$

which is a constant term.

Bounding $\sum_{t=1}^{T} \mathbb{1}_{\{N_{N,a_t}(t) \geq \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t))$. Using the fact that $g(n)$ is a decreasing function for $n \geq \tau$, we have

$$\mathbb{1}_{\{N_{N,a_t}(t) \geq \tau\}} g(N_{N,a_t}(t)) \leq g(\tau) , \tag{148}$$

$$\text{and} \qquad \sum_{t=1}^{T} \mathbb{1}_{\{N_{N,a_t}(t) \geq \tau\}} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) \overset{(148)}{\leq} g(\tau) \sum_{t=1}^{T} \max_{i \in [N]} h(N_{i,a_t}(t)) . \tag{149}$$

The sum $\sum_{t=1}^{T} \max_{i \in [N]} h(N_{i,a_t}(t))$ is the final critical piece in the proof. $h(n)$ is a non-increasing function, and a decreasing one for $n \geq \tau$. However, the argument of h in (149) is changing due to taking a maximum over N possible values. We will prove in the following lemma that this can be compensated by having a $\sqrt{2N}$ factor on top of the optimal scaling behavior $\sqrt{T}$.

Lemma 8 *The term $\sum_{t=1}^{T} \max_{i \in [N]} h(N_{i,a_t}(t))$ is bounded by*

$$\sum_{t=1}^{T} \max_{i \in [N]} h(N_{i,a_t}(t)) < 2N\left(\sqrt{2}\tau + \sqrt{\frac{2}{\kappa_{\min}}}\left(\sqrt{\frac{2T}{N}} + \sqrt{\tau}\log\left(\frac{T}{2N}\right) \right) \right) . \tag{150}$$

25

Proof See Appendix B. ∎

We are ready to combine the pieces to reach the final result. We apply Lemma 8 and (146) in (134), and then on (132) to obtain

$$\mathbb{1}_{\mathcal{E}_{\cup}^{\mathrm{c}}} \sum_{t=1}^{T} C_N(t)S(t) \leq \sum_{t=1}^{T} g(N_{N,a_t}(t)) \max_{i \in [N]} h(N_{i,a_t}(t)) \tag{151}$$

$$< \frac{4g(\tau)}{\sqrt{\kappa_{\min}}} \sqrt{NT} + 2\sqrt{2}(N+1)\tau g(\tau) + \frac{2\sqrt{2}N\sqrt{\tau}g(\tau)}{\sqrt{\kappa_{\min}}} \log\left(\frac{T}{2N}\right) , \tag{152}$$

where $g(\tau) = \sqrt{2}(\sqrt{\tau\kappa_{\max}} + \sqrt{\tau\kappa_{\min}} + 1)$ since τ makes the denominator in (133) equal to 1. Note that the upper bound we have just found is not a random variable but a constant. Then, $\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cup}^{\mathrm{c}}} \sum_{t=1}^{T} C_N(t)S(t)\right]$ is immediately upper bounded by this result. Also recall the result in (117), based on which we have

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cup}} \sum_{t=1}^{T} C_N(t)S(t)\right] \overset{(117)}{\leq} \frac{m}{T} + \frac{2m}{3} + 1 . \tag{153}$$

Therefore, the final result is

$$\lambda_T = \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cup}^{\mathrm{c}}} \sum_{t=1}^{T} C_N(t)S(t)\right] + \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_{\cup}} \sum_{t=1}^{T} C_N(t)S(t)\right] \tag{154}$$

$$< \frac{4g(\tau)}{\sqrt{\kappa_{\min}}} \sqrt{NT} + 2\sqrt{2}(N+1)\tau g(\tau) + \frac{2\sqrt{2}N\sqrt{\tau}g(\tau)}{\sqrt{\kappa_{\min}}} \log\left(\frac{T}{2N}\right) + \frac{m}{T} + \frac{2m}{3} + 1 . \tag{155}$$

Finally, note that $\alpha = \mathcal{O}(\sqrt{\log(dNT)})$ and $\tau = \mathcal{O}(\log(dNT))$. Also, $\kappa_{\max}$ and $\kappa_{\min}$ are independent of T. Hence, ignoring the logarithmic terms and constants,

$$\lambda_T = \tilde{\mathcal{O}}(\sqrt{NT}) . \tag{156}$$

∎

Finally, we combine Theorem 5 and 6 to obtain the regret bound of our algorithm, formalized next.

Theorem 9 *Under Assumption 1, the regret of LinSEM-UCB is*

$$\mathbb{E}[R(T)] = \tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT}) . \tag{157}$$

Proof Since $\beta_T = m_B + \sqrt{2\log(2NT) + (d+1)\log(1 + m^2T/(d+1))}$ and m_B is constant, ignoring the poly-logarithmic factors, $(\beta_T + m_B)^{L+1}$ contributes $(d+1)^{\frac{L+1}{2}}$ factor to the result of Theorem 5. Factoring the result of Theorem 6 for λ_T, we obtain $\mathbb{E}[R(T)] = \tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT})$. ∎

Algorithm 2 LinSEM-TS

1: **Input:** Causal graph $\mathcal{G}$, action set $\mathcal{A}$, hyperparameter σ, prior distribution π_0.
2: **Initialization:** Initialize parameters for $2N$ linear problems:
3: $[\mathbf{B}(0)]_i = \mathbf{0}_{(N+1)\times 1}, [\mathbf{B}^*(0)]_i = \mathbf{0}_{(N+1)\times 1}, \ \forall i \in [N]$ $\triangleright$ initiliaze estimates for parameter vectors
4: $\mathbf{V}_i(0) = I_{N+1}, \ \mathbf{V}_i^*(0) = I_{N+1}, \ g_i(0) = \mathbf{0}_{(N+1)\times 1}, \ g_i^*(0) = \mathbf{0}_{(N+1)\times 1}$ $\triangleright$ initiliaze auxiliary parameters
5: **for** $t = 1, 2, \ldots T$ **do**
6:     $\breve{\mathbf{W}} \sim \pi_{t-1}(\mathbf{W} \mid X(1), \ldots, X(t-1))$
7:     **for** $a \in \mathcal{A}$ **do**
8:        Construct $\breve{\mathbf{B}}_a$ from rows of $\breve{\mathbf{W}}$ similarly to (9).
9:        $\hat{\mu}_a \leftarrow f(\breve{\mathbf{B}}_a)$ $\triangleright$ expected reward under action a
10:     **end for**
11:     $a_t = \arg\max_a \hat{\mu}_a$ $\triangleright$ select the action that maximizes expected reward
12:     Pull a_t, observe $X(t) = (1, X_1(t), \ldots, X_N(t))^\top$.
13:     Update the posterior to π_t under the linear model, similarly to Lines 11-21 of Algorithm 1.
14: **end for**

---

5. LinSEM-TS Algorithm

Thompson sampling (TS)-based algorithms gradually refine the posterior distributions of the reward for each action and select actions by sampling from their posterior distributions. The actions are selected sequentially in a way that they balance the exploitation and the exploration processes. To this end, a TS algorithm needs to update the posterior distributions of all arms. Similar to LinSEM-UCB's improvement over UCB, the known causal structure can be leveraged to improve the performance of TS in causal bandits.

Assumption 2 (Bounded prior) *The domain of the prior distribution of parameters is bounded as*
$\mathcal{W} \triangleq \{\mathbf{W} \in R^{2N \times N} : \|[\mathbf{W}]_i\| \leq m_B \ \forall i \in [2N]\}.$

We denote the prior distribution of $\mathbf{W}$ by π_0, and denote its posterior given the filtration generated until time t by π_t for $t \in \mathbb{N}$. The LinSEM-TS algorithm at time t samples from the posterior distribution π_{t-1}. Subsequently, it constructs the weight matrix for each intervention action and computes the corresponding expected reward. Next, the intervention with the highest expected reward is selected, and the graph instance $X(t)$ is observed. Finally, the posterior distributions of the $2N$ independent weight vectors that constitute $\mathbf{W}$ are updated according to the linear model.

We note that LinSEM-TS has two advantages over LinSEM-UCB. First, the constrained optimization problem for computing $\mathrm{UCB}_a(t)$ has no closed-form solution to our knowledge. Hence, LinSEM-UCB is computationally expensive to implement in practical settings. LinSEM-TS does not suffer from this problem. Secondly, LinSEM-UCB requires to know bound m on $\|X\|$ and horizon T since the radius of the confidence intervals β_T depends on m and T. In contrast, LinSEM-TS does not require to know m or T. Hence, LinSEM-TS is an anytime algorithm. Furthermore, we can ensure a performance guarantee similar to that of Theorem 9 for the Bayesian setting. This guarantee is formalized in the next theorem.

Algorithm 3 LinSEM-TS-Gaussian

 1: **Input:** Causal graph $\mathcal{G}$, action set $\mathcal{A}$, hyperparameter σ.
 2: **Initialization:** Initialize parameters for $2N$ linear problems:
 3: $[\mathbf{B}(0)]_i = \mathbf{0}_{N \times 1}, [\mathbf{B}^*(0)]_i = \mathbf{0}_{N \times 1}, \forall i \in [N]$ ▷ initiliaze estimates for parameter vectors
 4: $\mathbf{V}_i(0) = I_{N+1}, \mathbf{V}_i^*(0) = I_{N+1}, g_i(0) = \mathbf{0}_{N \times 1}, g_i^*(0) = \mathbf{0}_{N \times 1}$ ▷ initiliaze auxiliary
      parameters
 5: **for** $t = 1, 2, \ldots T$ **do**
 6:     **for** $i \in \{1, \ldots, N\}$ **do**
 7:         $[\breve{\mathbf{B}}(t)]_i \sim \mathcal{N}([\mathbf{B}(t-1)]_i, \sigma^2 [\mathbf{V}_i(t-1)]^{-1})$
 8:         $[\breve{\mathbf{B}}^*(t)]_i \sim \mathcal{N}([\mathbf{B}^*(t-1)]_i, \sigma^2 [\mathbf{V}_i^*(t-1)]^{-1})$
 9:     **end for**
10:     **for** $a \in \mathcal{A}$ **do**
11:         Construct $\breve{\mathbf{B}}_a$ from rows of $\breve{\mathbf{W}} \triangleq [\breve{\mathbf{B}} \ \breve{\mathbf{B}}^*]$ similarly to (9).
12:         $\hat{\mu}_a \leftarrow f(\breve{\mathbf{B}}_a)$
13:     **end for**
14:     $a_t = \arg\max_a \hat{\mu}_a$ ▷ select the action that maximizes the expected reward
15:     Pull a_t, observe $X(t) = (X_1(t), \ldots, X_N(t))^\top$.
16:     Update the Gaussian posteriors under the linear model.
17: **end for**

Theorem 10 *Under Assumption 1 and Assumption 2, the Bayesian regret of LinSEM-TS is*

$$\mathrm{BR}(T) = \tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT}) . \tag{158}$$

Proof The main tools of the proofs, e.g., Lemma 2, are similar to that proof of frequentist result. However, the proof for Thompson sampling critically relies on the property that conditional distributions of different actions are almost surely the same given proper filtration. Importantly, this filtration is different than the one used for concentration inequalities in the proof of Lemma 7. Finally, the ranges of the parameters regarding the intervention space, e.g., $\kappa_{\min}$, are different from that of the frequentist analysis and are examined carefully. For detailed proof, see Appendix B. ∎

5.1 LinSEM-TS-Gaussian Algorithm

We have provided LinSEM-TS and its performance guarantees under the bounded noise and bounded prior distributions. Sampling from the posteriors under bounded priors can be done via different numerical techniques yet they can be time-consuming. For instance, commonly used Gibbs sampling has a long convergence time, especially in the high-dimensional regime. To circumvent this, in this section, we present a modified algorithm that leverages Gaussian priors, which we refer to as LinSEM-TS-Gaussian. We note that the theoretical guarantees presented in Section 5 do not apply to LinSEM-TS-Gaussian since boundedness assumptions are not satisfied.

Gaussian posterior computation. We use a Gaussian distribution for the prior distribution of the parameters and a Gaussian likelihood model for the data. We follow the standard Bayesian posterior inference (Agrawal and Goyal, 2013). To summarize, consider the posterior computation

of parameters $[\mathbf{B}]_i$. The likelihood function of $X_i(t)$ is given by $\mathcal{N}(\langle[\mathbf{B}]_i, X_{\overline{\mathrm{pa}}(i)}(t)\rangle, \sigma_i^2)$ according to the linear SEM. We start with the prior $\mathcal{N}(\mathbf{0}_{N\times 1}, I_{N+1})$ for $[\mathbf{B}]_i$. If the prior of $[\mathbf{B}]_i$ at round t is $\mathcal{N}([\mathbf{B}(t-1)]_i, \sigma^2[\mathbf{V}_i(t-1)]^{-1})$, then the posterior distribution at round t is computed as $\mathcal{N}([\mathbf{B}(t)]_i, \sigma^2[\mathbf{V}_i(t)]^{-1})$. A similar computation shows that the posterior distribution of $[\mathbf{B}^*]_i$ at round t is $\mathcal{N}([\mathbf{B}^*(t)], \sigma^2[\mathbf{V}_i^*(t)]^{-1})$.

Algorithm details. Algorithm 3 presents the LinSEM-TS-Gaussian algorithm. Since we have exact forms of the posteriors of weight vectors, we explicitly write the sampling of parameters in lines 7-8. Estimated weight matrices and intervention selection are all formed similarly to those of Algorithm 2. Finally, interventional and observational posteriors are updated according to the linear model specified earlier.

6. Lower Bounds

In this section, we provide minimax lower bounds on the regret. The main observation is that the dependence of these lower bounds on the graph parameters (maximum degree d and longest causal path length L) conforms to that of the achievable regret bounds. For this purpose, we consider a general class of graphs and prove a minimax lower bound by showing that there exists a causal bandit environment parameterized by a linear SEM, for which the regret of any algorithm is at least a constant factor of $d^{\frac{L}{2}-2}\sqrt{T}$.

A general class of graphs. Note that given values of d and L necessitate that the graph size N is at least $\max\{d+1, L+1\}$, and at most $\sum_{\ell=0}^{L} d^\ell < 2d^L$. On the other hand, d and L can vary independently without imposing restrictions on one another. Therefore, we aim to prove a lower bound for fixed d and L whereas N can vary freely. In the proof of the following theorem, we construct graph structures and bandit instances that are flexible enough to accommodate all possible choices of d and L, and the graph size N that can have varying sizes in the range $N \in \{(dL - d + 1), \ldots, d^L + (dL - d + 1)\}$. Since $N < 2d^L$, this class contains large enough graphs that have the same order as the largest possible graphs for fixed d and L.

Theorem 11 *Given the pair (d, L), for any $N \in \{d(L-1) + 2, \ldots, d^L + d(L-1) + 1\}$, there exists a causal bandit instance with graph parameters (d, L, N) such that expected regret of any algorithm is at least*

$$\mathbb{E}[R(T)] \geq cd^{\frac{L}{2}-2}\sqrt{T} , \tag{159}$$

where $c > 0$ is a constant.

Sketch of the proof. The central idea of the proof is as follows. Two linear SEM causal bandit instances that differ by only one edge parameter are hard to distinguish. At the same time, we can construct them to have different optimal actions, indicating that a selection policy cannot incur small regret for both at the same time under the same data realization. Note that, the difference of the rewards, or equivalently the regrets, observed by these two bandit instances under the same action can be computed by tracing the effect of that differing edge parameter over all the paths that end at the reward node. We use a hierarchical structure that consists of L layers and d nodes at each layer to maximize the number of such paths. We also use d-ary trees to control the number of nodes in the graph without affecting the lower bound analysis. See Appendix C for detailed proof.

Remark 12 *We remark that optimism in the face of uncertainty linear bandit algorithm (OFUL) of (Abbasi-Yadkori et al., 2011) has $\mathcal{O}(d\sqrt{T})$ regret, whereas the lower bound for linear bandit problem with finite arms is $\mathcal{O}(\sqrt{dT})$. Note that the linear bandit problem can be considered a special case of our framework with $L = 1$. Since our confidence intervals rely on the same principles as OFUL, the gap between lower and upper bounds established in the paper, i.e., $d^{\frac{L}{2}}$ factor, is due to the aggregated effect of L layers. Tighter regret results would have been possible by using a more sophisticated approach for building the confidence intervals along with elimination, such as the SupLinUCB algorithm of (Chu et al., 2011). However, this investigation is out of the scope of this paper.*

7. Simulations

The main novelty of our work is to use causal graph knowledge without knowing the interventional distributions. Furthermore, we use soft interventions on continuous variables. To the best of our knowledge, there does not exist any causal bandit algorithm that suits this setting. The standard linear bandit algorithms are also not applicable to causal bandits for linear SEMs due to the feature vectors being unknown. Therefore, we compare the performance of our method to that of the non-causal UCB algorithm. For practical implementation, we use Algorithm 3 for the simulations in this section. [1]

Parameterization. We consider linear SEMs with soft interventions under the Bayesian setting. For a given causal graph structure, the prior distributions of the non-zero entries of observational parameters $\mathbf{B}$ are sampled independently at random according to the uniform distribution on $[-1, -0.25] \cup [0.25, 1]$. Priors for interventional weights are set as negatives of observational weights, $\mathbf{B}^* = -\mathbf{B}$, so that the effect of an intervention is strong. To approximate the Bayesian setting, the true weights $[\mathbf{B}]_i = -[\mathbf{B}^*]_i$ are sampled 10 times from a Gaussian prior with small variances. The distribution of the individual noise terms ϵ_i is set to $\mathcal{N}(0, 1)$ for all nodes. We note that all parameters are unknown to the learner. Simulations for each sampled parameter set are repeated 10 times. We simulate causal graphs under two structural families.

Hierarchical Graphs: We consider hierarchical graph structures with L layers and degree d constructed as follows. Each layer $\ell \in \{1, \dots, L\}$ contains d nodes, each node at layer $\ell < L$ is a parent of the nodes in the next layer $(\ell + 1)$, and the nodes in the final layer L are the parents of the reward node. The total number of nodes is $N = dL + 1$. Figure 2a depicts the hierarchical structure for $d = 3$ and $L = 2$. Every node is set to be intervenable, leading to $|\mathcal{A}| = 2^{dL+1}$ number of possible interventions. We construct this structure by observing that both our regret upper bound and the minimax lower bound are a function of d and L.

Figure 2b compares the cumulative regret of LinSEM-TS-Gaussian and non-causal UCB algorithms for $d = 3$ and $L = 2$. We observe that LinSEM-TS-Gaussian algorithm significantly outperforms the standard bandit algorithm by exploiting the known causal structure. Next, we look into the graph structure's effect on the regret of LinSEM-TS-Gaussian. In Figure 2c, we plot the cumulative regrets at $T = 5000$ for different values of d and L. We note that the curves in Figure 2c become steeper as L increases. For instance, the green curve ($L = 4$) has a faster-growing regret than the blue curve ($L = 2$). On the other hand, the growth of regret is much slower than the growth

---

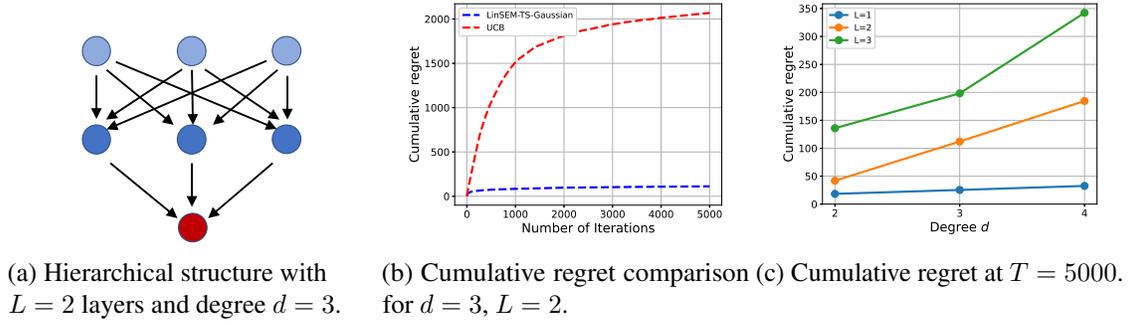1. The codebase for reproducing the simulations are available at `https://github.com/bvarici/causal-bandits-linear-sem`.

(a) Hierarchical structure with $L = 2$ layers and degree $d = 3$.

(b) Cumulative regret comparison for $d = 3$, $L = 2$.

(c) Cumulative regret at $T = 5000$.

Figure 2: Simulation results for hierarchical causal graphs.

(a) Sample structure with $N = 5$
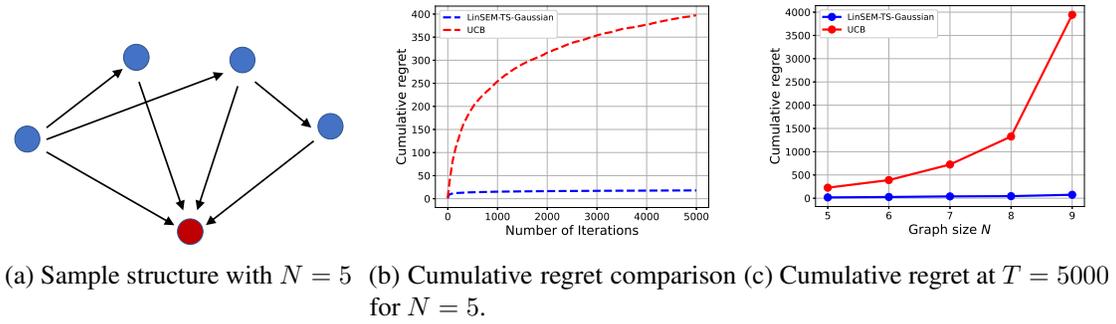
(b) Cumulative regret comparison for $N = 5$.

(c) Cumulative regret at $T = 5000$

Figure 3: Simulation results for enhanced parallel bandits.

of the cardinality $|\mathcal{A}| = 2^{dL+1}$. These observations imply that our theoretical upper bounds that scale with d^L instead of $|\mathcal{A}|$ are realized empirically.

Enhanced parallel bandits: We construct a structure similar to the parallel bandits model of Lattimore et al. (2016) with additional edges. In this model, each node $\{1, \ldots, N - 1\}$ is a parent of the reward node N. In addition, we assign one parent to each node $i \in \{2, \ldots, N - 1\}$, which is chosen uniformly at random from the nodes $\{1, \ldots, i - 1\}$. Every node is intervenable, which gives $|\mathcal{A}| = 2^N$ number of possible interventions. One example of this structure with 5 nodes is given in Figure 3a. We randomly generate 5 structures for each value of $N \in \{5, 6, 7, 8, 9\}$ and run LinSEM-TS-Gaussian and UCB for each graph.

Figure 3b compares the cumulative regret of LinSEM-TS-Gaussian and UCB algorithms for $N = 5$. Similarly to the previous setting, LinSEM-TS-Gaussian outperforms UCB that does not use the causal relationships among the arms. Secondly, we fix horizon T at 5000 and compare the performance of LinSEM-TS-Gaussian as N grows. We note that the maximum degree d is equal to $(N - 1)$ for enhanced parallel bandits. The largest path length L is not strictly controlled by N and can take values between 2 and $(N - 1)$. Figure 3c shows that UCB algorithm's regret grows exponentially as N grows, as $|\mathcal{A}|$ does, whereas LinSEM-TS-Gaussian incurs a much smaller regret.

Graph misspecification. In the simulations above, the learner is given the perfect knowledge of the causal structure following the theoretical analysis in previous sections. In the following simulations, we investigate a possible scenario in which the structure given to the learner deviates from the true structure by a few edges. This can be a practical scenario where the domain expert generously assigns
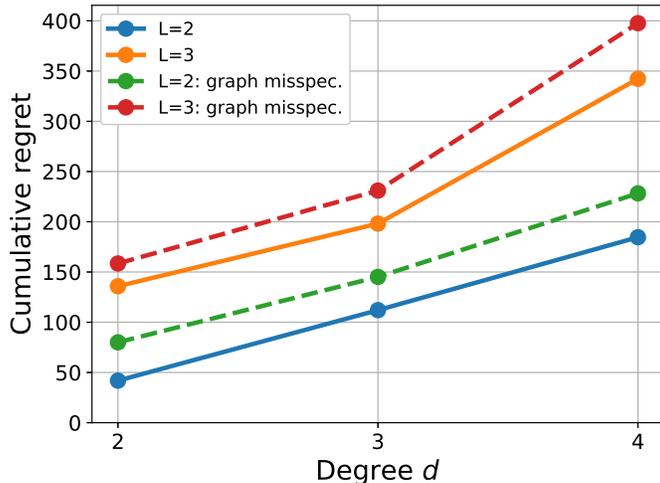
Figure 4: Cumulative regret comparison (at $T = 5000$) of the main setting with perfect knowledge of the graph and graph misspecification setting for varying d and L values. Our algorithm incurs sublinear regret even when there are misspecified edges.

edges from possible causes to effects. In other words, the learner can be given a graph that is a supergraph of the true graph with some additional edges. We simulate this graph misspecification setting in hierarchical graphs by randomly adding incorrect edges to the graph structure provided to the learner. The incorrect edges are chosen in a way that preserves acyclicity. The number of incorrect edges is set to 2 for each graph instance, which is reasonable given that the graph structures have moderate sizes.

Figure 4 shows that our method still works when the knowledge of graph structure is perturbed by a small number of edges. We interpret this result by recalling our theoretical analysis. Since the true parents of a node i is a subset of the assumed parents of it, the guarantees of estimating $[\mathbf{B}]_i$ and $[\mathbf{B}^*]_i$ would still follow through, albeit for a greater degree than the true degree of node i. The overall effect is a potential increase in the graph's d and L values. The results in Figure 4 align with our expectations as the regret under graph misspecification is only slightly worse than the case with perfect graph knowledge.

Knowledge of interventional distributions Finally, we assess the effect of knowing or lacking the interventional distributions through simulations. The interventional distributions of the reward's parents $\{\mathbb{P}_a(\overline{\mathrm{pa}}(N)) : a \in \mathcal{A}\}$ in linear SEMs correspond to knowing all of the weights $[\mathbf{B}_a]_i$ for non-reward nodes $i \in [N-1]$. In this case, the estimation problem reduces to estimating only the two weight vectors $[\mathbf{B}]_N$ and $[\mathbf{B}^*]_N$ corresponding to the reward node. Hence, the causal bandit problem becomes significantly easier when the interventional distributions are given a priori.

Figure 5 shows that the effect of L with a fixed value of d on the hierarchical graph's cumulative regret is more significant when the distributions $\{\mathbb{P}_a(\overline{\mathrm{pa}}(N)) : a \in \mathcal{A}\}$ are unknown. This observation can be explained by the fact that there is no compounded estimation error through L-length paths when the knowledge of the distributions is given, and the problem essentially reduces to a linear bandit. These empirical results illustrate the relative difficulty of our problem setting, i.e., lacking the knowledge of interventional distributions.
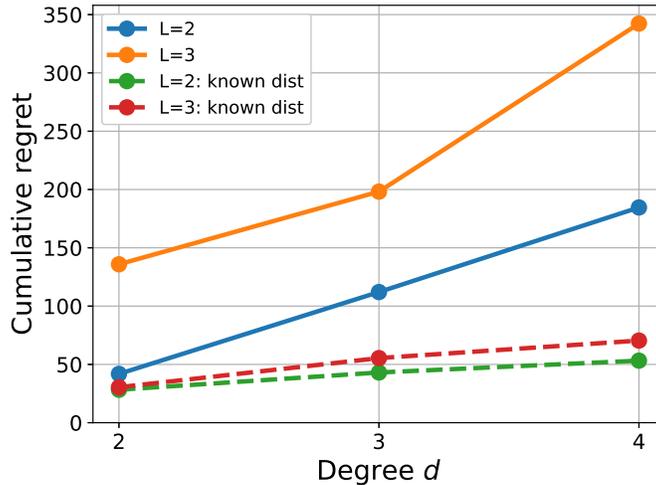
Figure 5: Cumulative regret at $T = 5000$ with knowledge and without the knowledge of the interventional distributions for varying d and L values.

8. Discussion

Background. There have been considerable recent advances in understanding the fundamental performance limits of causal bandit problems under various settings. The hallmark of the results is establishing theoretical advantages over non-causal methods. The extent of such advantages critically hinges on the extent of information known about the topology of the causal graph and the interventional distributions. Recent studies have advanced in the direction of having minimal assumptions on these two key factors. Some of these studies, nevertheless, rely on other tacit or simplifying assumptions such as adopting single-node interventions and Boolean variables. Motivated by the applications in which domain knowledge suffices to construct a causal graph (e.g., the computational advertising model given in Figure 1), an important direction of studies, which is also the focus of this paper, is assuming that the topology is known while the intervention distributions are unknown.

Contributions. Our focus has been on designing an optimal sequence of interventions when the intervention space can grow with the graph. Specifically, mainly for analytical clarity, we have assumed one observational and one interventional model per node. That, however, can be readily extended to multiple interventional models per node, albeit at the expense of a higher degree of parameterization. We have addressed the open theoretical question of finding an optimal sequence of interventions for optimizing a graph reward value. This is done for the class of linear systems in which the causal relationships follow linear models. This renders that the nodes' conditional distributions have locally linear properties. For the linear models, we have focused on previously understudied soft interventions, continuous variables, and exponentially large intervention space. Under these settings, we have characterized upper and lower bounds on the hitherto unknown cumulative regret under appropriate boundedness assumptions. We have shown that the achievable regret bounds depend on the properties of the causal graph but do not grow exponentially with the graph size, whereas the

cardinality of the intervention space does. Furthermore, we have established a lower bound to show that this dependence on the graph properties is information-theoretic.

Future directions. There are several open theoretical questions in the setting we have investigated. One pertains to relaxing the linear structural equation model. This can be done, for instance, by using reproducing kernel Hilbert space with proper approximation methods. Investigating the graphs with partially known structures, e.g., topological orders, is another important direction. Investigating the sensitivity of the results to model mismatch, contamination, or fluctuation is necessary to understand these algorithms' robustness. Finally, there is a scope for improving the dependence of our regret guarantees on the properties of intervention space. Given that regret scales with $\sqrt{T}$, a more sophisticated analysis would indicate that actions with autocorrelation matrix ($\Sigma_{i,a}$) that are closer to that of the optimal arm (Σ_{i,a^*}) are played overwhelmingly often. Then, $\frac{1}{\sqrt{\kappa_{\min}}}$ factor, in which $\kappa_{\min}$ is a lower bound over the intervention space, might be too conservative.

Acknowledgments

Appendix A. Notations

The following table summarizes the commonly used notations throughout the paper.

$$\mathcal{A}, \mathcal{E} : \text{calligraphic letters for sets and events}$$
$$X_i : \text{random variables of the causal graph}$$
$$X_i(t) : \text{value of node } i \text{ at time } t$$
$$X : (X_1, \ldots, X_N)^\top$$
$$X(t) : (1, X_1(t), \ldots, X_N(t))^\top \quad \text{data vector at time } t$$
$$X_{\mathcal{S}}(t) : X(t) \odot \mathbf{1}(\mathcal{S}) \quad \forall \mathcal{S} \subseteq \mathcal{V}$$
$$X_{\overline{\mathrm{pa}}(i)}(t) : \text{follows the previous line}$$
$$\epsilon_i : \text{noise variable for node } i$$
$$\epsilon : (1, \epsilon_1, \ldots, \epsilon_N)^\top$$
$$\epsilon(t) : (1, \epsilon_1(t), \ldots, \epsilon_N(t))^\top \quad \text{noise vector at time } t$$
$$\mathbf{A}, [\mathbf{A}]_i, [\mathbf{A}]_{i,j} : \text{a matrix } \mathbf{A}, \text{ its column } i, \text{ and its entry at row } i \text{ and column } j, \text{ respectively}$$
$$\mathbf{B}_a, [\mathbf{B}_a]_i : \text{weight matrix for intervention } a, \text{ and its } i\text{-th column}$$
$$\mathbf{B}_a(t), [\mathbf{B}_a(t)]_i : \text{estimate of } \mathbf{B}_a \text{ at time } t, \text{ and its } i\text{-th column}$$
$$\mathbf{B}_a^\ell, [\mathbf{B}_a^\ell]_i : \ell\text{-th power of } \mathbf{B}_a, \text{ and its } i\text{-th column}$$
$$\mathbf{B}_a^\ell(t), [\mathbf{B}_a^\ell(t)]_i : \ell\text{-th power of } \mathbf{B}_a(t), \text{ and its } i\text{-th column.}$$
$$\mathbf{\Delta}_a(t) : \mathbf{B}_a(t) - \mathbf{B}_a$$
$$[\mathbf{\Delta}_a(t)]_i : [\mathbf{B}_a(t)]_i - [\mathbf{B}_a]_i$$
$$\mathbf{\Delta}_a^{(\ell)}(t) : \mathbf{B}_a^\ell(t) - \mathbf{B}_a^\ell$$
$$[\mathbf{\Delta}_a^{(\ell)}(t)]_i : [\mathbf{B}_a^\ell(t)]_i - [\mathbf{B}_a^\ell]_i$$
$$\mathbf{V}_i(t), \mathbf{V}_i^*(t) : \text{regularized gram matrices for node } i, \text{ observational and interventional}$$
$$\mathbf{D}_i(t), \mathbf{D}_i^*(t) : \text{data matrices for parents of node } i$$
$$\mathbf{V}_{i,a}(t), \mathbf{D}_{i,a}(t) : \text{relevant gram and data matrices for node } i \text{ under intervention } a$$
$$N_i^*(t) : \text{number of times node } i \text{ is intervened up to time } t$$
$$N_i(t) : \text{number of times node } i \text{ is non-intervened up to time } t$$
$$\Sigma_{i,a} : \text{autocorrelation matrix for parents of node } i \text{ under intervention } a$$

Appendix B. Proofs of Our Regret Results

Proof of Lemma 2. Note that β_T is independent of $t \in [T]$. For simplicity, we use β as a shorthand for β_T. We will prove the desired result in three steps. Note that matrix $\mathbf{V}_{i,a}(t)$ has $(d+1)$-sparse vectors in its rows and columns, based on its definition in (37). This is a key property that will be used in all the following steps.

Step 1: First, we derive the bounds for vectors $[\mathbf{B}_a]_i$, $[\Delta_a(t)]_i$, and their norms. For any valid vector $[\mathbf{B}_a]_i$ and matrix $\mathbf{V}_{N,a}(t)$, we use the CS inequality to obtain

$$\|[\mathbf{B}_a]_i\|_{\mathbf{V}_{N,a}(t)} \leq \underbrace{\|[\mathbf{B}_a]_i\|}_{\leq m_B}\lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) \leq m_B\lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) . \tag{160}$$

Also, noting that $[\Delta_a(t)]_i = [\Delta_a(t)]_i \odot \mathbf{1}(\overline{\mathrm{pa}}(i))$, we have

$$\|[\Delta_a(t)]_i\|_{\mathbf{V}_{i,a}(t)} \geq \|[\Delta_a(t)]_i \odot \mathbf{1}(\overline{\mathrm{pa}}(i))\|\,\lambda_{\min}^{1/2}(\mathbf{V}_{i,a}(t)) = \|[\Delta_a(t)]_i\|\,\lambda_{\min}^{1/2}(\mathbf{V}_{i,a}(t)) . \tag{161}$$

Using the conditions of the lemma and the inequalities above, we also have

$$\|[\Delta_a(t)]_i\| \overset{(161)}{\leq} \underbrace{\|[\Delta_a(t)]_i\|_{\mathbf{V}_{i,a}(t)}}_{\leq\beta}\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) \leq \beta\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) . \tag{162}$$

Note that this lemma provides the result for the $\mathbf{V}_{N,a}(t)$-norm. Using (160), (161), and the lemma conditions, we have

$$\|[\Delta_a(t)]_i\|_{\mathbf{V}_{N,a}(t)} \overset{(160)}{\leq} \|[\Delta_a(t)]_i\|\,\lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) \tag{163}$$

$$\overset{(161)}{\leq} \underbrace{\|[\Delta_a(t)]_i\|_{\mathbf{V}_{i,a}(t)}}_{\leq\beta}\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t))\lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) \tag{164}$$

$$\leq \beta\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t))\lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) . \tag{165}$$

Next, we use the CS inequality, $|\overline{\mathrm{pa}}(i)| \leq d+1$, and $\|[\mathbf{B}_a]_i\| \leq m_B$ to obtain

$$\sum_{j\in\overline{\mathrm{pa}}(i)} [\mathbf{B}_a]_{j,i} \overset{(CS)}{\leq} \sqrt{|\overline{\mathrm{pa}}(i)| \sum_{j\in\overline{\mathrm{pa}}(i)} ([\mathbf{B}_a]_{j,i})^2} \leq \|[\mathbf{B}_a]_i\|\sqrt{d+1} \leq m_B\sqrt{d+1} . \tag{166}$$

For the error vector $\Delta_a(t)$, we use (161) and (162) to obtain

$$\sum_{j\in\overline{\mathrm{pa}}(i)} [\Delta_a(t)]_{j,i} \overset{(CS)}{\leq} \sqrt{(d+1) \sum_{j\in\overline{\mathrm{pa}}(i)} ([\Delta_a(t)]_{j,i})^2} = \sqrt{d+1}\,\|[\Delta_a(t)]_i\| \tag{167}$$

$$\overset{(161)}{\leq} \sqrt{d+1}\,\|[\Delta_a(t)]_i\|_{\mathbf{V}_{i,a}(t)}\,\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) \tag{168}$$

$$\overset{(162)}{\leq} \sqrt{d+1}\beta\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) . \tag{169}$$

Step 2: In (165), we have the term $\lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t))$. Let us define

$$S_a(t) \triangleq \max_{i\in[N]} \lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) = \frac{1}{\min_{i\in[N]} \lambda_{\min}^{1/2}(\mathbf{V}_{i,a}(t))} . \tag{170}$$

By the definitions in (41) and (42), $\mathbf{B}_a(t) = \Delta_a(t) + \mathbf{B}_a$, and $\Delta_a^{(\ell)}(t) = [\Delta_a(t) + \mathbf{B}_a]^\ell - \mathbf{B}_a^\ell$. Therefore, each term in the binomial expansion of $\Delta_a^{(\ell)}(t)$ is a product that consists of $\Delta_a(t)$ and $\mathbf{B}_a$

factors. For any $\ell \in [L+1]$ and $k \in [\ell] \cup \{0\}$, there are $\binom{\ell}{k}$ terms that contains the $\Delta_a(t)$ factor k times and the $\mathbf{B}_a$ factor $(\ell - k)$ times. We denote the set of such product terms by

$$\mathcal{H}_{\ell,k} \triangleq \{H : H \text{ has } \Delta_a(t) \text{ factor } k \text{ times and } \mathbf{B}_a \text{ factor } \ell - k \text{ times}\} . \tag{171}$$

For instance, set $\mathcal{H}_{3,1}$ consists of $\binom{3}{1} = 3$ terms:

$$\mathcal{H}_{3,1} = \{\Delta_a(t)\mathbf{B}_a\mathbf{B}_a, \mathbf{B}_a\Delta_a(t)\mathbf{B}_a, \mathbf{B}_a\mathbf{B}_a\Delta_a(t)\} . \tag{172}$$

Note that $\mathcal{H}_{\ell,0} = \{\mathbf{B}_a^\ell\}$, which cancels out the $\mathbf{B}_a^\ell$ term in the expansion of $\Delta_a^{(\ell)}(t)$. Therefore, sets $\mathcal{H}_{\ell,1}, \ldots, \mathcal{H}_{\ell,\ell}$ contain all valid products consisting of $\mathbf{B}_a$ and $\Delta_a(t)$. Hence,

$$\Delta_a^{(\ell)}(t) = \sum_{k=1}^{\ell} \sum_{H \in \mathcal{H}_{\ell,k}} H . \tag{173}$$

In this step, by induction, we show that for any $\ell \geq 1$, and $k \in [\ell] \cup \{0\}$,

$$\|[H]_i\|_{\mathbf{V}_{N,a}(t)} \leq (d+1)^{\frac{\ell-1}{2}} m_B^{\ell-k} \beta^k S_a^k(t) \lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) , \quad \forall H \in \mathcal{H}_{\ell,k} . \tag{174}$$

Consider $\ell = 1$. For $k = 0$, we have $\mathcal{H}_{1,0} = \{\mathbf{B}_a\}$, and from (160) we have $\|[\mathbf{B}_a]_i\|_{\mathbf{V}_{N,a}(t)} \leq m_B \lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t))$. For $k = 1$, $\mathcal{H}_{1,1} = \{\Delta_a(t)\}$, and

$$\|[\Delta_a(t)]_i\|_{\mathbf{V}_{N,a}(t)} \overset{(165)}{\leq} \beta \lambda_{\min}^{-1/2}(\mathbf{V}_{i,a}(t)) \lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) \overset{(170)}{\leq} \beta S_a(t) \lambda_{\max}^{1/2}(\mathbf{V}_{N,a}(t)) . \tag{175}$$

Therefore, (174) is correct for $\ell = 1$. Suppose that it is correct for $1, \ldots, \ell - 1$ values, for $\ell \geq 2$. Consider a product term $H \in \mathcal{H}_{\ell,k}$, for some $k \in [\ell] \cup \{0\}$. The first factor of H is either $\mathbf{B}_a$ or $\Delta_a(t)$, and we analyze the induction step for each of these possibilities separately.

1. If H starts with $\mathbf{B}_a$, represent it by $H = \mathbf{B}_a \bar{H}$, where $\bar{H} \in \mathcal{H}_{\ell-1,k}$ and $k \in [\ell - 1] \cup \{0\}$. Using the induction assumption for the elements of set $\mathcal{H}_{\ell-1,k}$ we obtain

$$\|[H]_i\|_{\mathbf{V}_{N,a}(t)}^2 = \|(\mathbf{B}_a \bar{H})_i\|_{\mathbf{V}_{N,a}(t)}^2 \tag{176}$$

$$= \sum_{u,v \in \overline{\mathrm{pa}}(i)} [\mathbf{B}_a]_{u,i} [\mathbf{B}_a]_{v,i} \bar{H}_v^\top \mathbf{V}_{N,a}(t) \bar{H}_u \tag{177}$$

$$\overset{(\mathrm{CS})}{\leq} (d+1) \sum_{u \in \overline{\mathrm{pa}}(i)} ([\mathbf{B}_a]_{u,i})^2 \|\bar{H}_u\|_{\mathbf{V}_{N,a}(t)}^2 \tag{178}$$

$$\overset{(174)}{\leq} (d+1)^{\ell-1} m_B^{2(\ell-k-1)} \beta^{2k} S_a^{2k}(t) \lambda_{\max}(\mathbf{V}_{N,a}(t)) \sum_{u \in \overline{\mathrm{pa}}(i)} ([\mathbf{B}_a]_{u,i})^2 \tag{179}$$

$$= (d+1)^{\ell-1} m_B^{2(\ell-k-1)} \beta^{2k} S_a^{2k}(t) \lambda_{\max}(\mathbf{V}_{N,a}(t)) \underbrace{\|[\mathbf{B}_a]_i\|^2}_{\leq m_B^2} \tag{180}$$

$$\leq (d+1)^{\ell-1} m_B^{2(\ell-k)} \beta^{2k} S_a^{2k}(t) \lambda_{\max}(\mathbf{V}_{N,a}(t)) . \tag{181}$$

2. If H starts with $\Delta_a(t)$ represent it by $H = \Delta_a(t)\bar{H}$, where $\bar{H} \in \mathcal{H}_{\ell-1,k-1}$ and $k \in [\ell]$. Similarly to the first case, we have

$$\|[H]_i\|^2_{\mathbf{V}_{N,a}(t)} = \left\|[\Delta_a(t)\bar{H}]_i\right\|^2_{\mathbf{V}_{N,a}(t)} \tag{182}$$

$$= \sum_{u,v\in\overline{\mathrm{pa}}(i)} [\Delta_a(t)]_{u,i}[\Delta_a(t)]_{v,i}[\bar{H}]_v^\top \mathbf{V}_{N,a}(t)[\bar{H}]_u \tag{183}$$

$$\overset{(\mathrm{CS})}{\leq} (d+1) \sum_{u\in\overline{\mathrm{pa}}(i)} (\Delta_a(t))_{u,i}^2 \left\|[\bar{H}]_u\right\|^2_{\mathbf{V}_{N,a}(t)} \tag{184}$$

$$\overset{(174)}{\leq} (d+1)^{\ell-1}m_B^{2(\ell-k)}\beta^{2k-2}S_a^{2k-2}(t)\lambda_{\max}\big(\mathbf{V}_{N,a}(t)\big) \sum_{u\in\overline{\mathrm{pa}}(i)} ([\Delta_a(t)]_{u,i})^2 \tag{185}$$

$$= (d+1)^{\ell-1}m_B^{2(\ell-k)}\beta^{2k-2}S_a^{2k-2}(t)\lambda_{\max}\big(\mathbf{V}_{N,a}(t)\big) \|[\Delta_a(t)]_i\|^2 \tag{186}$$

$$\overset{(162)}{\leq} (d+1)^{\ell-1}m_B^{2(\ell-k)}\beta^{2k-2}S_a^{2k-2}(t)\lambda_{\max}\big(\mathbf{V}_{N,a}(t)\big)\beta^2\underbrace{\lambda_{\min}^{-1}\big(\mathbf{V}_{i,a}(t)\big)}_{\leq S_a^2(t)} \tag{187}$$

$$\leq (d+1)^{\ell-1}m_B^{2(\ell-k)}\beta^{2k}S_a^{2k}(t)\lambda_{\max}\big(\mathbf{V}_{N,a}(t)\big) . \tag{188}$$

Taking the square-roots of both sides in (181) and (188) yields

$$\|[H]_i\|_{\mathbf{V}_{N,a}(t)} \leq (d+1)^{\frac{\ell-1}{2}}m_B^{\ell-k}\beta^k S_a^k(t)\lambda_{\max}^{1/2}\big(\mathbf{V}_{N,a}(t)\big) , \tag{189}$$

which is the desired inequality for all $k \in [\ell] \cup \{0\}$. This completes the proof of induction.

Step 3: Recall the binomial expansion of $\Delta_a^{(\ell)}(t)$ and focus on its i-th column:

$$\left[\Delta_a^{(\ell)}(t)\right]_i \overset{(173)}{=} \sum_{k=1}^{\ell} \sum_{H\in\mathcal{H}_{\ell,k}} [H]_i , \tag{190}$$

in which we aim to bound $\mathbf{V}_{N,a}(t)$-norm of each $[H]_i$ term in (189). The eigenvalues of $\mathbf{V}_{i,a}(t)$ are always at least 1, which makes $S_a(t) \leq 1$ by definition. Therefore, for all $H \in \bigcup_{1\leq k\leq\ell} \mathcal{H}_{\ell,k}$, we can replace $S_a^k(t)$ in (189) by $S_a(t)$,

$$S_a^k(t) \leq S_a(t) , \quad \forall k \in [\ell] , \tag{191}$$

and $\quad \|[H]_i\|_{\mathbf{V}_{N,a}(t)} \overset{(189)}{\leq} (d+1)^{\frac{\ell-1}{2}}m_B^{\ell-k}\beta^k S_a(t)\lambda_{\max}^{1/2}\big(\mathbf{V}_{N,a}(t)\big) . \tag{192}$

The final result follows by using (192) for each of the $(2^\ell - 1)$ terms in the sum (190) as follows

$$\left\| \left[\Delta_a^{(\ell)}(t) \right]_i \right\|_{\mathbf{V}_{N,a}(t)} = \left\| \sum_{k=1}^{\ell} \sum_{H \in \mathcal{H}_{\ell,k}} [H]_i \right\|_{\mathbf{V}_{N,a}(t)} \tag{193}$$

$$\leq \sum_{k=1}^{\ell} \sum_{H \in \mathcal{H}_{\ell,k}} \left\| [H]_i \right\|_{\mathbf{V}_{N,a}(t)} \tag{194}$$

$$\overset{(192)}{\leq} (d+1)^{\frac{\ell-1}{2}} S_a(t) \lambda_{\max}^{1/2} \left(\mathbf{V}_{N,a}(t) \right) \sum_{k=1}^{\ell} |\mathcal{H}_{\ell,k}| m_B^{\ell-k} \beta^k \tag{195}$$

$$= (d+1)^{\frac{\ell-1}{2}} S_a(t) \lambda_{\max}^{1/2} \left(\mathbf{V}_{N,a}(t) \right) \sum_{k=1}^{\ell} \binom{\ell}{k} m_B^{\ell-k} \beta^k \tag{196}$$

$$= (d+1)^{\frac{\ell-1}{2}} S_a(t) \lambda_{\max}^{1/2} \left(\mathbf{V}_{N,a}(t) \right) m_B^{\ell} \left(\left(\frac{\beta}{m_B} + 1 \right)^{\ell} - 1 \right) \tag{197}$$

$$< (d+1)^{\frac{\ell-1}{2}} S_a(t) \lambda_{\max}^{1/2} \left(\mathbf{V}_{N,a}(t) \right) (\beta + m_B)^{\ell} \tag{198}$$

$$= (d+1)^{\frac{\ell-1}{2}} (\beta + m_B)^{\ell} \left[\lambda_{\max}^{1/2} \left(\mathbf{V}_{N,a}(t) \right) \max_{i \in [N]} \lambda_{\min}^{-1/2} \left(\mathbf{V}_{i,a}(t) \right) \right] . \tag{199}$$

$\blacksquare$

Proof of Corollary 3. Note that while proving (189), we have only used the CS inequality, along with the properties $\| [\mathbf{B}_a]_i \| \leq 1$, $\| [\Delta_a(t)]_i \|_{\mathbf{V}_{i,a}(t)} \leq \beta$, $\| [\mathbf{B}_a]_i \|_0 \leq d+1$, and $\| [\Delta_a(t)]_i \|_0 \leq d+1$. Therefore, for a matrix $\mathbf{A}$ that satisfies the same conditions that $\mathbf{B}_a$ does, and matrix $\Delta_{\mathbf{A}}(t)$ that satisfies the same conditions that $\Delta_a(t)$ does, the result in (49) holds for $[\Delta_{\mathbf{A}}(t)]_N$ following similar steps. $\blacksquare$

The next result relates the singular values of a convex combination of positive semidefinite matrices. The subsequent corollary will be useful when we consider the singular values of the sum of the autocorrelation matrices later.

Lemma 13 *Let $\mathbf{A}_1, \mathbf{A}_2 \ldots \mathbf{A}_n \in \mathbb{R}^{d \times d}$ be positive semidefinite matrices, and denote the minimum and maximum singular values of $\mathbf{A}_i$ by $\sigma_{\min}(\mathbf{A}_i)$ and $\sigma_{\max}(\mathbf{A}_i)$ respectively. Also, let $\alpha_1, \alpha_2 \ldots \alpha_n \geq 0 : \sum_{i=1}^{n} \alpha_i = 1$. Then,*

$$\sigma_{\max} \left(\sum_{i=1}^{n} \alpha_i \mathbf{A}_i \right) \leq \max_{i \in [n]} \sigma_{\max}(\mathbf{A}_i) , \tag{200}$$

$$\text{and} \quad \sigma_{\min} \left(\sum_{i=1}^{n} \alpha_i \mathbf{A}_i \right) \geq \min_{i \in [n]} \sigma_{\min}(\mathbf{A}_i) . \tag{201}$$

Proof For the positive definite matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, its maximum singular value is

$$\sigma_{\max}(\mathbf{A}) = \max_{x \in \mathbb{R}^d : \|x\|=1} x^\top \mathbf{A} x . \tag{202}$$

Therefore, we have

$$\sigma_{\max}\left(\sum_{i=1}^{n}\alpha_i\mathbf{A}_i\right) = \max_{x\in\mathbb{R}^d:\|x\|=1} x^\top\left(\sum_{i=1}^{n}\alpha_i\mathbf{A}_i\right)x \leq \sum_{i=1}^{n}\alpha_i\sigma_{\max}\left(\mathbf{A}_i\right) \leq \max_{i\in[n]}\sigma_{\max}\left(\mathbf{A}_i\right) . \tag{203}$$

Similar arguments with inequalities flipped and using the definition of minimum singular value yield the second result of the lemma. ∎

Corollary 14 *Consider a sequence of interventions $\{a_s : s \in [t]\}$. Then,*

$$\sigma_{\max}\left(\sum_{s=1}^{t}\mathbb{1}_{\{i\in a_s\}}\Sigma_{i,a_s}\right) \leq \left(\sum_{s=1}^{t}\mathbb{1}_{\{i\in a_s\}}\right)\kappa_{i,\max} , \tag{204}$$

$$and \qquad \sigma_{\min}\left(\sum_{s=1}^{t}\mathbb{1}_{\{i\in a_s\}}\Sigma_{i,a_s}\right) \geq \left(\sum_{s=1}^{t}\mathbb{1}_{\{i\in a_s\}}\right)\kappa_{i,\min} . \tag{205}$$

Proof Recall that $N_i(t) \stackrel{(39)}{=} \sum_{s=1}^{t}\mathbb{1}_{\{i\in a_s\}}$. If $N_i(t) = 0$, each of the quantities in the corollary is zero. If $N_i(t) > 0$, the coefficients $\left\{\frac{\mathbb{1}_{\{i\in a_1\}}}{N_i(t)}, \ldots, \frac{\mathbb{1}_{\{i\in a_t\}}}{N_i(t)}\right\}$ constitute a sequence that sums up to 1. Then, we can apply Lemma 13 and the definitions of $\kappa_{i,\max}, \kappa_{i,\min}$ to obtain

$$\sigma_{\max}\left(\sum_{s=1}^{t}\frac{\mathbb{1}_{\{i\in a_s\}}}{N_i(t)}\Sigma_{i,a_s}\right) \leq \max_{s\in[t]}\sigma_{\max}\left(\Sigma_{i,a_s}\right) \leq \kappa_{i,\max} , \tag{206}$$

$$and \qquad \sigma_{\min}\left(\sum_{s=1}^{t}\frac{\mathbb{1}_{\{i\in a_s\}}}{N_i(t)}\Sigma_{i,a_s}\right) \geq \min_{s\in[t]}\sigma_{\min}\left(\Sigma_{i,a_s}\right) \geq \kappa_{i,\min} . \tag{207}$$

By dividing both sides of these two inequalities by $N_i(t)$, (206) and (207) imply (204) and (205). ∎

Lemma 15 *Consider matrices $\mathbf{D}$ and $\mathbf{A}$ that satisfy*

$$\left\|\mathbf{D}^\top\mathbf{D} - \mathbf{A}\right\| \leq \zeta . \tag{208}$$

Then we have,

$$\sigma_{\max}\left(\mathbf{D}\right) \leq \sqrt{\sigma_{\max}\left(\mathbf{A}\right)} + \frac{\zeta}{\sqrt{\sigma_{\max}\left(\mathbf{A}\right)}} , \tag{209}$$

$$and \qquad \sigma_{\min}\left(\mathbf{D}\right) \geq \max\left\{0, \sqrt{\sigma_{\min}\left(\mathbf{A}\right)} - \frac{\zeta}{\sqrt{\sigma_{\min}\left(\mathbf{A}\right)}}\right\} . \tag{210}$$

Equivalently, if

$$\sigma_{\max}\left(\mathbf{D}\right) \geq \sqrt{\sigma_{\max}\left(\mathbf{A}\right)} + \frac{\zeta}{\sqrt{\sigma_{\max}\left(\mathbf{A}\right)}} \;\; or \;\; \sigma_{\min}\left(\mathbf{D}\right) \leq \max\left\{0, \sqrt{\sigma_{\min}\left(\mathbf{A}\right)} - \frac{\zeta}{\sqrt{\sigma_{\min}\left(\mathbf{A}\right)}}\right\} , \tag{211}$$

then $\left\|\mathbf{D}^\top\mathbf{D} - \mathbf{A}\right\| \geq \zeta$.

Proof We prove it via bounding $\|\mathbf{D}x\|^2$. For vector x that satisfies $\|x\| = 1$ we have

$$| \|\mathbf{D}x\|^2 - x^\top \mathbf{A}x| = |\langle(\mathbf{D}^\top\mathbf{D} - \mathbf{A})x, x\rangle| \overset{\text{(CS)}}{\leq} \sigma_{\max}\left(\mathbf{D}^\top\mathbf{D} - \mathbf{A}\right) \overset{\text{(208)}}{\leq} \zeta. \tag{212}$$

We have the following immediate conclusions for (208):

$$\sigma_{\min}\left(\mathbf{A}\right) - \zeta \leq \min_x x^\top \mathbf{A}x - \zeta \leq \|\mathbf{D}x\|^2 \leq \max_x x^\top \mathbf{A}x + \zeta \leq \sigma_{\max}\left(\mathbf{A}\right) + \zeta, \tag{213}$$

$$\|\mathbf{D}x\| \leq \sqrt{\sigma_{\max}\left(\mathbf{A}\right) + \zeta}, \qquad \forall x, \tag{214}$$

$$\sigma_{\max}\left(\mathbf{D}\right) \leq \sqrt{\sigma_{\max}\left(\mathbf{A}\right) + \zeta} \leq \sqrt{\sigma_{\max}\left(\mathbf{A}\right)} + \frac{\zeta}{\sqrt{\sigma_{\max}\left(\mathbf{A}\right)}}, \tag{215}$$

$$\|\mathbf{D}x\|^2 \geq \max\left\{0, \sigma_{\min}\left(\mathbf{A}\right) - \zeta\right\}, \tag{216}$$

$$\|\mathbf{D}x\| \geq \sqrt{\max\left\{0, \sigma_{\min}\left(\mathbf{A}\right) - \zeta\right\}}, \quad \forall x, \tag{217}$$

$$\sigma_{\min}\left(\mathbf{D}\right) \geq \sqrt{\max\left\{0, \sigma_{\min}\left(\mathbf{A}\right) - \zeta\right\}} \tag{218}$$

$$\geq \max\left\{0, \sqrt{\sigma_{\min}\left(\mathbf{A}\right)} - \frac{\zeta}{\sqrt{\sigma_{\min}\left(\mathbf{A}\right)}}\right\}. \tag{219}$$

For the second statement of the lemma, denote the events in the lemma by $\mathcal{Z}_1 \triangleq \{\|\mathbf{D}^\top\mathbf{D} - \mathbf{A}\| \leq \zeta\}$ and

$$\mathcal{Z}_2 \triangleq \left\{\sigma_{\max}\left(\mathbf{D}\right) \leq \sqrt{\sigma_{\max}\left(\mathbf{A}\right)} + \frac{\zeta}{\sqrt{\sigma_{\max}\left(\mathbf{A}\right)}} \quad \text{and} \right.$$

$$\left. \sigma_{\min}\left(\mathbf{D}\right) \geq \max\left\{0, \sqrt{\sigma_{\min}\left(\mathbf{A}\right)} - \frac{\zeta}{\sqrt{\sigma_{\min}\left(\mathbf{A}\right)}}\right\}\right\}. \tag{220}$$

In the first step we showed $\mathcal{Z}_1 \subseteq \mathcal{Z}_2$, which implies $\mathcal{Z}_2^{\text{c}} \subseteq \mathcal{Z}_1^{\text{c}}$. ■

Freedman's inequality. Freedman (1975, Theorem 1.6) is a martingale extension of Bernstein-type concentration inequalities. An earlier extension of Freedman's inequality to matrix martingales is given by Oliveira (2009, Theorem 1.2). The result of Oliveira (2009) requires a uniform bound $\sigma_{\max}\left(\mathbf{Z}(k)\right) \leq R$ while the final result similarly involves $\lambda_{\max}\left(\mathbf{Y}(k)\right)$. Tropp (2011, Theorem 1.2) achieves a similar concentration result for $\lambda_{\max}\left(\mathbf{Y}(k)\right)$ while requiring $\lambda_{\max}\left(\mathbf{Z}(k)\right) \leq R$, which is a weaker condition compared to $\sigma_{\max}\left(\mathbf{Z}(k)\right) \leq R$. However, when we define our matrix martingale sequence $\mathbf{Y}(k)$ in our proofs, we seek bounds for $\sigma_{\max}\left(\mathbf{Y}(k)\right)$. Therefore, we derive a parallel result to Tropp (2011, Theorem 1.2) with a uniform bound condition on $\sigma_{\max}\left(\mathbf{Z}(k)\right)$ and a final result with $\sigma_{\max}\left(\mathbf{Y}(k)\right)$.

Lemma 16 (Matrix Freedman) *Consider a matrix martingale* $\{\mathbf{Y}(k) : k = 0, 1, \dots\}$ *whose values are self-adjoint matrices with dimension* n, *and let* $\{\mathbf{Z}(k) : k = 1, 2, \dots\}$ *be the difference sequence. Assume that the difference sequence is uniformly bounded in the sense that*

$$\sigma_{\max}\left(\mathbf{Z}(k)\right) \leq R, \quad \text{almost surely} \quad \forall k \in \mathbb{N}^+. \tag{221}$$

Define the predictable quadratic variation process of the martingale as

$$\mathbf{W}(k) \triangleq \sum_{j=1}^{k} \mathbb{E}[\mathbf{Z}(j)^2 \mid \mathcal{F}_{j-1}] , \quad \forall k \in \mathbb{N}^+ . \tag{222}$$

Then, for all $\varepsilon \geq 0$ and $\sigma^2 > 0$,

$$\mathbb{P}\{\exists k : \sigma_{\max}(\mathbf{Y}(k)) \geq \varepsilon \ \text{and} \ \|\mathbf{W}(k)\| \leq \sigma^2\} \leq n \exp\left\{\frac{-\varepsilon^2/2}{\sigma^2 + R\varepsilon/3}\right\} \tag{223}$$

$$\leq n \exp\left\{-\frac{3}{8}\min\left(\frac{\varepsilon^2}{\sigma^2}, \frac{\varepsilon}{R}\right)\right\} . \tag{224}$$

Proof The proof for the most part follows from the steps of Tropp (2011, Theorem 1.2), and we present only the necessary changes to obtain the desired result. For a $c : (0, \infty) \to [0, \infty]$ function and positive number θ, the real-valued function of two self-adjoint matrices is defined as

$$G_\theta(\mathbf{Y}, \mathbf{W}) \triangleq tr \exp(\theta\mathbf{Y} - c(\theta)\mathbf{W}) . \tag{225}$$

Tropp (2011, Lemma 2.2) shows that if $\lambda_{\max}(\mathbf{Y}) \geq t$ and $\lambda_{\max}(\mathbf{W}) \leq w$, then for all $\theta > 0$, $G_\theta(\mathbf{Y}, \mathbf{W}) \geq \exp(\theta t - c(\theta)w)$. We change the conditions to $\sigma_{\max}(\mathbf{Y}) \geq t$ and $\sigma_{\max}(\mathbf{W}) \leq w$, and the proof of Tropp (2011, Lemma 2.2) follows through to show $G_\theta(\mathbf{Y}, \mathbf{W}) \geq \exp(\theta t - c(\theta)w)$ for all $\theta > 0$. By changing all instances of $\lambda_{\max}(\mathbf{Y}(k))$ and $\lambda_{\max}(\mathbf{W}(k))$ to $\sigma_{\max}(\mathbf{Y}(k))$ and $\sigma_{\max}(\mathbf{W}(k))$ in the proof of Tropp (2011, Theorem 1.2), the desired result in Lemma 16 follows directly.

Finally, note that zero-padding a $n \times n$ matrix with additional $(N - n)$ zero rows and $(N - n)$ zero columns leaves its maximum singular value and maximum eigenvalue unchanged. Therefore, the final result will hold for a matrix martingale $\mathbf{Y}(k)$ with dimension N that has only a $n \times n$ non-zero submatrix. ∎

Proof of Lemma 7. We prove the result for $\mathbb{P}(\mathcal{E}_{i,n}(t))$ and the result for $\mathbb{P}(\mathcal{E}_{i,n}^*(t))$ will follow similarly. The core of the proof is using Freedman's concentration inequality for matrix martingales.

Martingale construction. Let us consider node i. We define two interventional and observational martingale sequences for node i. Let $\mathcal{F}_{s-1} \triangleq \sigma(a_1, X(1), \ldots, a_{s-1}, X(s-1), a_s)$ denote the filtration for $s \in [T]$. Define Σ_{i,a_t} as the autocorrelation matrix of $X_{\overline{\mathrm{pa}}(i)}$, which is distributed according to $\mathbb{P}_{a_t}$. Furthermore, define

$$\mathbf{Z}_i(s) \triangleq \mathbb{1}_{\{i \notin a_s\}} \left(X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s) - \Sigma_{i,a_s}\right) , \ \forall s \in [T] , \tag{226}$$

$$\mathbf{Y}_i(k) \triangleq \sum_{s=1}^{k} \mathbf{Z}_i(s) , \ \forall k \in [T] , \tag{227}$$

and $$\mathbf{W}_i(k) \triangleq \sum_{s=1}^{k} \mathbb{E}[\mathbf{Z}_i^2(s) \mid \mathcal{F}_{s-1}] , \ \forall k \in [T] . \tag{228}$$

$\mathbf{Z}_i(k)$ is the difference sequence and $\mathbf{W}_i(k)$ is the predictable quadratic variation of the process. We show that the $\mathbf{Y}_i(k)$ sequence is a martingale, i.e., $\mathbb{E}[\mathbf{Y}_i(k) \mid \mathcal{F}_{k-1}] = \mathbf{Y}_i(k-1)$.

$$\mathbb{E}[\mathbf{Y}_i(k) \mid \mathcal{F}_{k-1}] = \mathbb{E}[\mathbf{Y}_i(k-1) + \mathbf{Z}_i(k) \mid \mathcal{F}_{k-1}] \tag{229}$$

$$= \mathbf{Y}_i(k-1) + \mathbb{E}\left[\mathbb{1}_{\{i \notin a_k\}} \left(X_{\overline{\mathrm{pa}}(i)}(k) X_{\overline{\mathrm{pa}}(i)}^\top(k) - \Sigma_{i,a_k}\right) \mid \mathcal{F}_{k-1}\right]. \tag{230}$$

Action a_k is $\mathcal{F}_{k-1}$-measurable. Hence, the randomness in the expectation after conditioning on $\mathcal{F}_{k-1}$ is induced by $X(k) \sim \mathbb{P}_{a_k}$. Therefore, the expected value in (230) is zero by definition of $\Sigma_{i,a}$ in (13), and the sequence $\mathbf{Y}_i(k)$ defined in (227) is a martingale.

Defining the events. We will use n to denote realizations of $N_i(t)$. For all $n \in [T]$, recall ε_n in (89), and also define σ_n^2 as

$$\sigma_n^2 \triangleq 2m^4 n, \quad \forall n \in [T]. \tag{231}$$

Then, we define the following events for each triplet $\{(i, t, n) : i \in [N], t \in [T], n \in [t]\}$:

$$\mathcal{U}_{i,n}(t) \triangleq \{N_i(t) = n\}, \tag{232}$$

$$\mathcal{Y}_{i,n}(t) \triangleq \{\sigma_{\max}(\mathbf{Y}_i(t)) \geq \varepsilon_n\}, \tag{233}$$

$$\mathcal{Q}_{i,n}(t) \triangleq \{\|\mathbf{W}_i(t)\| \leq \sigma_n^2\}, \tag{234}$$

$$\mathcal{D}_{i,n}(t) \triangleq \left\{\sigma_{\min}(\mathbf{D}_i(t)) \leq \max\left\{0, \sqrt{n\kappa_{\min}} - \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\} \quad \text{or}\right.$$

$$\left. \sigma_{\max}(\mathbf{D}_i(t)) \geq \sqrt{n\kappa_{\max}} + \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\}. \tag{235}$$

We will show the desired result, that is $\mathbb{P}(\mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) \leq (d+1)\exp\left(-\frac{3\alpha^2}{16}\right)$ in four steps.

Step 1: Show that $\mathbb{P}(\mathcal{Q}_{i,n}(t) \mid \mathcal{U}_{i,n}(t)) = 1$. The summands of $\mathbf{W}_i(t)$ are the conditional expectations of the following $\mathbf{Z}_i^2(s)$ terms

$$\mathbf{Z}_i^2(s) = \mathbb{1}_{\{i \notin a_s\}}\left([X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s)]^2 - 2 X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s) \Sigma_{i,a_s} + \Sigma_{i,a_s}^2\right). \tag{236}$$

Note that Σ_{i,a_s} is $\mathcal{F}_{s-1}$-measurable. Hence,

$$\mathbb{E}[X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s) \Sigma_{i,a_s} \mid \mathcal{F}_{s-1}] = \Sigma_{i,a_s} \mathbb{E}[X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s)] = \Sigma_{i,a_s}^2. \tag{237}$$

Using Assumption 1, we also have

$$\left\|[X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s)]^2\right\| \leq \left\|X_{\overline{\mathrm{pa}}(i)}(s) X_{\overline{\mathrm{pa}}(i)}^\top(s)\right\|^2 \leq m^4. \tag{238}$$

43

Taking the norm of the expected values on both sides in (236), we obtain

$$\left\| \mathbb{E}[\mathbf{Z}_i^2(s) \mid \mathcal{F}_{s-1}] \right\| \overset{(237)}{=} \left\| \mathbb{E}[\mathbb{1}_{\{i \notin a_s\}}(X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s))^2 \mid \mathcal{F}_{s-1}] - \Sigma_{i,a_s}^2 \mathbb{E}[\mathbb{1}_{\{i \notin a_s\}} \mid \mathcal{F}_{s-1}] \right\|$$

(239)

$$\leq \left\| \mathbb{E}[\mathbb{1}_{\{i \notin a_s\}}(X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s))^2 \mid \mathcal{F}_{s-1}] \right\| + \underbrace{\Sigma_{i,a_s}^2 \mathbb{E}[\mathbb{1}_{\{i \notin a_s\}} \mid \mathcal{F}_{s-1}]}_{\leq m^4}$$

(240)

$$\overset{(238)}{\leq} 2m^4 \mathbb{E}[\mathbb{1}_{\{i \notin a_s\}} \mid \mathcal{F}_{s-1}]$$

(241)

$$= 2m^4 \mathbb{1}_{\{i \notin a_s\}} .$$

(242)

Note that (242) is correct since $\mathbb{1}_{\{i \notin a_s\}}$ is $\mathcal{F}_{s-1}$-measurable. Subsequently, $\|\mathbf{W}_i(t)\|$ is bounded by

$$\|\mathbf{W}_i(t)\| = \left\| \sum_{s=1}^t \mathbb{E}[\mathbf{Z}_i^2(s) \mid \mathcal{F}_{s-1}] \right\|$$

(243)

$$\leq \sum_{s=1}^t \left\| \mathbb{E}[\mathbf{Z}_i^2(s) \mid \mathcal{F}_{s-1}] \right\|$$

(244)

$$\overset{(242)}{\leq} \sum_{s=1}^t 2m^4 \mathbb{1}_{\{i \notin a_s\}}$$

(245)

$$= 2m^4 N_i(t) .$$

(246)

Given that under the event $\mathcal{U}_{i,n}(t)$ we have $N_i(t) = n$, we obtain

$$\|\mathbf{W}_i(t)\| \overset{(246)}{\leq} 2m^4 N_i(t) = 2m^4 n = \sigma_n^2 ,$$

(247)

$$\text{and} \qquad \mathbb{P}(\mathcal{Q}_{i,n}(t) \mid \mathcal{U}_{i,n}(t)) = 1 ,$$

(248)

since event $\mathcal{U}_{i,n}(t)$ implies event $\mathcal{Q}_{i,n}(t)$.

Step 2: Show that $\mathbb{P}(\mathcal{Y}_{i,n}(t) \mid \mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) = 1$. Let us define $\mathbf{A} = \sum_{s=1}^t \mathbb{1}_{\{i \notin a_s\}} \Sigma_{i,a_s}$. From the definition of the martingale sequence $\mathbf{Y}_i(k)$ in (227) we have

$$\mathbf{Y}_i(t) \overset{(227)}{=} \sum_{s=1}^t \mathbb{1}_{\{i \notin a_s\}} \left(X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s) - \Sigma_{i,a_s} \right)$$

(249)

$$\overset{(34)}{=} \mathbf{D}_i^\top(t)\mathbf{D}_i(t) - \sum_{s=1}^t \mathbb{1}_{\{i \notin a_s\}} \Sigma_{i,a_s}$$

(250)

$$= \mathbf{D}_i^\top(t)\mathbf{D}_i(t) - \mathbf{A} .$$

(251)

Given that under the event $\mathcal{U}_{i,n}(t)$ we have $N_i(t) = n$, Corollary 14 indicates that

$$\sigma_{\max}(\mathbf{A}) = \sigma_{\max}\left(\sum_{s=1}^{t} \mathbb{1}_{\{i \notin a_s\}} \Sigma_{i,a_s}\right) \leq N_i(t)\kappa_{\max} = n\kappa_{\max} , \tag{252}$$

$$\text{and} \qquad \sigma_{\min}(\mathbf{A}) = \sigma_{\min}\left(\sum_{s=1}^{t} \mathbb{1}_{\{i \notin a_s\}} \Sigma_{i,a_s}\right) \geq N_i(t)\kappa_{\min} = n\kappa_{\min} . \tag{253}$$

Therefore, the event $\mathcal{D}_{i,n}(t)$ implies that at least one of the following two inequalities is correct:

$$\sigma_{\min}(\mathbf{D}_i(t)) \leq \max\left\{0, \sqrt{n\kappa_{\min}} - \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}}\right\} \overset{(252)}{\leq} \max\left\{0, \sqrt{\sigma_{\min}(\mathbf{A})} - \frac{\varepsilon_n}{\sqrt{\sigma_{\min}(\mathbf{A})}}\right\} , \tag{254}$$

$$\text{and} \quad \sigma_{\max}(\mathbf{D}_i(t)) \geq \sqrt{n\kappa_{\max}} + \frac{\varepsilon_n}{\sqrt{n\kappa_{\min}}} \overset{(253)}{\geq} \sqrt{\sigma_{\max}(\mathbf{A})} + \frac{\varepsilon_n}{\sigma_{\min}(\mathbf{A})} . \tag{255}$$

Given the events $\mathcal{U}_{i,n}(t)$ and $\mathcal{D}_{i,n}(t)$, we invoke the second statement of Lemma 15 to obtain

$$\left\|\mathbf{D}_i^\top(t)\mathbf{D}_i(t) - \mathbf{A}\right\| \geq \varepsilon_n , \tag{256}$$

which is the event $\mathcal{Y}_{i,n}(t)$. Therefore, we have

$$\mathbb{P}(\mathcal{Y}_{i,n}(t) \mid \mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) = 1 . \tag{257}$$

Step 3: Show that $\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{Q}_{i,n}(t)) \leq (d+1)\exp\left(-\frac{3\alpha^2}{16}\right)$**.** The norm of the difference sequence $\mathbf{Z}_i(s)$ for martingale $\mathbf{Y}_i(k)$ is bounded as

$$\|\mathbf{Z}_i(s)\| = \left\|\mathbb{1}_{\{i \notin a_s\}}\left(X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s) - \Sigma_{i,a_s}\right)\right\| \tag{258}$$

$$\leq \left\|X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s) - \Sigma_{i,a_s}\right\| \tag{259}$$

$$\leq \left\|X_{\overline{\mathrm{pa}}(i)}(s)X_{\overline{\mathrm{pa}}(i)}^\top(s)\right\| + \|\Sigma_{i,a_s}\| \tag{260}$$

$$\leq \left\|X_{\overline{\mathrm{pa}}(i)}(s)\right\|^2 + \underbrace{\kappa_{i,\max}}_{\leq m^2} \tag{261}$$

$$\leq 2m^2 . \tag{262}$$

Next, we apply Lemma 16 (matrix Freedman) with $R = 2m^2$ to obtain

$$\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{Q}_{i,n}(t)) = \mathbb{P}\{\sigma_{\max}(\mathbf{Y}_i(t)) \geq \varepsilon_n \ \text{ and } \ \|\mathbf{W}_i(t)\| \leq \sigma_n^2\} \tag{263}$$

$$\leq \mathbb{P}\{\exists k \in [T] : \sigma_{\max}(\mathbf{Y}_i(k)) \geq \varepsilon_n \ \text{ and } \ \|\mathbf{W}_i(k)\| \leq \sigma_n^2\} \tag{264}$$

$$\leq (d+1)\exp\left(-\frac{3}{8}\min\left\{\frac{\varepsilon_n^2}{\sigma_n^2}, \frac{\varepsilon_n}{2m^2}\right\}\right) \tag{265}$$

$$= (d+1)\exp\left(-\frac{3}{8}\min\left\{\frac{\alpha^2 \max\{n, \alpha^2\}}{2n}, \frac{\alpha \max\{\sqrt{n}, \alpha\}}{2}\right\}\right) \tag{266}$$

$$= (d+1)\exp\left(-\frac{3\alpha^2}{16}\right) . \tag{267}$$

45

Finally, note that it can be easily verified that in both cases of $\alpha \geq \sqrt{n}$ and $\alpha < \sqrt{n}$ we have

$$\frac{\alpha^2}{2} = \min \left\{ \frac{\alpha^2 \max\{n, \alpha^2\}}{2n}, \frac{\alpha \max\{\sqrt{n}, \alpha\}}{2} \right\} , \tag{268}$$

which, in turn, implies (267).

Step 4: Show that $\mathbb{P}(\mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) \leq (d+1) \exp\left(-\frac{3\alpha^2}{16}\right)$. We are ready to combine the last three steps and establish the desired result. Since $\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{Q}_{i,n}(t)) \leq (d+1) \exp\left(-\frac{3\alpha^2}{16}\right)$, it suffices to show that $\mathbb{P}(\mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) \leq \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{Q}_{i,n}(t))$. First, due to the Step 1 result $\mathbb{P}(\mathcal{Q}_{i,n}(t) \mid \mathcal{U}_{i,n}(t)) = 1$, we have

$$\mathbb{P}(\mathcal{Q}_{i,n}(t) \mid \mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t)) = \mathbb{P}(\mathcal{Q}_{i,n}(t) \mid \mathcal{U}_{i,n}(t)) = 1 , \tag{269}$$

$$\text{and} \quad \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t), \mathcal{Q}_{i,n}(t)) = \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t))\mathbb{P}(\mathcal{Q}_{i,n}(t) \mid \mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t)) \tag{270}$$

$$\stackrel{(269)}{=} \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t)) . \tag{271}$$

Furthermore, using the Step 3 result $\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{Q}_{i,n}(t)) \leq (d+1) \exp\left(-\frac{3\alpha^2}{16}\right)$, we have

$$\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t)) \stackrel{(271)}{=} \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t), \mathcal{Q}_{i,n}(t)) \tag{272}$$

$$\leq \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{Q}_{i,n}(t)) \tag{273}$$

$$\leq (d+1) \exp\left(-\frac{3\alpha^2}{16}\right) . \tag{274}$$

Next, using the Step 2 result $\mathbb{P}(\mathcal{Y}_{i,n}(t) \mid \mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) = 1$, we obtain

$$\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) = \mathbb{P}(\mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t))\mathbb{P}(\mathcal{Y}_{i,n}(t) \mid \mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) \tag{275}$$

$$\stackrel{(257)}{=} \mathbb{P}(\mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) . \tag{276}$$

Finally, using $\mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t)) \leq (d+1) \exp\left(-\frac{3\alpha^2}{16}\right)$, we have

$$\mathbb{P}(\mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) \stackrel{(276)}{=} \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{D}_{i,n}(t), \mathcal{U}_{i,n}(t)) \tag{277}$$

$$\leq \mathbb{P}(\mathcal{Y}_{i,n}(t), \mathcal{U}_{i,n}(t)) \tag{278}$$

$$\stackrel{(274)}{\leq} (d+1) \exp\left(-\frac{3\alpha^2}{16}\right) , \tag{279}$$

which is the desired result. The interventional counterpart result, i.e., $\mathbb{P}(\mathcal{E}_{i,n}^*(t)) \leq (d+1) \exp\left(-\frac{3\alpha^2}{16}\right)$, can be shown similarly. $\blacksquare$

Proof of Lemma 8. We will use the fact that h is a non-increasing function. First, note that there may exist multiple nodes that achieve $\max_{i \in [N]} h(N_{i,a_t}(t))$. Without loss of generality, we select the smallest solution as $\arg\max$ (or $\arg\min$) when the $\max$ (or $\min$) of a function over set $[N]$ has more than one solution. Then, since h is a non-increasing function, we have

$$\arg\max_{i \in [N]} h(N_{i,a_t}(t)) = \arg\min_{i \in [N]} N_{i,a_t}(t) . \tag{280}$$

46

Also, by definition,

$$N_{i,a_t}(t) \overset{(40)}{=} \mathbb{1}_{\{i \in a_t\}} N_i^*(t) + \mathbb{1}_{\{i \notin a_t\}}(t - N_i^*(t)) . \tag{281}$$

Therefore, the argument of (280) is the node that has the smallest relevant counter variable $N_i^*(t)$ for the nodes $i \in a_t$, and $(t - N_i^*(t))$ for the nodes $i \notin a_t$. We denote this node by i_t as follows

$$i_t \triangleq \arg\max_{i \in [N]} h(N_{i,a_t}(t)) \overset{(280)}{=} \arg\min_{i \in [N]} N_{i,a_t}(t) . \tag{282}$$

Note that i_t does not capture whether $i \in a_t$ or $i \notin a_t$. In other words, i_t does not specify whether $N_{i_t,a_t}(t) = N_{i_t}^*(t)$ or $N_{i_t,a_t}(t) = N_{i_t}(t)$. Driven by addressing the challenge that it causes, define the set of time indices where $i_t = i$ for each of these two cases as follows

$$\mathcal{S}_i \triangleq \{t \in [T] : i_t = i, \, i \notin a_t\} , \quad \forall i \in [N] , \tag{283}$$

$$\text{and} \quad \mathcal{S}_i^* \triangleq \{t \in [T] : i_t = i, \, i \in a_t\} , \quad \forall i \in [N] . \tag{284}$$

Subsequently, $h(N_{i_t,a_t}(t))$ becomes

$$\max_{i \in [N]} h(N_{i,a_t}(t)) = h(N_{i_t,a_t}(t)) = \begin{cases} h(N_i^*(t)) , & \forall t \in \mathcal{S}_i^*(t) \\ h(N_i(t)) , & \forall t \in \mathcal{S}_i(t) \end{cases} . \tag{285}$$

Denote the elements of $\mathcal{S}_i$ by $S_{i,1}, \ldots, S_{i,|\mathcal{S}_i|}$. Note that until time $S_{i,n}$, the event $\{i_t = i, i \notin a_t\}$ occurs exactly n times. Similarly $\{i_t = i, i \in a_t\}$ occurs n times until time $S_{i,n}^*$. Then,

$$n = \sum_{t=1}^{S_{i,n}} \mathbb{1}_{\{i_t=i,i\notin a_t\}} \leq \sum_{t=1}^{S_{i,n}} \mathbb{1}_{\{i\notin a_t\}} = N_i(S_{i,n}) , \tag{286}$$

$$\text{and} \quad n = \sum_{t=1}^{S_{i,n}^*} \mathbb{1}_{\{i_t=i,i\in a_t\}} \leq \sum_{t=1}^{S_{i,n}^*} \mathbb{1}_{\{i\in a_t\}} = N_i^*(S_{i,n}^*) . \tag{287}$$

Using the results above and noting that h is a non-increasing function, we obtain

$$\sum_{t=1}^{T} \max_{i \in [N]} h(N_{i_t,a_t}(t)) = \sum_{t=1}^{T} h(N_{i_t,a_t}(t)) \tag{288}$$

$$= \sum_{i=1}^{N} \sum_{t:t\in\mathcal{S}_i} h(N_i(t)) + \sum_{i=1}^{N} \sum_{t:t\in\mathcal{S}_i^*} h(N_i^*(t)) \tag{289}$$

$$= \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i|} \underbrace{h(N_i(S_{i,n}))}_{\overset{(286)}{\leq} h(n)} + \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i^*|} \underbrace{h(N_i^*(S_{i,n}^*))}_{\overset{(287)}{\leq} h(n)} \tag{290}$$

$$\leq \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i|} h(n) + \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i^*|} h(n) . \tag{291}$$

47

To bound the discrete sums through integrals, we define

$$H(y) = \int_{x=0}^{y} h(x)dx , \quad y \geq 0 . \tag{292}$$

Since $h(x)$ is a positive, non-increasing function, for any $k \in \mathbb{N}^+$ we have

$$\sum_{n=1}^{k} h(n) \leq \int_{x=1}^{k+1} h(x)dx \leq \int_{x=0}^{k} h(x)dx = H(k) , \tag{293}$$

and subsequently,

$$\sum_{t=1}^{T} \max_{i \in [N]} h(N_{i,a_t}(t)) \leq \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i|} h(n) + \sum_{i=1}^{N} \sum_{n=1}^{|\mathcal{S}_i^*|} h(n) \tag{294}$$

$$\overset{(293)}{\leq} \sum_{i=1}^{N} H(|\mathcal{S}_i|) + \sum_{i=1}^{N} H(|\mathcal{S}_i^*|) . \tag{295}$$

We verify that H is a concave function as follows

$$H'(n) = h(n) - h(0) \tag{296}$$

$$= \sqrt{2} \left(\frac{1}{\max\left\{0, \sqrt{n\kappa_{\min}} - \sqrt{\tau\kappa_{\min}}\right\} + 1} - 1 \right) \tag{297}$$

$$= \begin{cases} 0 , & 0 \leq n < \tau \\ \sqrt{2}(\sqrt{n\kappa_{\min}} - \sqrt{\tau\kappa_{\min}} + 1)^{-1} , & n \geq \tau \end{cases} , \tag{298}$$

$$\text{and} \quad H''(n) = \begin{cases} 0 , & 0 \leq n < \tau \\ -\sqrt{\frac{\kappa_{\min}}{2n}}(\sqrt{n\kappa_{\min}} - \sqrt{\tau\kappa_{\min}} + 1)^{-2} , & n \geq \tau \end{cases} . \tag{299}$$

Since $H''(n) \leq 0$ for the domain of H, H is a concave function. Also, $\sum_{i=1}^{N} |\mathcal{S}_i| + \sum_{i=1}^{N} |\mathcal{S}_i^*| = T$. Then, we use Jensen's inequality for concave functions to obtain

$$\sum_{i=1}^{N} H(|\mathcal{S}_i|) + \sum_{i=1}^{N} H(|\mathcal{S}_i^*|) \leq 2N \times H\left(\frac{1}{2N} \sum_{i=1}^{N} |\mathcal{S}_i| + \frac{1}{2N} \sum_{i=1}^{N} |\mathcal{S}_i^*|\right) \tag{300}$$

$$= 2N \times H\left(\frac{T}{2N}\right) . \tag{301}$$

Next, we note that $h(x) = \sqrt{2}$ for $x \leq \tau$ and decompose $H\left(\frac{T}{2N}\right)$ into two parts as

$$H\left(\frac{T}{2N}\right) = \int_{x=0}^{\frac{T}{2N}} h(x)dx \tag{302}$$

$$= \int_{x=0}^{\tau} h(x)dx + \int_{x=\tau}^{\frac{T}{2N}} h(x)dx \tag{303}$$

$$= \sqrt{2}\tau + \int_{x=\tau}^{\frac{T}{2N}} h(x)dx . \tag{304}$$

48

Next, we need to bound $\int_{x=\tau}^{\frac{T}{2N}} h(x)dx$. Let us define $y \triangleq \sqrt{\tau} - \frac{1}{\sqrt{\kappa_{\min}}}$. Hence,

$$\int_{x=\tau}^{\frac{T}{2N}} h(x)dx = \sqrt{2} \int_{x=\tau}^{\frac{T}{2N}} \frac{1}{\sqrt{x\kappa_{\min}} - \sqrt{\tau\kappa_{\min}} + 1}dx \tag{305}$$

$$= \sqrt{\frac{2}{\kappa_{\min}}} \int_{x=\tau}^{\frac{T}{2N}} \frac{1}{\sqrt{x} - y}dx \tag{306}$$

$$= 2\sqrt{\frac{2}{\kappa_{\min}}} \left(\sqrt{\frac{T}{2N}} + y\log\left(\sqrt{\frac{T}{2N}} - y\right) - \sqrt{\tau} - y\log(\sqrt{\tau} - y)\right) \tag{307}$$

$$< \sqrt{\frac{2}{\kappa_{\min}}} \left(\sqrt{\frac{2T}{N}} + y\log\left(\frac{T}{2N}\right)\right) \tag{308}$$

$$< \sqrt{\frac{2}{\kappa_{\min}}} \left(\sqrt{\frac{2T}{N}} + \sqrt{\tau}\log\left(\frac{T}{2N}\right)\right) . \tag{309}$$

Finally, plugging (309) into (304), we obtain the desired result

$$\sum_{t=1}^{T} \max_{i\in[N]} h(N_{i,a_t}(t)) \leq 2N \times \left(\tau + \int_{x=\tau}^{\frac{T}{2N}} h(x)dx \right) \tag{310}$$

$$< 2N\left(\sqrt{2}\tau + \sqrt{\frac{2}{\kappa_{\min}}} \left(\sqrt{\frac{2T}{N}} + \sqrt{\tau}\log\left(\frac{T}{2N}\right)\right)\right) . \tag{311}$$

■

Proof of Theorem 10. Before starting the proof, we comment on the changes from the frequentist setting.

Global lower and upper bounds for singular values. We have defined global bounds for singular values in (14) and (15) for the frequentist analysis. However, the probability measure $\mathbb{P}_a$ changes with respect to the sampled parameters $\mathbf{W}$ for the Bayesian setting. Therefore, we need to expand the definition of lower and upper bounds for singular values to the domain of $\mathbf{W}$, i.e., $\mathcal{W}$. We redefine the probability measure for an intervention a by also accounting for the dependence on parameters $\mathbf{W}$, and denote it by $\mathbb{P}_a^{\mathbf{W}}$. Accordingly, we define $\Sigma_{i,a}^{\mathbf{W}} \triangleq \mathbb{E}_{X\sim\mathbb{P}_a^{\mathbf{W}}}[X_{\overline{\mathrm{pa}}(i)}X_{\overline{\mathrm{pa}}(i)}^{\top}]$ and denote the lower and upper bounds of these moments' singular values by

$$\kappa_{\max}^{\mathbf{W}} \triangleq \max_{i\in[N]} \max_{a\in\mathcal{A}} \sigma_{\max}\left(\Sigma_{i,a}^{\mathbf{W}}\right) , \tag{312}$$

$$\kappa_{\max}^{\mathcal{W}} \triangleq \max_{\mathbf{W}\in\mathcal{W}} \kappa_{\max}^{\mathbf{W}} , \tag{313}$$

$$\kappa_{\min}^{\mathbf{W}} \triangleq \min_{i\in[N]} \min_{a\in\mathcal{A}} \sigma_{\min}\left(\Sigma_{i,a}^{\mathbf{W}}\right) , \tag{314}$$

$$\text{and} \qquad \kappa_{\min}^{\mathcal{W}} \triangleq \min_{\mathbf{W}\in\mathcal{W}} \kappa_{\min}^{\mathbf{W}} . \tag{315}$$

Similar to the definition of τ in (120) for the frequentist setting, we define

$$\tau_{\mathbf{W}} \triangleq \frac{\alpha^2 m^4}{(\kappa_{\min}^{\mathbf{W}})^2} , \quad \text{and} \quad \tau_{\mathcal{W}} \triangleq \frac{\alpha^2 m^4}{(\kappa_{\min}^{\mathcal{W}})^2} . \tag{316}$$

We start by finding bounds on the expected regret for fixed $\mathbf{W}$ using the tools that we have developed in Section 4. Then, we analyze the terms in the result that are affected by the choice of $\mathbf{W}$, and obtain the final Bayesian regret. Consider

$$\mathbb{E}_\epsilon[R_\mathbf{W}(T)] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[\mu_{a^*} - \mu_{a_t} \mid \tilde{\mathcal{F}}_{t-1}]\right], \tag{317}$$

where $\tilde{\mathcal{F}}_t = \sigma(a_1, X(1), \dots, a_t, X(t))$. Note that $\tilde{\mathcal{F}}_t$ is different from $\mathcal{F}_t$ defined and used earlier. The next few steps are similar to those of the proof of Theorem 5. We use the same β as before, i.e.,

$$\beta = m_B + \sqrt{2\log(2NT) + (d+1)\log(1 + m^2 T/(d+1))}, \tag{318}$$

and the upper confidence bound $\text{UCB}(t)$ in (46). Defining the event $\mathcal{E}_\cap$ similarly to (60) in the proof of Theorem 5, we have $\mathbb{P}(\mathcal{E}_\cap^c) \leq \frac{1}{T}$. Next, we decompose the regret in (317) as

$$\sum_{t=1}^T \mathbb{E}[\mu_{a^*} - \mu_{a_t} \mid \tilde{\mathcal{F}}_{t-1}] = \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_\cap^c}\underbrace{(\mu_{a^*} - \mu_{a_t})}_{\leq 2m} \mid \tilde{\mathcal{F}}_{t-1}] + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_\cap}(\mu_{a^*} - \mu_{a_t}) \mid \tilde{\mathcal{F}}_{t-1}] \tag{319}$$

$$\leq 2mT\underbrace{\mathbb{P}(\mathcal{E}_\cap^c \mid \tilde{\mathcal{F}}_{t-1})}_{\leq 1/T} + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_\cap}(\mu_{a^*} - \mu_{a_t}) \mid \tilde{\mathcal{F}}_{t-1}] \tag{320}$$

$$\leq 2m + \sum_{t=1}^T \mathbb{E}[\mathbb{1}_{\mathcal{E}_\cap}(\mu_{a^*} - \mu_{a_t}) \mid \tilde{\mathcal{F}}_{t-1}]. \tag{321}$$

Thompson sampling has the property that $\mathbb{P}(a^* = a \mid \tilde{\mathcal{F}}_{t-1}) = \mathbb{P}(a_t = a \mid \tilde{\mathcal{F}}_{t-1})$. Therefore,

$$\text{UCB}_{a^*}(t) = \text{UCB}_{a_t}(t), \tag{322}$$

and consequently,

$$\mathbb{E}[\mathbb{1}_{\mathcal{E}_\cap}(\mu_{a^*} - \mu_{a_t}) \mid \tilde{\mathcal{F}}_{t-1}] = \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap}\underbrace{\left(\mu_{a^*} - \text{UCB}_{a^*}(t)\right.}_{\leq 0} + \text{UCB}_{a_t}(t) - \mu_{a_t}) \mid \tilde{\mathcal{F}}_{t-1}\right] \tag{323}$$

$$\leq \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap}\left(\text{UCB}_{a_t}(t) - \mu_{a_t}\right) \mid \tilde{\mathcal{F}}_{t-1}\right]. \tag{324}$$

Define $\text{UCB}_a(t) = f(\tilde{\mathbf{B}}_a)$. By combining (321) and (324), we obtain

$$\mathbb{E}_\epsilon[R_\mathbf{W}(T)] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}[\mu_{a^*} - \mu_{a_t} \mid \tilde{\mathcal{F}}_{t-1}]\right] \tag{325}$$

$$\leq 2m + \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap}\left(\text{UCB}_{a_t}(t) - \mu_{a_t}\right) \mid \tilde{\mathcal{F}}_{t-1}\right]\right] \tag{326}$$

$$= 2m + \mathbb{E}\left[\sum_{t=1}^T \mathbb{E}\left[\mathbb{1}_{\mathcal{E}_\cap}(f(\tilde{\mathbf{B}}_{a_t}) - f(\mathbf{B}_{a_t})) \mid \tilde{\mathcal{F}}_{t-1}\right]\right] \tag{327}$$

$$= 2m + \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}_{\mathcal{E}_\cap}(f(\tilde{\mathbf{B}}_{a_t}) - f(\mathbf{B}_{a_t}))\right] . \tag{328}$$

Following similar steps to the proof of Theorem 5, we can bound the expected value in (328), and obtain

$$\mathbb{E}_\epsilon[R_\mathbf{W}(T)] \leq 2m + 2(\beta + m_B)^{L+1}(d+1)^{\frac{L}{2}}\lambda_{\mathbf{W},T} , \tag{329}$$

where

$$\lambda_{\mathbf{W},T} < \frac{4g(\tau_\mathbf{W})}{\sqrt{\kappa_{\mathbf{W},\min}}}\sqrt{NT} + 3(N+1)\tau_\mathbf{W} g(\tau_\mathbf{W}) + \frac{2\sqrt{2}N\sqrt{\tau_\mathbf{W}}g(\tau_\mathbf{W})}{\sqrt{\kappa_{\mathbf{W},\min}}}\log\left(\frac{T}{2N}\right)$$
$$+ \frac{m}{T} + \frac{2m}{3} + 1 . \tag{330}$$

Finally, we replace $\mathbf{W}$-specific terms with global $\mathcal{W}$ terms as follows

$$\kappa_{\mathcal{W},\min} \leq \kappa_{\mathbf{W},\min} , \tag{331}$$

$$\tau_\mathbf{W} \overset{(316)}{\leq} \tau_\mathcal{W} , \tag{332}$$

$$\sqrt{\tau_\mathbf{W}}g(\tau_\mathbf{W}) \leq \sqrt{\tau_\mathcal{W}}g(\tau_\mathcal{W}) , \tag{333}$$

$$\lambda_{\mathbf{W},T} \leq \lambda_{\mathcal{W},T} . \tag{334}$$

Since $\lambda_{\mathbf{W},T}$ is the only term in (329) that depends on $\mathbf{W}$, we obtain

$$\text{BR}(T) = \mathbb{E}_\mathcal{W}\mathbb{E}_\epsilon[R_\mathbf{W}(T)] \tag{335}$$

$$\leq 2m + 2(\beta + m_B)^{L+1}(d+1)^{\frac{L}{2}}\lambda_{\mathcal{W},T} . \tag{336}$$

Similar to the proof of Theorem 9, using the same β as in Theorem 5, and ignoring poly-logarithmic terms and constants, we obtain $\text{BR}(T) = \tilde{\mathcal{O}}(d^{L+\frac{1}{2}}\sqrt{NT})$. ∎

Appendix C. Proof of Theorem 11

Sketch of the proof. To find a lower bound on the minimax regret, we provide a proof as follows. Let Π be the set of all policies on a set of stochastic bandit environments $\mathcal{I}$. The minimax regret is

(a) Hierarchical block with $d = 3$.

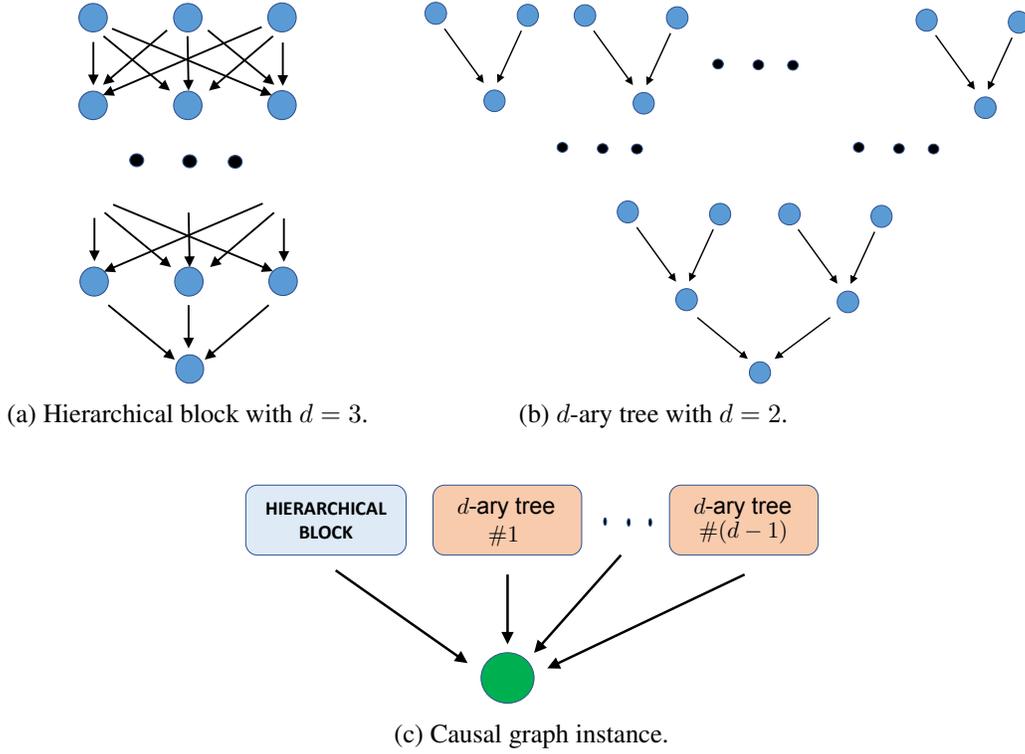(b) d-ary tree with $d = 2$.

(c) Causal graph instance.

Figure 6: Sample diagrams for the graphs used in the proof of Theorem 11.

defined as

$$\inf_{\pi \in \Pi} \sup_{\mathcal{I}^* \in \mathcal{I}} \mathbb{E}_{\pi, \mathcal{I}^*}[R(T)] \,, \tag{337}$$

in which $\mathbb{E}_{\pi, \mathcal{I}^*}[R(T)]$ denotes the expected regret of policy π on the bandit instance $\mathcal{I}^*$. We will consider a set $\tilde{\mathcal{I}}$ that contains two bandit instances. By definition of minimax regret, a lower bound for the regret of any policy on $\tilde{\mathcal{I}}$ also is a lower bound for the minimax regret since

$$\inf_{\pi \in \Pi} \sup_{\mathcal{I}^* \in \mathcal{I}} \mathbb{E}_{\pi, \mathcal{I}^*}[R(T)] \geq \inf_{\pi \in \Pi} \sup_{\mathcal{I}^* \in \tilde{\mathcal{I}}} \mathbb{E}_{\pi, \mathcal{I}^*}[R(T)] \,. \tag{338}$$

After this intuition, the central idea of the proof is as follows. Two linear SEM causal bandit instances that differ by only one edge parameter are hard to distinguish. At the same time, we can construct them to have different optimal actions, indicating that a selection policy cannot incur small regret for both at the same time under the same data realization. Note that, the difference of the rewards, or equivalently the regrets, observed by these two bandit instances under the same action can be computed by tracing the effect of that differing edge parameter over all the paths that end at the reward node. We carefully build graphs to maximize the number of such paths for given d and L. In this section, we provide details of these steps.

We consider two linear SEM causal bandit instances that share the same graph $\mathcal{G}$ and are parameterized by $\mathcal{I}_1 \triangleq \{\mathbf{H}, \mathbf{H}^*, \nu, \nu^*, \epsilon\}$, and $\mathcal{I}_2 \triangleq \{\bar{\mathbf{H}}, \bar{\mathbf{H}}^*, \bar{\nu}, \bar{\nu}^*, \epsilon\}$. For both instances, let the weights of all observational edges be b, and all interventional edges be $b - \delta$ such that $0 < \delta < b \leq 1$.

In other words, if for $i, j \in [N]$

$$(i \to j) \in \mathcal{G} \,, \tag{339}$$

then

$$[\mathbf{H}]_{i,j} = [\bar{\mathbf{H}}]_{i,j} = b \,, \qquad \text{and} \qquad [\mathbf{H}^*]_{i,j} = [\bar{\mathbf{H}}^*]_{i,j} = b - \delta \,. \tag{340}$$

We assume the noise terms have standard Gaussian distributions, i.e., $\epsilon_i \sim \mathcal{N}(0,1)$ for all $i \in [N]$. Furthermore, for all $k \in \{2, \ldots, N\}$ set

$$\nu_k = \nu_k^* = \bar{\nu}_k = \bar{\nu}_k^* = 0 \,. \tag{341}$$

The only difference between the two instances is the affine terms of node 1, for which $\nu_1 = \bar{\nu}_1^* = 1$, and $\nu_1^* = \bar{\nu}_1 = 1 - \delta$. Note that this parameterization can be applied to any graph $\mathcal{G}$. Next, consider a fixed bandit policy π that generates the following filtration over time

$$\mathcal{F}_t \triangleq \{a_1, X(1), \ldots, a_t, X(t)\} \,. \tag{342}$$

The decision of π at time t is $\mathcal{F}_{t-1}$-measurable. Accordingly, define $\mathbb{P}_t$ and $\bar{\mathbb{P}}_t$ as the probability measures induced by $\mathcal{F}_t$ by t rounds of interaction between π and the two bandit instances. When it is clear from context, we use the shorthand terms $\mathbb{P}$ and $\bar{\mathbb{P}}$ for $\mathbb{P}_T$ and $\bar{\mathbb{P}}_T$, respectively. We will show that π cannot suffer small regret in both instances at the same time and under the same filtration $\mathcal{F}_T$.

Next, we construct our sample graph for fixed d and L values using the following two building blocks.

1. One hierarchical block as depicted in Figure 6a, which consists of $(L-1)$ layers each with d nodes. Adjacent layers are fully connected. There exists a final layer with one node fully connected to layer $L-1$.

2. One d-ary tree with L layers as depicted in Figure 6b.

We use one hierarchical block and $(d-1)$ number of d-ary trees and connect their sink nodes to form a reward node. Hence, in this graph, the hierarchical block consists of $d(L-1)+1$ nodes, a d-ary tree consists of $\sum_{\ell=0}^{L-1} d^\ell$ nodes, and the total number of nodes is

$$N = d(L-1) + 1 + (d-1) \sum_{\ell=0}^{L-1} d^\ell = d^L + d(L-1) \,. \tag{343}$$

The nodes in the hierarchical blocks are labeled by $\{1, \ldots, d(L-1)+1\}$, beginning from the top layer. All weights are in the set $\{b, b-\delta\}$ and are positive in both bandit instances. Hence, by Lemma 1, the optimal action is the one that maximizes the value of each entry of $\mathbf{B}_a$ and $\bar{\mathbf{B}}_a$. As a result, the optimal actions are $a^* = \emptyset$ for the bandit instance $\mathcal{I}_1$ and $\bar{a}^* = \{1\}$ for bandit instance $\mathcal{I}_2$. Define $\mathcal{E}_{\text{lb}}$ as the event in which node 1 is intervened at least $\frac{T}{2}$ times after T rounds, i.e.,

$$\mathcal{E}_{\text{lb}} \triangleq \left\{ N_1^*(T) \geq \frac{T}{2} \right\} \,. \tag{344}$$

We note that the event $\mathcal{E}_{\text{lb}}$ is defined on the sigma-algebra defined by the filtration $\mathcal{F}_t$, that induces both $\mathbb{P}_t$ and $\bar{\mathbb{P}}_t$. We compute the expected instantaneous regret when node 1 is intervened in the first bandit. Note that since $\nu_k = \nu_k^*$ and ϵ_k have a zero mean for $k \geq 2$, only the paths that start at node 1 and end at the reward node N contribute to the expected regret. Furthermore, since every weight is positive, in $\mathcal{I}_1$, we have $\mu_{\{1\}} \geq \mu_a$ for any a that contains 1. Therefore, for $\mathcal{I}_1$, the expected value of instantaneous regret is at least $\mu_{\emptyset} - \mu_{\{1\}}$ if 1 is intervened. Then, by definition of $\mathcal{E}_{\text{lb}}$, we have

$$\mathbb{E}_{\mathbb{P}}[R(t)] = \mathbb{E}_{\mathbb{P}}\left[\sum_{t=1}^{T} r(t)\right] \tag{345}$$

$$\geq \mathbb{E}_{\mathbb{P}}\left[\sum_{t \in [T]: 1 \notin a_t} (\mu_{\emptyset} - \mu_{\{1\}})\right] \tag{346}$$

$$\geq \mathbb{P}(\mathcal{E}_{\text{lb}})\frac{T}{2}(\mu_{\emptyset} - \mu_{\{1\}}) \tag{347}$$

$$= \mathbb{P}(\mathcal{E}_{\text{lb}})\frac{T}{2}\delta b^L d^{L-2} \ . \tag{348}$$

The final equality holds because there exists d^{L-2} paths from node 1 to node N, and the difference between $\nu_1 - \nu_1^* = \delta$ is multiplied with a b factor for every edge along a path. Similarly, for $\mathcal{I}_2$, we have $\bar{\mu}_{\emptyset} \geq \bar{\mu}_a$ for any a that does not contain 1, and expected value of instantaneous regret is at least $\bar{\mu}_{\{1\}} - \bar{\mu}_{\emptyset}$ if 1 is not intervened. Applying the same steps, we obtain

$$\mathbb{E}_{\bar{\mathbb{P}}}[R(t)] = \mathbb{E}_{\bar{\mathbb{P}}}\left[\sum_{t=1}^{T} r(t)\right] \tag{349}$$

$$\geq \mathbb{E}_{\bar{\mathbb{P}}}\left[\sum_{t \in [T]: 1 \in a_t} (\bar{\mu}_{\{1\}} - \bar{\mu}_{\emptyset})\right] \tag{350}$$

$$\geq \bar{\mathbb{P}}(\mathcal{E}_{\text{lb}}^{\text{c}})\frac{T}{2}(\bar{\mu}_{\{1\}} - \bar{\mu}_{\emptyset}) \tag{351}$$

$$= \bar{\mathbb{P}}(\mathcal{E}_{\text{lb}}^{\text{c}})\frac{T}{2}\delta b^L d^{L-2} \ . \tag{352}$$

By combining (348) and (352) we have

$$\mathbb{E}_{\mathbb{P}}[R(t)] + \mathbb{E}_{\bar{\mathbb{P}}}[R(t)] \geq \frac{T}{2} \delta b^L d^{L-2} [\mathbb{P}(\mathcal{E}_{\text{lb}}) + \bar{\mathbb{P}}(\mathcal{E}_{\text{lb}}^{\text{c}})] \ . \tag{353}$$

By setting $b = \sqrt{1/d}$, we ensure that $m_B = 1$, and (353) becomes

$$\mathbb{E}_{\mathbb{P}}[R(t)] + \mathbb{E}_{\bar{\mathbb{P}}}[R(t)] \geq \frac{T}{2} \delta d^{\frac{L}{2}-2} [\mathbb{P}(\mathcal{E}_{\text{lb}}) + \bar{\mathbb{P}}(\mathcal{E}_{\text{lb}}^{\text{c}})] \ . \tag{354}$$

Next, we characterize a lower bound on $(\mathbb{P}(\mathcal{E}_{\text{lb}}) + \bar{\mathbb{P}}(\mathcal{E}_{\text{lb}}^{\text{c}}))$ that involves the Kullback-Leibler (KL) divergence between $\mathbb{P}$ and $\bar{\mathbb{P}}$, denoted by $\text{D}_{\text{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}})$. For this purpose, we leverage the following theorem.

Theorem 17 (Bretagnolle-Huber inequality) *Let $\mathbb{P}$ and $\bar{\mathbb{P}}$ be probability measures on the same measurable space $(\Omega, \mathcal{F})$ and let $A \in \mathcal{F}$ be an arbitrary event. Then,*

$$\mathbb{P}(A) + \bar{\mathbb{P}}(A^c) \geq \frac{1}{2} \exp(-D_{\mathrm{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}})) . \tag{355}$$

By invoking Theorem 17, from (354) we obtain

$$\mathbb{E}_{\mathbb{P}}[R(t)] + \mathbb{E}_{\bar{\mathbb{P}}}[R(t)] \geq \frac{T}{4} \, \delta d^{\frac{L}{2} - 2} \, \exp(-D_{\mathrm{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}})) . \tag{356}$$

It remains to compute $\exp(-D_{\mathrm{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}}))$ to conclude our proof, for which we leverage the following result.

Proposition 18 *The KL divergence between $\mathbb{P}$ and $\bar{\mathbb{P}}$ is equal to*

$$D_{\mathrm{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}}) = T\delta^2 . \tag{357}$$

Proof Note that a Bayesian network factorizes as

$$\mathbb{P}(X_1, \ldots, X_N) = \prod_{i=1}^{N} \mathbb{P}(X_i \mid X_{\mathrm{pa}(i)}) . \tag{358}$$

Additionally, the two bandit instances differ only in the mechanism of node 1, which is a root node. Then, $D_{\mathrm{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}})$ can be simplified as

$$D_{\mathrm{KL}}(\mathbb{P} \parallel \bar{\mathbb{P}}) = \sum_{i=1}^{N} D_{\mathrm{KL}}(\mathbb{P}(X_i \mid X_{\mathrm{pa}(i)}) \parallel \bar{\mathbb{P}}(X_i \mid X_{\mathrm{pa}(i)})) = D_{\mathrm{KL}}(\mathbb{P}(X_1) \parallel \bar{\mathbb{P}}(X_1)) . \tag{359}$$

Hence, we only need to consider $D_{\mathrm{KL}}(\mathbb{P}(X_1) \parallel \bar{\mathbb{P}}(X_1))$ under two cases: when 1 is observed, and 1 is intervened. We have that

$$X_1 \sim \begin{cases} \mathcal{N}(\nu_1, 1) , & \text{under } \mathbb{P} \text{ when } 1 \notin a \\ \mathcal{N}(\nu_1 - \delta, 1) , & \text{under } \mathbb{P} \text{ when } 1 \in a \\ \mathcal{N}(\nu_1 - \delta, 1) , & \text{under } \bar{\mathbb{P}} \text{ when } 1 \notin a \\ \mathcal{N}(\nu_1, 1) , & \text{under } \bar{\mathbb{P}} \text{ when } 1 \notin a \end{cases} . \tag{360}$$

By noting that

$$D_{\mathrm{KL}}(\mathcal{N}(\nu_1, 1) \parallel \mathcal{N}(\nu_1 - \delta, 1)) = D_{\mathrm{KL}}(\mathcal{N}(\nu_1 - \delta, 1) \parallel \mathcal{N}(\nu_1, 1)) = \frac{\delta^2}{2} , \tag{361}$$

from (360) we obtain

$$D_{\mathrm{KL}}(\mathbb{P}(X_1) \parallel \bar{\mathbb{P}}(X_1))$$
$$= \sum_{t \in [T]: 1 \notin a_t} D_{\mathrm{KL}}(\mathcal{N}(\nu_1, 1), \mathcal{N}(\nu_1 - \delta, 1)) + \sum_{t \in [T]: 1 \in a_t} D_{\mathrm{KL}}(\mathcal{N}(\nu_1 - \delta, 1), \mathcal{N}(\nu_1, 1)) \tag{362}$$
$$= N_1^*(T) \frac{\delta^2}{2} + (T - N_1^*(T)) \frac{\delta^2}{2} \tag{363}$$
$$= \frac{T\delta^2}{2} . \tag{364}$$

■

By applying Proposition 18 on (356) and setting $\delta = \sqrt{2/T}$, we obtain

$$\max\{\mathbb{E}_{\mathbb{P}}[R(t)], \mathbb{E}_{\bar{\mathbb{P}}}[R(t)]\} \geq \frac{1}{2}(\mathbb{E}_{\mathbb{P}}[R(t)] + \mathbb{E}_{\bar{\mathbb{P}}}[R(t)]) \tag{365}$$

$$\geq \frac{T}{8}\, \delta d^{\frac{L}{2}-2} \, \exp(-T\delta^2/2) \tag{366}$$

$$= \frac{\exp(-1)}{8\sqrt{2}} \, d^{\frac{L}{2}-2} \, \sqrt{T}\,. \tag{367}$$

Hence, for $c = \frac{\exp(-1)}{8\sqrt{2}}$, the policy π incurs a regret $cd^{\frac{L}{2}-2}\sqrt{T}$ in at least one of the two bandit instances. Finally, note that removing the nodes from the tree blocks of the constructed graph does not affect the analysis. Hence, N can take any value in the range $\{(d(L-1)+2), \ldots, d^L+d(L-1)+1\}$.
■

Appendix D. A Depiction of the Non-linearity

In Lemma 1, we have shown that the reward X_N is a linear function of ϵ variables but not a linear function of the edge weights. Specifically, the effect of a node on the reward is compounded via the edge weights along the paths from that node to the reward node. This renders our problem completely different from the linear bandit problem. Consider the example in Figure 7. Since the dummy noise variable is $\epsilon_0 = 1$ and ν values are put into the dummy 0-th row of $\mathbf{B}$, the reward node X_5 can be written as
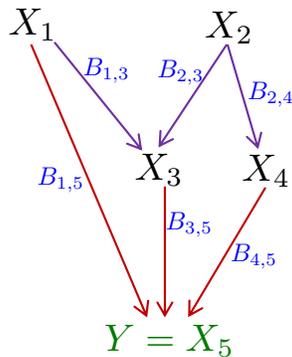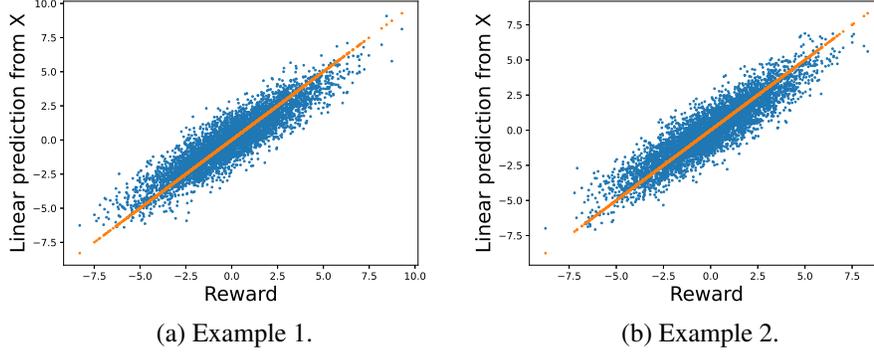

Figure 7: Sample graph with $N = 5$ nodes and its edge weights.

56

(a) Example 1. (b) Example 2.

Figure 8: Sample results of estimating reward from X.

$$X_5 = [\mathbf{B}]_{1,5}X_1 + [\mathbf{B}]_{3,5}X_3 + [\mathbf{B}]_{4,5}X_4 + (\nu_5 + \epsilon_5) \tag{368}$$

$$= [\mathbf{B}]_{1,5}(\nu_1 + \epsilon_1) + [\mathbf{B}]_{3,5}\left([\mathbf{B}]_{1,3}(\nu_1 + \epsilon_1) + [\mathbf{B}]_{2,3}(\nu_2 + \epsilon_2) + \nu_3 + \epsilon_3\right)$$

$$+ [\mathbf{B}]_{4,5}\left([\mathbf{B}]_{2,4}(\nu_2 + \epsilon_2) + \nu_4 + \epsilon_4\right) + (\nu_5 + \epsilon_5) \tag{369}$$

$$= \underbrace{\epsilon_0}_{=1}\left([\mathbf{B}]_{0,1}[\mathbf{B}]_{1,5} + [\mathbf{B}]_{0,1}[\mathbf{B}]_{1,3}[\mathbf{B}]_{3,5} + [\mathbf{B}]_{0,2}[\mathbf{B}]_{2,3}[\mathbf{B}]_{3,5} + [\mathbf{B}]_{0,2}[\mathbf{B}]_{2,4}[\mathbf{B}]_{4,5}\right.$$

$$\left. + [\mathbf{B}]_{0,3}[\mathbf{B}]_{3,5} + [\mathbf{B}]_{0,4}[\mathbf{B}]_{4,5} + [\mathbf{B}]_{0,5}\right)$$

$$+ \epsilon_1\left([\mathbf{B}]_{1,5} + [\mathbf{B}]_{1,3}[\mathbf{B}]_{3,5}\right)$$

$$+ \epsilon_2\left([\mathbf{B}]_{2,3}[\mathbf{B}]_{3,5} + [\mathbf{B}]_{2,4}[\mathbf{B}]_{4,5}\right)$$

$$+ \epsilon_3[\mathbf{B}]_{3,5}$$

$$+ \epsilon_4[\mathbf{B}]_{4,5}$$

$$+ \epsilon_5 . \tag{370}$$

We also demonstrate this non-linearity aspect via a simple simulation. Note that, we observe only vector X. Hence, if we are to estimate reward X_N as a linear function, our only choice is to use X. To do so, we create a hierarchical graph with $N = 10$ nodes, degree $d = 3$, and $L = 3$ layers. The parameters are randomly chosen similarly to the simulations in Section 7. We generate 5000 training and 5000 test samples. Then, we perform linear regression to estimate X_N from $\{1, X_1, \ldots, X_{N-1}\}$ using training data. We use the estimated parameters to predict X_N on test data. Figure 8 shows the distribution of actual reward versus predicted reward on two different trials. Unsurprisingly, even for a simple model with $N = 10$ nodes and 5000 data samples, the reward is highly non-linear with respect to X_N, and equivalently to edge weight parameters.

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Proc. Advances in Neural Information Processing Systems*, pages 2312–2320, Granada, Spain, December 2011.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proc. International Conference on Machine Learning*, pages 127–135, Atlanta, GA, June 2013.

E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *Proc. Advances in Neural Information Processing Systems*, Montréal, Canada, December 2015.

B. Bilodeau, L. Wang, and D. M. Roy. Adaptively exploiting d-separators with causal bandits. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

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

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proc. International Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, April 2011.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Proc. Conference on Learning Theory*, pages 355–366, Helsinki, Finland, July 2008.

A. de Kroon, J. Mooij, and D. Belgrave. Causal bandits without prior knowledge using separating sets. In *Conference on Causal Learning and Reasoning*, pages 407–427, Eureka, CA, April 2022.

S. Feng and W. Chen. Combinatorial causal bandits. In *Proc. of the AAAI Conference on Artificial Intelligence*, Washington, DC, February 2023.

D. A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Proc. Advances in Neural Information Processing Systems*, pages 9551–9561, December 2020.

F. Lattimore, T. Lattimore, and M. D. Reid. Causal bandits: Learning good interventions via causal inference. In *Proc. Advances in Neural Information Processing Systems*, Barcelona, Spain, December 2016.

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020.

S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng. Reinforcement learning for clinical decision support in critical care: Comprehensive review. *Journal of Medical Internet Research*, 22 (7):e18477, July 2020.

Y. Lu, A. Meisami, A. Tewari, and W. Yan. Regret analysis of bandit problems with causal background knowledge. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 141–150, August 2020.

Y. Lu, A. Meisami, and A. Tewari. Causal bandits with unknown graph structure. In *Proc. Advances in Neural Information Processing Systems*, pages 24817–24828, December 2021.

A. Maiti, V. Nair, and G. Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *Proc. Conference on Uncertainty in Artificial Intelligence*, Eindhoven, Netherlands, August 2022.

V. Nair, V. Patil, and G. Sinha. Budgeted and non-budgeted causal bandits. In *Proc. International Conference on Artificial Intelligence and Statistics*, pages 2017–2025, April 2021.

R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv:0911.0600*, 2009.

P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

N. Sawant, C. B. Namballa, N. Sadagopan, and H. Nassif. Contextual multi-armed bandits for causal marketing. *arXiv:1810.01859*, 2018.

R. Sen, K. Shanmugam, A. G. Dimakis, and S. Shakkottai. Identifying best interventions through online importance sampling. In *Proc. International Conference on Machine Learning*, pages 3057–3066, Sydney, Australia, August 2017.

W. Shen, J. Wang, Y.-G. Jiang, and H. Zha. Portfolio choices with orthogonal bandit learning. In *Proc. International Joint Conference on Artificial Intelligence*, Buenos Aires, Argentina, July 2015.

A. Tewari and S. A. Murphy. *From Ads to Interventions: Contextual Bandits in Mobile Health*, pages 495–517. Springer International Publishing, Cham, 2017.

J. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer. Scalable intervention target estimation in linear models. In *Proc. Advances in Neural Information Processing Systems*, pages 1494–1505, December 2021.

N. Xiong and W. Chen. Combinatorial pure exploration of causal bandits. In *Proc. International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, Kigali, Rwanda, May 2023.

A. Yabe, D. Hatano, H. Sumita, S. Ito, N. Kakimura, T. Fukunaga, and K. Kawarabayashi. Causal bandits with propagating inference. In *Proc. International Conference on Machine Learning*, pages 5508–5516, Stockholm, Sweden, July 2018.

Q. Zhou, X. Zhang, J. Xu, and B. Liang. Large-scale bandit approaches for recommender systems. In *Proc. International Conference on Neural Information Processing*, pages 811–821, Guangzhou, China, November 2017.